



University of Bergamo & University of Pavia
Ph.D. Program in Linguistics – XXXII Cycle

A UD Literary Treebank For German

**A Case Study on the Fragments from *Athenaeum* and *Lyceum*:
Processing, Annotation, and Treebank-Based Analysis**

Ph.D. Thesis

Ph.D. Candidate: Alessio Salomoni

Supervisor: Silvia Luraghi

Academic Year 2018-2019

Wenn seine Werke auch nicht übermäßig viel Bildung enthalten, so sind sie doch gebildet: das Ganze ist wie das Einzelne und umgekehrt; kurz, er ist fertig.

F. Schlegel, *Athenaeum*, Fragment 421

Acknowledgements

Thanks to Silvia Luraghi and Marco Passarotti for supervising this project. Many thanks to Sandra Kübler for all her helpful observations. I am grateful to Daniel Zeman, who welcomed me at ÚFAL in Prague, when the October's colours made this city marvellous. He showed great interest toward this research project right from the beginning, and promptly helped me solve different problems many times in these years. Many thanks to Cristina Mariotti, who gave me the opportunity to teach English at the faculty of Political Sciences in Pavia. It was definitely a breath of fresh air, which let me find new energy and motivation. Thanks to all the people with whom I shared beautiful moments both in Bergamo and in Pavia, in particular Shan Huang, Fabio Lo Iodice, Federica Massia, and Deokhyun Nam. A special thank goes to my friend and colleague Marco Budassi. From Amsterdam to Oltrepò, from the relaxing glasses of wines to the hot tennis courts, many beautiful memories come to mind. I do believe that they will echo for a long time, much more than any linguistic notion will. Thanks to all the friends from the Robbiate area. Times have changed, we have changed, and each one is now facing new great challenges. But it is always beautiful and inspiring talking and meeting when possible. And it was amazing to see almost all of you, all together in my small flat, in that rainy evening, after go-karting. I am very grateful to my parents, Manuela and Luigi, for their endless patience and support. Our beautiful Sunday lunches have been more than a light in many dark weeks. I also thank my brother Riccardo. It is always beautiful and stimulating hearing about architectural theories, foreign friends, new bars, new recipes, and whatever. Finally, the most important thought goes to Silvia. We shared so much beauty in these years, that any word would be useless to evoke even a minimal part of it. Thank you for standing beside me, with your smiles and encouragements. You definitely made these years rich and joyful far more than everything.

Contents

1	Introduction	8
1.1	What is a Treebank and What is it for?	8
1.2	Dependency Treebanks for German: State of the Art and Problems	9
1.3	Goals of the Thesis	13
1.4	Structure of the Thesis	14
2	Treebank Development and NLP Accuracy	16
2.1	Selecting the Target Texts: Fragments of the Early Romanticism	16
2.2	Building the Source Corpus	18
2.3	Data Format: CoNLL-U	19
2.4	Preserving Information of the Original Texts in CoNLL-U	21
2.5	Processing the Fragments	22
2.5.1	Tokenization	22
2.5.2	Building a Test Set	23
2.5.3	Best-Performing NLP Pipeline	24
2.6	NLP-Tools Tests on German	26
2.7	Accuracy Metrics	27
2.8	Tests on Lemmatization	28
2.9	Tests on POS-Tagging	31
2.9.1	Universal Tagset (UTS)	31
2.9.2	Stuttgart-Tübingen Tagset (STTS)	35
2.10	Tests on Dependency Parsing	40
2.10.1	In-Depth Evaluation	44
2.11	First Release of the Treebank	56
3	Linguistic Annotation	58
3.1	Introduction to Dependency Grammar	58
3.2	The Annotation Scheme: An Overview on Universal Dependencies	60
3.2.1	Parts of Speech	61
3.2.2	Syntactic Relations	64
3.3	The Linguistic Annotation of the Fragments	68
3.3.1	Relations within a Single Clause	70

3.3.1.1	Main Verb in Simple Form, Nominal Subject, Direct Object, Interrogative Adverb, Nominal Modifier, Adjectival Modifier, Preposition, Determiner	70
3.3.1.2	Main Verb in Complex Form, Auxiliary for Past Form, Oblique Argument.....	72
3.3.1.3	Nonverbal Predicate, Nominal Modifier of a Nonverbal Predicate	74
3.3.1.4	Indirect Object, Possessive Determiner.....	75
3.3.1.5	Modal Verb.....	77
3.3.1.6	Main Verb in Passive Voice	78
3.3.1.7	Modal Verb Depending on a Main Verb in Passive Voice	80
3.3.1.8	Modal Verb Modifying Another Modal Verb, with the Main Verb in Passive Voice.....	81
3.3.1.9	Expletive Element, Coordination between Nominals	82
3.3.1.10	Series of Coordinate Items.....	84
3.3.1.11	Copula-Like Verb	86
3.3.1.12	Nominal Sentence, Adverb Coordinated to a Verb	87
3.3.2	Relations between Clauses: Coordination	89
3.3.2.1	Syndetic Coordination Between Verbs	89
3.3.2.2	Asyndetic Coordination Between Verbs	90
3.3.3	Relations between Clauses: Subordination	91
3.3.3.1	Relative Clause (with Secondary Predication in the Main Clause)	91
3.3.3.2	Subordinating Conjunction, Adverbial Clause	93
3.3.3.3	Non-Finite Subordinate Clause.....	95
3.3.3.4	Non-Finite Subordinate Clause with Auxiliary	97
3.3.3.5	Clausal Subject, Verb Followed by Bare Infinitive.....	99
3.3.3.6	Verb Followed by Infinitive with <i>Zu</i>	100
3.3.3.7	Clause as Predicative Part of a Nonverbal Predicate, (Parenthetical Clause)	103
3.3.3.8	Clausal Subject of a Nonverbal Predicate	104
3.3.3.9	Clausal Complement, Adjectival Clause Modifying a Pronoun, (<i>Etwas</i> Followed by a Substantive)	106
3.3.4	Comparative	107
3.3.4.1	Comparative: <i>Als</i> as Subordination Marker.....	107
3.3.4.2	Comparative: <i>Als</i> as Comparative Conjunction Introducing a Noun Phrase	109
3.3.4.3	Comparative: <i>Wie</i> Introducing an Oblique Argument.....	110
3.3.5	Ellipsis	111
3.3.5.1	Gapping, (Appositional Modifier)	111
3.3.5.2	Modal Verb Promotion as Head in <i>Afinite Konstruktion</i>	113

4 Analysis 116

4.1	Goals and Methodology.....	116
4.2	Datasets	118
4.3	Overall Distribution of Parts of Speech (UPOS).....	120
4.4	Overall Distribution of Dependency Relations.....	134
4.5	An Investigation of Predicates	139
4.5.1	Distribution of Verbal and Nonverbal Predicates	139
4.5.2	Verbal Forms.....	161
2.1.1	Existential Clauses	167
4.5.3	Modal Verbs	178
4.5.4	Position of Subordinate Clauses.....	194
5	Conclusions.....	208
6	References.....	223

1 Introduction

1.1 What is a Treebank and What is it for?

A treebank can be defined as a digital linguistically annotated corpus in which syntactic information is encoded beyond the level of parts of speech through machine-readable metadata¹. The first treebanks were developed during sixties, especially for the English language, therefore they have a long-standing tradition in the field of digital linguistic resources. They are sorted in two main groups, according to the syntactic formalism through which syntax is encoded in texts: constituency treebanks and dependency treebanks. In a constituency formalism, syntax consists of *phrases*, such as noun phrases, verbal phrases, or prepositional phrases, which combine with each other to build larger structural units. This process culminates in shaping clauses and sentences. In a tree-like syntactic representation, this formalism generates abstract non-terminal nodes, such as a noun phrases consisting of a noun phrase combined with a prepositional phrase, while the words of the sentence are the terminal nodes. This causes the syntactic tree to be a complex structure, in which the number of nodes is much higher with respect to the number of words standing in the sentence (Cf. Nivre 2005). On the contrary, in a dependency formalism, syntax consists of binary hierarchical relations that directly occur between pair of words, i.e. *dependency relations* linking a head to a dependent (Kübler, McDonald, and Nivre 2009). Therefore, sentences are represented in tree-like form without any non-terminal node: each node of the tree is a word, while each edge connecting a pair of nodes is a syntactic relation (for a detailed illustration of a dependency tree, see Chapter 3). Each dependency relation usually encodes a grammar function, which is usually the syntactic function played by the dependent with respect to the head, such as, among the others, nominal subject, direct object, or adjectival modifier. The core of the whole sentence is usually the main predicate, which works as an *atom with hooks* (Tesnière 1959) toward the other elements of the sentence. In fact, all the other words can be considered as either direct or indirect dependents of the main predicate. Constituency treebanks were very common when treebanks began spreading as linguistic resources in the field of corpora linguistics. However, even if constituency syntax (or phrase-structure grammar) is still used in both syntactic annotation of corpora and descriptive linguistics, dependency syntax (or dependency grammar) has absolutely surged up, over the last years, as the major formalism to design syntactically annotated corpora. This is due to different reasons, including, among the others, the opportunity to better parse those languages showing free word order or relatively free word order, such as German (Nivre 2005), (Kübler and Prokic, 2006). In fact, as said above, the dependency representation

¹ For a definition of treebank, Cf. (Nivre 2009). The term *treebank* appears to have been coined by Leech (Cf. Sampson 2003).

turns out to be simpler with respect to the constituency one, in terms of both formal representation and computation.

As linguistic resources, dependency treebanks have two fundamental purposes. On the one hand, they are used to train and test statistical models for data-driven natural language processing (NLP), especially, even if not only, for syntactic dependency parsing (Kübler, McDonald, and Nivre 2009). On the other hand, they are at disposal for all those corpus-based (or corpus-driven²) linguistic analysis for which the automatic retrieval of linguistic information is needed, especially information concerning syntax. For instance, they can be used to study the distribution of syntactic functions in a textual genre, to compare syntax across different textual genres, to search for empirical evidence of specific syntactic phenomena in order to validate linguistic theories and hypothesis, or even to automatically induce lexical resources, such as verb subcategorization frames. In fact, the syntactic information in dependency treebanks can be quickly searched and retrieved through specific formal query languages, such as PML-TQ (Štěpánek and Pajas 2010). Moreover, depending on the implemented annotation layers, any linguistic information encoded, ranging from POS-tag to semantic information, can be automatically extracted as well (or exploited to train and test NLP models alike), making these corpora very powerful resources for a wide range of both empirical investigations and NLP tasks. For a recent overview on treebanks and their usage for different kinds of linguistic analysis, see e.g. (Abeillé 2012), (Scheible et al. 2013), (Kübler and Zinsmeister 2015), (Ide and Pustejovsky 2017), (Zeman 2018). For an overview on linguistic corpora in general and their exploitation for linguistic research, see e.g. (O’Keeffe and McCarthy 2010). The research project that is here discussed introduces a new dependency treebank for the German language.

1.2 Dependency Treebanks for German: State of the Art and Problems

Let us now focus on the German dependency treebanks available for research purposes. Table 1 summarizes the reference dependency treebanks of the written German language that are freely available online for research purposes. Overall, they provide a considerable amount of data, but they all share two noticeable limits, due to the variety of data that they collect.

² For the difference between corpus-based and corpus-driven approaches, see (Biber 2012)

Treebank³	Tokens	Genre	Historical Variety
TüBa Treebank	1,959,474	General News	Contemporary
TIGER Treebank	900,000	General News	Contemporary
UD - GSD	290,000	General News + Wikipedia	Contemporary
UD - HDT	3,000,000	Sectoral Web News (technology)	Contemporary

Table 1 Reference dependency treebanks of the written German language currently available for research purposes.

First, they only represent contemporary varieties. Second, they mainly represent a single macro textual genre, i.e. news texts, which are gathered both from the internet and newspapers in electronic format. Traditionally, this is the domain which is mostly used to build and use syntactically annotated corpora. However, solid linguistic motivation behind this choice have never been adduced in literature. In fact, the real reasons appear to be mainly functional, that is both the high availability and the easy access of these linguistic data. As a consequence, news texts have become the *de facto* standard variety to work with treebanks without any solid linguistic reason or purpose, also for German (Cf. Dipper, Lüdeling, and Reznicek 2013). Of course, this is a strong bottleneck toward extending the scientific scope of these linguistic resources. This issue involves both the two main applications of these corpora, i.e. data-driven NLP and linguistic analysis. As I anticipated above, dependency treebanks are used for training and testing statistical models for parsing. These models notoriously suffer from the so-called domain change, see e.g. (Gildea 2001), (Petrov and McDonald 2012). In fact, the distribution of syntactic features can remarkably vary across corpora collecting data from different varieties and domains (Biber 1993). This problem involves not only those varieties from the same period of the language, for instance legal texts and newswire texts, but also those varieties belonging to different stages of the history of the language, for instance web texts and letters of the 18th century. Consequently, models trained on the treebanks of contemporary news are expected to attain high or adequate accuracy on data that come from the same variety and related ones. Conversely, when tested on different varieties, they usually show a severe degradation in performance. For a recent work addressing this issue on English, see e.g. (Mukherjee, Kübler, and Scheutz 2017), while for a recent evaluation of parsing models trained on contemporary news data and tested on samples of texts types of the 18th century, see (Salomoni 2017a). As a consequence, semi-automatic approaches are still the most trustable methodology to design treebanks that aim to offer accurately annotated data other than news. In fact, in a semi-automatic approach, the output of the dependency parser is manually checked by one or more annotators. However, this is a very time-consuming and labor-intense task, especially in the case of those varieties requiring trained annotators able to deal with a specific lexicon and complex syntactic constructions. This is the case, for

³ I here report the reference for these treebanks: TüBa (Telljohann et al. 2004), TIGER (Brants et al. 2002), GSD (McDonald et al. 2013a), HDT (Foth et al. 2014). For a recent overview of both TIGER and TüBa, also see (Stefanie Dipper and Kübler 2017).

instance, of some literary varieties or specialized varieties. Consequently, this has especially discouraged the development of historical dependency treebanks for German, i.e. those collecting data from earlier stages of the language. As far as annotated historical corpora without any syntactic annotation is concerned, there are actually different reference corpora available for German. Among them, it is worth mentioning the *Referenzkorpus Mittelhochdeutsch* (REM) with its sub-corpora (Petran et al. 2016). For a recent overview of the historical corpora for German, see (Dipper 2015). By contrast, historical syntactically annotated corpora are still very scarce, especially those originally⁴ annotated with dependency formalism. In fact, only some pilot projects concerning the syntactic annotation of German historical varieties have been reported in literature over the last years. It is worth mentioning the *Mercurius-Baumbank* (Demske 2007), in which texts from a magazine of 1667 called *Nordische Mercurius* (premodern German, 30,000 tokens) was syntactically annotated through a constituency-based scheme⁵; the *Deutsche Diachrone Baumbank* (DDB), (Hirschmann and Linde 2011), which hosts some religious texts from different pre-modern varieties (8,000 tokens) that are syntactically annotated with the same scheme implemented in *Mercurius-Baumbank*; the *Anselm Corpus* (Dipper and Schultz-Balluff 2013), in which a small portion of a single text by St. Anselm, i.e. *Lords's Passion* (Early New High German, 7,000 tokens) was annotated with a dependency scheme; a constituency-annotated literary corpus collecting a small portion of Kafka's *Der Prozess* (Modern German, 7,000 tokens), (Dipper, Lüdeling, and Reznicek 2013). Apart from these projects, no historical dependency treebank for German is currently available for research purposes. As a result, a dramatic amount of varieties of the German literary treasure are still excluded from the benefits offered by this kind of resources. In fact, dependency parsing models trained on literary data are still missing, and comprehensive treebank-based linguistic analysis of genres, authors or works of the German literary history have never been conducted. Such a mechanism does nothing but increasing the long-standing rift between computational linguistics (CL) and many branches of descriptive linguistics, which have massively avoided treebank-based approaches so far. On the one hand, currently existing treebanks will be still merely used as a general test bed for training and testing NLP tools, putting aside the fact these data only represent a very limited portion of the German language. On the other hand, most linguistic studies, especially on literary varieties, will go ahead discarding these resources *in toto* because of the absence of both available annotated data and methodological perspectives, missing the opportunity to extend both the scope and the methodology of their research.

This panorama of the German dependency treebanks clashes with some objective facts. First, there is a huge amount of historical raw texts that is now freely available online for research purposes, thanks to

⁴ By originally, I mean that the dependency annotation is not obtained through any conversion script run on a constituency annotation.

⁵ It is a hybrid scheme (Albert et al. 2003), which is based on constituency syntax but encodes grammar functions as well. It was originally implemented in the Tiger Corpus.

many important digital archives and libraries developed over the last years, such as *Zeno.org*⁶, or *Deutsches Texts Archiv*⁷ (Geyken 2013), to name but a few. Therefore, the raw material for the linguistic annotation of historical data is conspicuous. Second, as mentioned above, despite treebanks have been being used for decades, the importance of dependency treebanks has definitely surged up over the last few years in the wealth of linguistic resources, most notably thanks to the international project Universal Dependencies (UD) (Nivre et al. 2016), which is now counting hundreds of collaborators all over the world.⁸ UD aims at building a large online repository of multilingual dependency treebanks, which is freely accessible for research purposes. Most importantly, the UD community has been constantly focusing on developing a consistent annotation scheme that is able to work across different human languages. Therefore, for the first time in the history of CL, treebanks are not developed according to subjective criteria adopted by local, isolated initiatives, but in compliance with standard guidelines that are shared by a wide, international scientific community. For the importance of standards in the development of linguistic resources, see (Ide et al. 2017). Third, historical dependency treebanks can surely fill a gap in the macro research area of the linguistic approaches to literature. For instance, they can extend the scope of the so-called corpus-based genre analysis. Such a discipline has a long-standing tradition in corpus linguistics for the English language, both on literary corpora and, especially, on contemporary specialized corpora, see e.g. (Biber and Finegan 1989), (D. Biber and Conrad 2001), (Studer 2008), (Short and Leech 2013). However, especially on the literary side, the investigations have mostly been limited to analysis that cannot go beyond the level of parts of speech. Moreover, there are other disciplines within linguistics that could integrate treebank-based approaches in their studies on literary varieties, such as stylistics, see e.g. (Burke 2017), or computational stylometry, see e.g. (Daelemans 2013). In this respect, treebank-based and dependency-based stylistic analysis of literary texts are still a totally unexplored research area, which could actually detect many hidden characteristics of the language, especially concerning, but not limited to, syntax. Furthermore, an interesting debate about the need of integrating literary analysis with linguistic methodologies has risen in the field of German studies over the last years (Bär, Mende, and Steen 2015), which is traditionally bound to intuition-based and speculative approaches. It seems therefore that times are more than mature for the development of resources that can offer new empirical research perspectives on the literary language.

⁶ <http://www.zeno.org> (last access on 23rd September 2019).

⁷ <http://www.deutschestextarchiv.de> (last access on 23rd September 2019).

⁸ Cf. <https://universaldependencies.org/> (last access on 23rd September 2019).

1.3 Goals of the Thesis

The research project that is here discussed aims to develop and analyze a new historical dependency treebank for German, within the UD framework. The treebank is specifically designed to host literary texts, in which information about author, work, and genre is preserved for each sentence, therefore I defined it a literary treebank. In particular, I focused on a case study based on a specific literary genre from the late 18th century, i.e. the Fragments of the early Romanticism. I chose these texts for both their linguistic and cultural importance in the German literary history, as well as for the fact that they could be dealt with as a cohesive textual genre with respect to other literary genres. Fragments are very short texts, often in aphorism-like form, which deal with philosophical issues concerning art, poetry, beauty, and related ones. They are often ironic and cryptic, and they are all written in a very concise style, to deliberately clash with the long and elaborated prose of the neoclassical authors of the same age, such as F. Schiller. They were published in literary magazines. In particular, I focused on two main collections of Fragments mainly written by Fredrich Schlegel, i.e. the Fragments from *Lyceum*, and, most importantly, the Fragments from *Athenaeum*. In particular, *Athenaeum*, published between 1798 and 1800, is considered a milestone of the German literary history, since it was the reference magazine of the early German Romanticism, which enormously contributed to boost the romantic values all over Europe. The Fragments were chosen by the early romantic authors as the perfect textual genre to embody their values, as well as their stylistic principles. For all these reasons, this genre was absolutely worthy to be represented in a dependency treebank for the first time.

The literary treebank was developed through a semi-automatic approach. In this respect, I conducted some tests of NLP tools on a sample of Fragments in order to detect the best-performing pipeline on this genre. The tools were trained on the available annotated data of contemporary varieties, and then tested on a test set of Fragments. Results are reported and discussed. I then illustrate the application of the UD scheme to the Fragments, therefore I discuss the annotation of a lot of different syntactic phenomena. Most importantly, the thesis attempts to show the benefits offered by a dependency treebank to the linguistic analysis of a literary genre. In this respect, I aim at exploiting the UD dependency annotation to empirically investigate some peculiar characteristics of the Fragments, which would be hardly investigated through traditional methods⁹, as well as through common corpus-based approaches that cannot exploit any dependency annotation. By providing new empirical evidence about the language of the Fragments, the thesis attempts therefore to lay the foundation for the future development of a proper treebank-based, and especially dependency-based, linguistic analysis of the literary language. Moreover, the Fragments are compared against two contemporary genres that are represented in the two

⁹ Methods that do not contemplate any corpus-based approach.

main UD treebanks for German, i.e. web news and web texts. In doing so, I aim to shed light on some linguistic differences, if any, between these two *de facto* standard varieties used in treebanking, and the literary variety that is here represented in a dependency treebank for the first time.

1.4 Structure of the Thesis

The thesis is structured as follows. Chapter 2 deals with the development of the treebank, especially in the earlier stages of the process. I first describe the selection of the target data to build the source corpus, and I introduce the data format. Then, I describe the tests conducted on a set of NLP tools, trained on available contemporary data for German, and tested an initial test set of Fragments. These tests primarily aimed to detect the best-performing NLP pipeline on Fragments. I first introduce and explain the best-performing NLP pipeline, which I then used to extend the treebanked data through a semi-automatic approach until the current release (UD 2.6 release). Then, I report and discuss all the evaluations previously conducted on the candidate tools for each task, i.e. lemmatization, POS-tagging and dependency parsing. Finally, I mention the validation of the treebank through the official automatic UD procedure, and the first official release of the treebank within UD (UD 2.4).

Chapter 3 illustrates the process of linguistic annotation, in which the UD scheme version 2 was applied to the Fragments. After a brief introduction on the dependency grammar, I provide an overview on the main characteristics of the UD scheme. I then describe the application of the UD scheme in detail, basing the explanations on dependency trees reported from the first release of the treebank. Examples are grouped into five macro areas, according to the type of syntactic phenomena that are dealt with, i.e. relations within a single clause, coordination between predicates, subordination, comparative constructions and ellipsis. In each of these groups, the annotation of different syntactic relations is reported and discussed.

Chapter 4 deals with the treebank-based analysis of Fragments. The analysis has a twofold goal. On the one hand, different aspects of Fragments are here empirically investigated exploiting the dependency annotation, in order to highlight the benefits offered by a treebank-based approach to the linguistic analysis of a literary variety, especially with respect to common corpus-based approach which cannot exploit any dependency annotation. On the other hand, the features of the Fragments are investigated against those of two contemporary genres which are represented in the main UD treebanks for German. I therefore also highlighted the differences between the Fragments and these two *de facto* standard varieties. After introducing the tool used for extracting data from the treebanks, I first focus on the

distribution of parts of speech, and the overall distribution of dependency relations. I then investigate different features of predicates and some of their direct dependents, thus highlighting the opportunities offered by a predicate-centered analysis. In each of these investigations, I reported quantitative results and I discussed them through concrete examples extracted from the datasets. All the formal queries used to extract data are reported as well.

Finally, the conclusions summarize all the results obtained in this thesis, in the development phase, as well as, especially, in the treebank-based analysis. Moreover, I suggest possible future work which can take advantage of the results produced by this research project.

2 Treebank Development and NLP Accuracy

2.1 Selecting the Target Texts: Fragments of the Early Romanticism

To set the case study for the first version of the literary treebank, I had to define and select the target data. I had different categories at disposal. In fact, I could have focused on a specific author, as well as on specific literary work, or on a specific literary genre, among the others. I opted for the category of genre¹⁰, which is traditionally used in corpora design as reference category to select data (D. Biber 2010). In particular, I aimed to find a specific microgenre which was cohesive enough to be dealt with as a homogeneous unit. For instance, the genre of the German Romances of the late 19th century is too vast and too variable to be dealt with in a single corpus, since each romance could actually be considered as a different micro variety on its own. In fact, each one can be written in a different style with respect to another one from the same period, and even works by the same author can differ a lot. Moreover, romances can deal with a variety of divergent topics, also causing the lexicon to vary a lot from a text to another. Before detecting the target genre, I had to define the historical stage of the German language within which searching for data. I focused on the second half of the 18th century. I made this choice for three reasons. First, this is regarded as one of the most prolific ages ever of the German literary history, since it saw the definitive emergence of the literary language, thanks to the widespread availability of literary magazines, as well as through the rise of new literary genres, such as letters, or Fragments. For an overview on the literary language of this age, see e.g. (Blackall 2011), (Scherer 2014). Therefore, a resource collecting data of this period looked like more than well grounded. Second, a lot of raw texts from this period were available online in different sources. Third, the historical variety of that period was the Modern German (*Moderne Hochdeutsch*), which does not show any substantial variation with respect to the contemporary historical variety (*Gegenwärtige Hochdeutsch*), especially in terms of morphology and spelling, see e.g. (Besch et al. 1998), (Polenz 2009). In the perspective of a treebank-based analysis (see Chapter 4), this would have allowed me to focus on the lexical and syntactic features of this genre, rather than considering spelling and morphological variations. In the end, I focused on the last decade of the 18th century, and I chose the genre of Fragments as case study for the literary treebank.

¹⁰ I do not adhere here to the distinction proposed by Biber (2010) between *genre* and *register*. Therefore, by genre, I mean a literary genre as text type, such a letter, a novel, an essay or a piece of drama.

Fragments are very short texts, often in aphorism-like form, which were adopted as favorite text type by some of the most important authors of the early German Romanticism (*Frühromantik*), such as Schlegel brothers and Novalis. They mainly deal with philosophical issues concerning beauty, art, poetry and related topics, often in a witty and cryptic form, but always in a very concise style. This makes them a cohesive literary genre, since they tend to deal with the same topics, and they are stylistically very similar to each other, regardless of the authorship. Moreover, they are considered a very important genre in the German literary history, since they perfectly embodied the spirit of the new-born Early Romanticism. In fact, they deliberately clashed with the long and elaborated prose of the neoclassical authors of the same age, such as Shiller and Goethe, to name but a few. Therefore, both their linguistic and cultural importance made them particularly worthy to be represented in the format of a dependency treebank for the first time. The Fragments were mainly gathered in collections (*Fragmentsammlungen*) that were released in literary magazines. In particular, I was interested in those Fragments published between 1797 and 1798 in two very important magazines from that period, that is *Lyceum der Schönen Künste* (or simply *Lyceum*) and, most importantly, *Athenäum*. The first *Fragmentsammlung* is called *Kritische Fragmente*, while the second one is called *Athenäums-Fragmente*. Both collections were written by F. Schlegel, even if the *Athenäums-Fragmente* also collect some Fragments by other authors, especially by Novalis and W.A. Schlegel. In particular, *Athenäum*, founded in 1798, has a crucial importance in the German literary history, since it is considered as the reference magazine of the early Romanticism, and represented a moment of transition in the German literary and intellectual panorama, which paved the way for the development of the Romanticism during the first half of the 19th century. Among the other things, the definition of Romantic Poetry is enucleated in the *Athenäums-Fragmente*, precisely in the Fragment 123. For an overview on the importance of the *Athenäum*, see e.g. (Blanchot, Esch, and Balfour 1983). For an overview on the literature of the *Frühromantik*, see e.g. (Behler 2011). A portrait of the target collections of Fragments is outlined in

Table 2. I report the original title of the first critical edition of each *Fragmentsammlung* on footnote. A Fragment is reported in (1), followed by the English translation.¹¹

1 Man nennt viele Künstler, die eigentlich Kunstwerke der Natur sind.¹²

1. Many so-called artists are really products of nature's art.

Work	Author	Year of Publication	Magazine
<i>Kritische Fragmente</i> ¹³	F. Schlegel	1797	<i>Lyceum der schönen Künste</i>
<i>Athenäums-Fragmente</i> ¹⁴	F. Schlegel, A.W. Schlegel et al.	1798	<i>Athenäum</i>

Table 2 The target texts of the literary treebank.

2.2 Building the Source Corpus

I searched for digital raw texts of the genre of Fragments to build the source corpus of the treebank. After an evaluation of the available sources, I focused on the digital library *zeno.org*¹⁵, in which all the hosted texts are public-domain and they are simply displayed as textual content of web pages. They can therefore be easily copied in a text editor to generate plain text files in .txt format. Moreover, both the target collections of Fragments were entirely available. The texts were therefore copied from the web page and pasted in a text editor, then encoded UTF-8 without BOM and saved as plain text files (.txt). A portrait of the raw texts collected in the source corpus is outlined in Table 3. For each text, I reported the information concerning the printed edition from which it was originally digitized on footnote, followed by the permalink to the webpage from which it was obtained. In addition, two further

¹¹ Friedrich Schlegel's *Lucinde and the Fragments*, University of Minnesota Press, 1971. ProQuest Ebook Central, <http://ebookcentral.proquest.com/lib/unibg-ebooks/detail.action?docID=345421>.

¹² F. Schlegel, *Kritische Fragmente*, fragment 1.

¹³ *Kritische Friedrich-Schlegel-Ausgabe*. Erste Abteilung: *Kritische Neuausgabe*, Band 2, München, Paderborn, Wien, Zürich 1967, S. 147-164. Erstdruck in: *Lyceum der schönen Künste* (Berlin), 1. Bd., 2. Teil, 1797.

¹⁴ *Kritische Friedrich-Schlegel-Ausgabe*. Erste Abteilung: *Kritische Neuausgabe*, Band 2, München, Paderborn, Wien, Zürich 1967, S. 165-256. Erstdruck in: *Athenäum* (Berlin), 1. Bd., 2. Stück, 1798

¹⁵ <http://www.zeno.org>

collections of Fragments were found, i.e. *Ideen* by F. Schlegel and *Blüthenstaub* by Novalis. However, both were released in *Athenaeum* in 1798, therefore they are often included in the Fragments from *Athenäum*.

Work	Author	Source	Tokens
<i>Athenäums-Fragmente</i> ¹⁶	F. Schlegel, A.W. Schlegel et al.		
<i>Kritische Fragmente</i> ¹⁷	F. Schlegel	zeno.org	51.900
<i>Ideen</i> ¹⁸	F. Schlegel		
<i>Blüthenstaub</i> ¹⁹	Novalis		

Table 3 Raw texts collected from the online source and included in the source corpus of the treebank.

2.3 Data Format: CoNLL-U

The standard data format that is used to encode treebanks in UD is the CoNLL-U format²⁰, which is an evolution of the CoNLL-X format (Buchholz and Marsi 2006a). A sentence in CoNLL-U format is shown in Table 4, followed by an explanation.

As shown in Table 3, a text file in CoNLL-U format displays each sentence one token per line. There are ten fields for each line, that are separated by tabs from each other. Each field bears specific linguistic information about each token. Such information is encoded through specific metadata, in compliance with the UD annotation guidelines. In the first field, we have a univocal numerical ID, which can range from 1 to n, where n is a decimal number. It marks the position of each token in the sentence. The last ID corresponds to the final punctuation mark. Then, the other fields follow. The type of metadata that is hosted in each of them is indicated in the first row of the file displayed in Table 10 (such row does

¹⁶ Kritische Friedrich-Schlegel-Ausgabe. Erste Abteilung: Kritische Neuausgabe, Band 2, München, Paderborn, Wien, Zürich 1967, S. 165-256. Erstdruck in: Athenäum (Berlin), 1. Bd., 2. Stück, 1798.

Permalink: <http://www.zeno.org/nid/20005618908>

¹⁷ Kritische Friedrich-Schlegel-Ausgabe. Erste Abteilung: Kritische Neuausgabe, Band 2, München, Paderborn, Wien, Zürich 1967, S. 147-164. Erstdruck in: Lyceum der schönen Künste (Berlin), 1. Bd., 2. Teil, 1797.

Permalink: <http://www.zeno.org/nid/20005618886>

¹⁸ Kritische Friedrich-Schlegel-Ausgabe. Erste Abteilung: Kritische Neuausgabe, Band 2, München, Paderborn, Wien, Zürich 1967. Erstdruck in: Athenäum (Berlin), 3. Bd., 1. Stück, 1798.

Permalink: <http://www.zeno.org/nid/20005618916>

¹⁹ Novalis: Schriften. Die Werke Friedrich von Hardenbergs. Band 2, Stuttgart 1960–1977, S. 413-464.

Entstanden 1797/98. Erstdruck in: Athenäum (Berlin), 1. Bd., 1. Stück, 1798. Vier Fragmente stammen von Friedrich Schlegel. Permalink: <http://www.zeno.org/nid/20005446929>

²⁰<http://universaldependencies.org/format.html>

not stand in the original file format, but it was added here for the matter of clarity). I provide a brief explanation.

ID	TOKEN	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC
1	Man	man	PRON	PIS	_	2	nsubj	_	_
				VVFI					
2	nennt	nennen	VERB	N	_	0	root	_	_
3	viele	vieler	PRON	PIAT	_	2	obj	_	_
4	Künstler	Künstler	NOUN	NN	_	2	xcomp	_	SpaceAfter=No
				PUNC					
5	,	--	T	\$,	_	8	punct	_	_
6	die	der	PRON	PRELS	_	8	nsubj	_	_
							advmo		
7	eigentlich Kunstwerk	eigentlich Kunstwerk	ADV	ADV	_	8	d	_	_
8	e	k	NOUN	NN	_	3	acl	_	_
9	der	der	DET	ART	_	10	det	_	_
10	Natur	Natur	NOUN	NN	_	8	nmod	_	_
				VAFI					
11	sind	sein	AUX	N	_	8	cop	_	SpaceAfter=No
				PUNC					
12	.	--	T	\$.	_	2	punct	_	_

Table 4 Representation of the sentence “Man nennt viele Künstler, die eigentlich Kunstwerke der Natur sind.” from the literary treebank in CoNLL-U format.

The field LEMMA hosts the lemma of the token, that is the basic form of a word, as it is represented in a lexicon (Kübler and Zinsmeister 2015). The field UPOS hosts the universal part of speech (UTS). The field XPOS hosts the fine-grained part of speech (STTS). FEATS stand for morphological features. This field hosts tags that specify, for instance, the case of a token, the degree of an adjective and many other features that are encoded through morphemes. They have not been considered in this work, thus this column in the literary treebank is filled with ‘_’, which stands for unspecified value. In fact, it corresponds to an empty value, but this hyphen has necessarily to be present, since no field can be left totally empty. The field HEAD hosts the ID of the token which is the head of the token in that line. 0 is the ID assigned to the fictional head of the main predicate. The field DEPREL hosts information concerning the type of dependency relation of which the token is the dependent. The value could be, for instance, *nsubj* for nominal subject, or *obj* for direct object. The field DEPS should be used for reporting the head of the enhanced dependencies, if any. In this version of the treebank, this field is filled with underscore, since these dependencies were not considered. The last field, MISC, stands for miscellaneous, and it can be used to add any extra annotation. In this work, this field is filled with underscore also, since no extra

metadata was considered. In spite of this, a technical value necessarily stays in the field MISC, that is *SpaceAfter=No*. It is used to automatically turn the text in CoNLL-U format into the original linear raw text. In particular, this value has to occur each time there is no empty space after the token in the original untokenized text. For instance, punctuation marks, such as full stops or commas, are always attached to the preceding word in raw texts. Therefore, the value *SpaceAfter=No* occupies the MISC field of those tokens preceding each of these punctuation marks in the raw text. Finally, sentence boundaries in CoNLL-U files are marked by blank lines. After a blank line, a new sentence begins.

2.4 Preserving Information of the Original Texts in CoNLL-U

Usually, treebanks of the written language collect general samples of textual data, and no information concerning the source of the data or the original texts themselves is encoded. On the contrary, preserving a part of such information was crucial for the literary treebank. In fact, this treebank aims to offer a form of interaction with data other than that usually offered by the treebanks of the contemporary varieties.

```
# newdoc id = lyceum
# newpar id = lyceum-fl
# genre = fragments
# author = Friedrich Schlegel
# work = Lyceum Fragmente
# sent id = lyceum-fl-s1
# text = Man nennt viele Künstler, die eigentlich Kunstwerke der Natur sind.
1 Man man PRON PIS _ 2 nsubj _ _
2 nennt nennen VERB VVFIN _ 0 root _ _
3 viele vieler DET PIAT _ 2 obj _ _
4 Künstler Künstler NOUN NN _ 2 xcomp _ SpaceAfter=No
5 , -- PUNCT $, _ 8 punct _ _
6 die der PRON PRELS _ 8 nsubj _ _
7 eigentlich eigentlich ADV ADV _ 8 advmod _ _
8 Kunstwerke Kunstwerk NOUN NN _ 3 acl _ _
9 der der DET ART _ 10 det _ _
10 Natur Natur NOUN NN _ 8 nmod _ _
11 sind sein AUX VAFIN _ 8 cop _ SpaceAfter=No
12 . -- PUNCT $. _ 2 punct _ _
```

Figure 1 A sentence from the literary treebank in CoNLL-U format. The information concerning author, work and textual genre, as well as the univocal sentence ID, are highlighted in red.

Therefore, I aimed to safeguard the usual categories that are traditionally adopted to interact with the literary texts, especially author, work and genre. This is especially useful in a long-term perspective, when this resource could host new data for further analysis, from new genres or authors. This concerns each single sentence, since the textual units in a treebank are the sentence themselves. The treebank was therefore intended as a library of trees, in which all the above-mentioned categories denote each single tree. In a long-term perspective, this allow to retrieve data on specific works, authors or genres. The information concerning author, word and genre was encoded in the treebank file through machine-readable comments introduced by #, which precede each single sentence. Such comments are allowed

by the CoNLL-U format²¹. Precisely, some of these comments are mandatory, and they are used by all the UD treebanks, for instance, to encode information concerning the beginning of a new document (`# newdoc_id =`), to report the univocal sentence identifier (`# sent_id =`), or to report the linear version of the untokenized sentence (`# text =`). Conversely, other comments are optional. I exploited these ones to encode additional information concerning the author of the sentence, the work from which it comes from, and the textual genre to which it belongs. An example is shown in Figure 1, in which the additional comments are highlighted in red. Each time a new collection of Fragments (i.e. a work) begins, the first sentence of such collection is introduced by the comment `# newdoc id =`, followed by an ID that I chose for each collection, e.g. `lyceum` for *Lyceum Fragmente (Kritische Fragmente)*. In addition, when a new fragment (i.e. a text) in a given collection begins, it is signalled by the comment `# newpar id =`. For example, `# newpar id = lyceum-f1` means that the first fragment (f1) of the work *Lyceum Fragmente* begins, and from that on, all the successive sentences will be part of this fragment, until a new `# newpar id =` is reported. Then, three comments follows, bearing information about the genre, the work and the author, respectively (`# author =`, `# work =`, `# genre =`). In addition, I encoded the information about both the work and the fragment from which the sentence comes from in the sentence ID. The sentence ID (`sent_id`) univocally denotes each sentence in the treebank. In doing so, each syntactic tree is always mapped onto the position that the sentence has in the original work from which it comes from. In this way, the parallelism between the original text and the annotated one was maintained for each single sentence.²²

2.5 Processing the Fragments

2.5.1 Tokenization

The raw texts were tokenized through regular expressions in a text editor, i.e. each text was turned into a format where both each word and each punctuation mark are separated from the preceding and the following item, either a word or a punctuation mark, by a blank space. According to the UD principles, a word is intended from a lexicalist perspective, i.e. it is never split into morphemes, but considered as a syntactic (not orthographic neither phonological) unit (Nivre et al. 2016). I hereby report an example of a tokenized sentence from the source corpus:

²¹ Cf. <https://universaldependencies.org/format.html>.

²² Many thanks to Daniel Zeman at the Institute of Formal and Applied Linguistics (ÚFAL), Charles University, Prague, who helped me implement this solution.

Es ist eine unbesonnene und unbescheidne Anmaßung, aus der Philosophie etwas über die Kunst lernen zu wollen²³.

Some occurrences of the personal pronoun *es* in the elided form *‘s*. For all these cases, I considered *‘s* as a single token, such as in the following sentence:

Manche fangen *‘s* so an, als ob sie hofften hier etwas Neues zu erfahren²⁴;

2.5.2 Building a Test Set

I aimed to test different statistical-based Natural Language Processing (NLP) tools on a sample of Fragments, in order to find the best-performing ones on this genre. I therefore selected an initial test set of tokenized Fragments, and I encoded it in CoNLL-U format, thanks to the online tool UD pipe (Straka and Straková 2017). I manually annotated a copy of such test set to have a gold standard, against which I could evaluate the output of the tools. I annotated it manually according to the UD guidelines version 2. For an illustration of the linguistic annotation, see Chapter 3. In the test set, I included Fragments by all the authors represented in the source corpus, and I encoded it in CoNLL-U format. I assigned the ID *frag1_tok* to this test set. The test set is outlined in Table 5. When the tested tools required a different format, the initial test set was converted into the required format through a Python script.

Author	Work	ID	Tokens	Sentences
F. Schlegel et al.	<i>Athenauem Fragmente</i> (from 1 to 50)			
F. Schlegel	<i>Lyceum Fragmente</i> (from 1 to 90)	frag1_tok	7.737	381
Novalis	<i>Blüthenstaub</i> (from 1 to 20)			

Table 5 The initial test set of Fragments.

²³ F. Schlegel, *Kritische Fragmente*, fragment 123.

²⁴ F. Schlegel, *Kritische Fragmente*, fragment 123.

2.5.3 Best-Performing NLP Pipeline

The test set *frag1_tok* was used to test different statistical-based NLP tools, i.e. the output of each NLP tool was evaluated against the gold standard. These tests aimed to find the best-performing pipeline of NLP tools on the Fragments, which could help me extend the annotated data of the Fragments through a semi-automatic method. By pipeline of NLP tools, I mean a combination of different tools, each one performing a different task, but which work in series. By task, I mean the automatic assignment of linguistic metadata to each token in the input text. In a pipeline approach, such tools are used in sequence, which means that the output of one task is fed as input to the following task in the process. The overall pipeline adopted to perform the tests is illustrated in Figure 2.



Figure 2 NLP pipeline used to conduct the first tests the initial test set of Fragments.

The tests of different tools on specific tasks are reported and discussed in the next sections. I here only focus on the best-performing pipeline detected through these tests, which was the Mate Tools suite consisting of lemmatizer, POS-tagger, both based on Support Vector Machines (SVM), and a dependency graph-based syntactic parser (Bohnet 2010a), (Björkelund et al. 2010), in the Anna 3.6 implementation. This was the pipeline that I then adopted to process other Fragments from the source corpus, and to semi-automatically extend the treebank until the current status (UD 2.6). I here provide a general explanation of the methodology adopted to process the Fragments through this pipeline, and I provide the overall accuracy attained on the initial test by this pipeline for each task.

First of all, before feeding the lemmatizer with the raw text, *frag1_tok.txt* was brought into CoNLL-2009 format (Hajič et al. 2009), which is the required data format by the Anna 3.6 pipeline. I brought *frag1_tok.txt* into a one-sentence-per-line format²⁵ (*frag1_tok_ospl.txt*) through regular expressions. Then, I used the function *is2.util.split* inside the Anna 3.6 package²⁶, and I applied it to *frag1_tok_ospl.txt*. I therefore obtained a new file, named *frag1.conll09*. Then, this file in CoNLL-2009 format was lemmatized, POS-tagged and finally parsed. The POS-tagger was fed with the output of the lemmatizer,

²⁵ From now on, I will use the acronym OSPL to refer to the one-sentence-per-line data format, both in explanations and in file names.

²⁶ All the tasks were run through command lines in a shell in Windows environment. In this phase, all the tasks were performed by keeping Anna3.6 in its standard configuration.

i.e. *frag1_lem.conll09*, while the parser was fed with the output of the POS-tagger. As for POS-tagging, I fed the POS-tagger twice with the lemmatized text, the first time to assign UPOS, while the second time to assign XPOS, obtaining two distinct outputs, i.e. *frag1_uts.conll09* and *frag1_stts.conll09*, respectively. Then, I fed the parser with the two different POS-tagged files, in two different sessions. I made this choice since the CoNLL 2009 format does not have two distinct fields to host both UPOS and XPOS at the same time²⁷. Indeed, XPOS could have been included in the FEATS field of the CoNLL-2009 format, as SubPOS feature (i.e. after the tag *SubPos=*)²⁸, while UPOS could have been put in the same file in the POS field. But I discarded this option for three reasons. First, I did not want to train a model to process the target data that had metadata in FEATS field, because morphological features were not meant to be annotated in the first version of the literary treebank, as stated above. Second, adding POS-tag through a morphological analyser requires additional evaluations. In addition, the morphological analyser should not be compared with other POS-taggers, since it is engineered for a different task. Third, the accuracy by the syntactic parser may drop by adding morphological features to the training data, because training the parsing model is more complex. Since the accuracy of the syntactic annotation was crucial in this work, I preferred avoiding the risk of a degradation of the parsing performances. Therefore, I opted for POS-tagging the lemmatized texts in two separate sessions, and in both of them the POS-tag occupied the POS field of the CoNLL 2009 format. Two separate output were obtained, i.e. *frag1_par_uts.conll09* and *frag1_par_stts.conll09*. The Anna 3.6 graph-based parser attained higher accuracy when fed with STTS rather than with UTS. Moreover, the POS-tagging accuracy was also higher with STTS rather than with UTS (see 2.8). Therefore, I finally opted for using a simplified version of the pipeline previously illustrated in Figure 2. In this new pipeline, POS-tagging is performed with STTS only, with the POS-tag standing in the POS-field of the CoNLL-2009 format. The UPOS in UTS were then obtained automatically at the end of the annotation, through a conversion script. The ultimate NLP pipeline is illustrated in Figure 3.



Figure 3 NLP pipeline used for the semi-automatic development of the whole treebank.

As for the models used for processing the initial test set though the ultimate Mate Tool pipeline, they are summarized for each task in Table 12. I also report the input file, and the overall accuracy of each task.

²⁷ Cf. <https://ufal.mff.cuni.cz/conll2009-st/task-description.html>

²⁸ Cf. <https://ufal.mff.cuni.cz/conll2009-st/task-description.html>

I here provide a brief explanation of the models. As for lemmatization and POS-tagging with XPOS (STTS tag set), I used two pre-trained models supplied with the Mate Tools suite, which were trained on the full Tiger Corpus, as stated in the Mate Tools web page²⁹. On the contrary, the model used for syntactic parsing was trained on the whole training file of the UD 2.0 German treebank³⁰, which had been previously converted into CoNLL 2009 format thanks to a Python script. To build the parsing model, I removed all the morphological features from the training file, for the reasons explained above. All the models reported are reported in Table 6. These are the modes used for processing the whole treebank³¹.

Task	Model	Tool	Test Set	Accuracy %
Lemmatization	Pre-trained on <i>Tiger Corpus</i>		frag1.conll09	97.6
POS-tagging (STTS)	Pre-trained on <i>Tiger Corpus</i>	Mate Tools –	frag1_lem.conll09	97.3
Dependency Parsing	Trained on <i>de-ud2.0- train.conllu</i>	Anna 3.6	frag1_xpos.conll09	67.2 (LAS ³²)

Table 6 Models of the Mate Tool pipeline and their accuracy on the initial test set.

2.6 NLP-Tools Tests on German

Over the last years, evaluation of NLP tools and models on German has mainly focused on contemporary varieties, both in POS-tagging (Ivanova and Kübler 2008), (Giesbrecht and Evert 2009), (Rehbein 2013), (Horsmann, Erbs, and Zesch 2015), (Horsmann and Zesch 2016), and in dependency parsing (Kübler and Prokic 2006), (Kübler, Hinrichs, and Maier 2006), (Rafferty and Manning 2008), (McDonald and Nivre 2007), (Sennrich, Volk, and Schneider 2013), (Maier et al. 2014). Conversely, evaluations of NLP tools on historical varieties are much rarer in literature. As for POS-tagging, some tests conducted on pre-modern varieties of German are reported, such as on Middle High German (Dipper 2010), (Scheible

²⁹ Cf. <https://code.google.com/archive/p/mate-tools/>.

³⁰ I used this version of the UD German Treebank as training set, since it was the most recent one when these tests were performed.

³¹ The model used for the POS-tagging with the UTS is not reported. Cf. 2.8.

³² LAS stands for Labeled Attachment Score, i.e. the percentage of correctly assigned heads and deprel. For the accuracy metrics, see 2.7.

et al. 2011) or on Middle Low German (Schulz and Kuhn 2016), (Koleva et al. 2017). For a recent evaluation of different POS-taggers on different historical varieties, also see (Paluch et al. 2017). As for dependency parsing, some results on modern literary varieties are reported in (Salomoni 2017c). The tests that were here conducted and discussed are twofold. On the one hand, they primarily aim at detecting the best-performing pipeline on the target genre, i.e. Fragments, in order to speed up the treebank development through a semi-automatic approach. On the other hand, they provide empirical results concerning the processing of this literary variety of the Modern German, which has never undergone any NLP-tool test before. All the tools were trained on the available training data, which all belong to contemporary news or web varieties. Consequently, these tests also provide results concerning the portability on the models trained on the contemporary varieties, when tested on this specific historical variety.

2.7 Accuracy Metrics

I applied different metrics to evaluate the overall accuracy. As for lemmatization and POS-tagging, I evaluated the overall accuracy (a) by comparing the number of correctly assigned labels (n) to the whole number of labels assigned to the test set (t). In fact, this is the metric usually adopted to evaluate POS-tagging accuracy (Paroubek 2007).

$$a = n/t$$

For instance, let's assume that an input file consists of 100 tokens, it means that 100 POS-tags are expected. Let's assume the correctly assigned labels are 80, thus:

$$a = n/t = 80/100 = 0.8$$

As for dependency parsing, I adopted the metrics introduced by the CoNLL-X shared task (Buchholz and Marsi 2006b), (Jurafsky 2000):

- LAS, which stands for label attachment score. It measures the number of both correctly assigned heads and dependency relation
- UAS, which stands for unlabelled attachment score. It measures the number of correctly assigned heads only.
- LA, which stands for label attachment. It measures the number of correctly assigned dependency relations only.

In this work, LAS was the main metric I adopted to measure overall accuracy by the candidate parsing systems. Furthermore, I used an additional metric, i.e. the F-score, to evaluate some performances on

syntactic parsing. The F-score is the harmonic mean of precision and recall, which are two metrics which are usually adopted to measure the accuracy of information retrieval systems (Jurafsky 2000) (Van Rijsbergen 1974), (Buckland and Gey 1994). I hereby provide a general explanation of these metrics.

- Precision, which is the number of correctly assigned labels compared with the whole number of actually assigned labels by the automatic task.

$$\textit{Precision} = (\textit{true positives})/(\textit{true positives} + \textit{false positives})$$

- Recall, which is the percentage of correctly assigned metadata compared with the whole number of metadata that should have been assigned by the automatic task.

$$\textit{Recall} = (\textit{true positives})/(\textit{true positives} + \textit{true negatives})$$

- F-Score, which is the harmonic mean of precision and recall.

$$\textit{F - score} = (2 \times \textit{precision} \times \textit{recall})/(\textit{precision} + \textit{recall})$$

2.8 Tests on Lemmatization

As for lemmatization, I considered a single candidate tool only, which is the lemmatizer from the Mate Tools suite in the Anna 3.6 implementation (Bohnet et al. 2016). In particular, it is a Support Vector Machine (SVM) lemmatizer (Cortes and Vapnik 1995) implementing the Margin-infused relaxed algorithm (MIRA) algorithm (Crammer and Singer 2001). I ran the tool in its standard configuration. As for the model, I first tested a pre-trained model on the whole Tiger Corpus. I evaluated the output against the manually annotated gold standard. Experimental design and overall accuracy are reported in Table 7.

Tool	Training Set	Test Set	Accuracy %
Mate Tools Lemmatizer Anna 3.6	Pre-trained model (Tiger Corpus) ³³	frag1.conll09	97.0

Table 7 Experimental design and overall accuracy of the candidate lemmatizer.

As shown in Table 7, the candidate tool attained a very high overall accuracy on the test set. Therefore, I opted for selecting this tool as reference tool for lemmatization without conducting any other test. Nevertheless, it is worth highlighting some problems concerning some mistakenly assigned lemmas in the output. Table 8 reports some of the errors returned by the lemmatization. ‘Assigned lemma’ stands for the lemma that was automatically assigned by the lemmatizer. ‘Expected lemma’ stands for the lemma that was expected³⁴. ‘AFT’ stands for the absolute frequency of the wrongly assigned lemma in the output. ‘RFT’ stands for relative frequency of the wrongly assigned lemma in the test set. I reported this value to help clarify the impact of the mistakenly assigned lemmas with respect to the whole bulk of data in the test set. Finally, ‘AFTR.’ stands for absolute frequency in the training set, i.e. the Tiger Corpus. I reported this value to show whether the error in lemmatization was caused by out-of-vocabulary (OOV) words.

Word	Assigned Lemma	Expected Lemma	AFT	RFT%	AFTR
seyen sey	seyen sey	sein	11	0.15	0
Gute	gute	gut	3	-	8
andre	andre	ander	6	0.08	0
seltner	seltner	selten	4	0.05	0
Unbedingten	Unbedingter	unbedingt	2	-	0
kömmt	kömmt	kommen	3	-	0
Kürzeste	Kürzeste	kurz	1	-	0
öfter	öfter	oft	1	-	3

Table 8 Mistakenly assigned lemmas in the test set of Fragments by the Anna 3.6 lemmatizer.

³³ According to what declared on the Mate Tools web page:, all the provided pre-trained models were trained on the whole Tiger Corpus., which consists in 938.709 tokens (aug07 release): <https://code.google.com/archive/p/mate-tools/>, last access on November 23rd September 2018.

³⁴ That is, according to the features of the training set and to the current German spelling.

First of all, the diphthong *ie* was often replaced by *ye* in the late-18th-century texts. In the target texts, it occurs e.g. in the non-finite form of the verb ‘sein’, which was spelled ‘seyen’. To each form of ‘seyen’ in the test set, the lemmatizer often assigned the lemma with the current spelling ‘seien’. Obviously, this is due to the absence of the spelling with ‘y’ in the training data, as shown by the absolute frequency in the training set (AFT_{TR}). The same happens for other word forms, such as the verb ‘freyen’ (AFT = 2, where AFT stands for absolute frequency in the test set), or the noun ‘Bewußtseyn’ (AFT = 1). To cope with this issue, a common approach would be to normalize the target data in order to reduce the gap with the training set, see e.g. (Dipper 2010). In this case, I should pre-process all the target data that are supposed to be processed and hosted in the treebank, by replacing all the diphthongs ‘ye’ with the current spelling ‘i’. But I opted differently, since normalization at tokens level cause a loss of linguistic information. In fact, by normalizing, I would miss observations that univocally mark the variety of data hosted in the treebank. Therefore, I decided to preserve the original spelling at tokens level, while I opted for normalizing the lemma of the corresponding word forms in post-processing, through regular expressions. By contrast, I opted differently for another quite frequent spelling variant, which is ‘ß’. In fact, some words are spelled with ‘ß’ in target data, such as the subordinating conjunction ‘daß’ (AFT = 19), or some forms of the modal verb ‘müssen’ (AFT = 6). In the current spelling, they are usually spelled with ‘ss’. In this case, I opted for preserving the original spelling with ‘ß’ both in word form and in lemma, since such spelling variation is still allowed in German dictionaries. Moreover, it is used in lemmas in Tiger Corpus as well, which is the source variety for these experiments in the lemmatization.

Other common mistakes in lemmatization concerns de-adjectival nouns, which can be really common in philosophical or literary lexicon, such as the female noun ‘Gute’ from the adjective ‘gut’ (beautiful), or the neutral substantive ‘Unbedingten’ from the adjective ‘unbedingt’ (absolute). The same happens for the comparative form in the substantive ‘Kürzeste’. In general, these words have been lemmatized as nouns, keeping the word form unchanged, while they should have been lemmatized as adjectives. For most of them, errors are due to pure OOV reasons, since such forms have a.f.tr. = 0. Looking at Table 2, ‘Gute’ looks like an exception. But in this case the value of a.f.tr. turns out to be tricky. In fact, by querying Tiger Corpus, occurrences of ‘Gute’ are mostly adjectives lying at the beginning of a sentence, such as ‘Gute Empfindung’. Just one occurrence out of ten is a deadjectival noun ‘die Gute’ indeed. Therefore, it seems that a single occurrence in training data is not enough to train an accurate model to predict the deadjectival use of ‘gut’ in unseen data. Apart from the manual checking of the gold standard of the test set ‘frag1_lem_conll09.txt’, forms involved in this error pattern have not been manually corrected in the treebank, once annotated data were extended.

Lemmas have been wrongly assigned in case of some altered adverbs and adjectives as well, such as ‘seltner’, which is the comparative of ‘selten’ (rare), or ‘öfter’, which is the comparative of ‘oft’ (often). In this first case, it is a matter of OOV (a.f.tr. = 0), therefore the word has been reproduced in the lemma as it occurs in the input texts. In the case of ‘öfter’, there are three occurrences of this form in Tiger Corpus, but just one of them is lemmatized correctly, while the other two show ‘öfter’ in lemma field. These errors were not checked in the treebank after the extension of the annotated data either.

2.9 Tests on POS-Tagging

As for POS-tagging, two candidate tools were tested on the Fragments, and two different tag sets were tested as well: the Universal Tag Set (UTS), (Petrov, Das, and McDonald 2012), and the *Stuttgart Tuebingen Tag Set* (STTS), (Schiller, Teufel, and Thielen 1995). For a comparison between the tag sets, see Chapter 3. The accuracy attained by the candidate tools on the target variety is compared with that by the same tools on a test set from the source variety. Results are illustrated and discussed.

2.9.1 Universal Tagset (UTS)

First, I tested two candidate tools on the Fragments with the UTS: Anna 3.6 POS-tagger from the suite Mate Tools and the POS-tagger UD Pipe 1.1 (Straka and Straková 2017). Anna 3.6, is a SVM tagger implementing MIRA algorithm, as the lemmatizer. The tagger in the UD Pipe 1.1 suite is the MorphoDiTa tagger (Straková, Straka, and Hajič 2014). It works in two main phases. First, the morphological dictionary suggests all possible lemma-tag candidates for each form in the text, then these lemma-tag pairs are disambiguated by the tagger. The tagger is implemented as supervised, rich feature averaged perceptron (Collins 2002), (Straková, Straka, and Hajič 2014). Both taggers were run in their standard configuration, in order to attain the baseline accuracy first. In this configuration, training epochs are automatically set on 9, and no additional parameter is specified. The taggers were fed with the lemmatized test set as input. Lemmatization was performed by the Anna 3.6 lemmatizer. As for the model, both POS-taggers were trained on the training file of the UD German Treebank 2.0, i.e. ‘de-ud2.0-train.conllu’, after converting it into CoNLL 2009 format. At the end of the experiment, Anna 3.6 attained the highest accuracy on the target data. Experimental design and overall accuracy are reported in Table 9.

Training Set	Test Set	Tool	Accuracy% (baseline)
de-ud2.0-train	frag1_lem_conll09	Anna 3.6 (Mate Tools)	90.0
		MorphoDiTa (UD Pipe 1.1)	88.5

Table 9 Overall Accuracy by two candidate POS-taggers on the test set of Fragments.

After this first experiment, I tested the Anna 3.6 POS-tagger on a test set of the source variety too, i.e. web texts from the UD German GSD 2.0. I maintained the tool in the same standard configuration used to for the test on the target variety. I performed a ten-fold cross-validation. First, I split the de-ud2.0-train.conllu file as follows: 10% for testing and 90% for training. Second, I trained a model on the 90% and I tested it on the 10%. I repeated the experiment 10 times, varying each time the 90% and the 10%, and I obtaining ten different outputs. Finally, I measured the overall accuracy by Anna 3.6 on the source variety by doing an arithmetic average of the ten measures. Results are reported in Table 10, while the accuracy gap between the source variety and the target variety is highlighted in Table 11.

Training Set	Test Set	Tool	Accuracy %
90% de-ud-train2.0	10% de-ud-train2.0	Mate Tools Anna 3.6	93.6

Table 10 Overall average accuracy attained by the Anna 3.6 POS-tagger on a test set from UD German GSD 2.0.

Tool	Accuracy % on Source Variety	Accuracy % on Target Variety	Accuracy Gap %
Mate Tools Anna 3.6	93.6	90.0	3.6

Table 11 Accuracy gap shown by the Anna 3.6 POS-tagger between the target variety and the source variety.

Usually, the coarser the set of labels to be assigned in POS-tagging, the higher the portability of the models trained on these labels, see e.g. (Maier et al. 2014), (Horsmann and Zesch 2016). Since I tested a coarse-grained tag set, i.e. UTS, this would explain the rather low accuracy gap shown by Anna 3.6 across the source and the target variety (3.6%). In any case, I performed an in-depth evaluation of the results on both the data sets, in order to detect the most problematic lexical classes. The results are sorted in two tables: the first one reports the results concerning the open lexical classes (Table 12), while the second one those concerning the closed classes (Table 13).

Tool	UPOS	Accuracy % on Source Variety	Accuracy % on Target Variety
Mate Tools Anna 3.6	NOUN	93.7	94.2
	PROPN	84.0	83.6
Mate Tools Anna 3.6	VERB	95.1	93.5
	ADV	90.7	83.0
	ADJ	91.2	94.0

Table 12 Accuracy attained by the Anna 3.6 POS-tagger on single UPOS (open classes), (UTS).

Tool	UPOS	Accuracy %	Accuracy %
		on Source Variety	on Target Variety
Anna 3.6 POS-tagger (Mate Tools)	AUX	83.9	38.9
	CCONJ	96.1	95.5
	SCONJ	89.1	79.0
	ADP	94.9	88.3
	DET	96.3	98.9
	PRON	93.2	93.0

Table 13 Accuracy attained by the Anna 3.6 POS-tagger on single UPOS (closed classes), (UTS).

As shown in Table 12, the most problematic open classes are proper nouns and adverbs. Proper nouns are notoriously problematic for POS-tagging, since the tokens of this class are often out-of-vocabulary (OOV) tokens. Therefore, instances of pairs of data and metadata can be very scarce in the training set for this class. In particular, all the proper nouns that were mistakenly POS-tagged were assigned the POS ‘NOUN’. Adverbs also showed a low accuracy with respect to the average accuracy attained on other parts of speech. In this case, many adverbs were assigned the POS ‘ADJ’. This is due to one of the morphological properties of the German adverbs, which are mostly obtained from adjectives through conversion (or zero-derivation), as shown in (1) and (2).

- 1) Das Bier ist gut. (The beer is good)
Er spielt gut. (He plays well)
- 2) Er bewegt sich ständig (he continuously move)
Der Verfahren ist ständig (the process is continuous)

Furthermore, both adverbs and adjectives are usually assigned the same lemma, which is used as input feature by the POS-tagger to assign the POS-tag. Therefore, the information brought by lemmas cannot help the model disambiguate the POS. These could be the reason behind the error rate for the POS ‘ADV’.

As for the closed classes, auxiliaries attained a really low accuracy. In this case, two factors could have played a role. Firstly, the different spelling. In fact, as highlighted in lemmatization, the auxiliary verb ‘sein’ always occurs as ‘seyn’ in non-finite form. Since there is no occurrence of this word form in the training data, this auxiliary is a OOV token. But this cannot be the sole reason, since AF of the form

‘seyen’ is rather low in the test set. In fact, 12.5% of auxiliaries in the test set of target data were mistakenly assigned ‘VERB’ as UPOS. Consequently, I thought that such an error rate could be caused by annotation error in the training set. I therefore queried the UD 2.0 German treebank through the SETS platform ³⁵. Indeed, I found out that there were noticeable annotation mistakes concerning the annotation of copula verbs in this version of the treebank. In fact, 3,293 occurrences of *sein* playing the role of copula in nominal predicates were assigned the UPOS ‘VERB’ rather than ‘AUX’. This can have caused to accuracy to drop on auxiliaries.

Finally, subordinating conjunctions also showed a lower accuracy (79) with respect to other elements of this lexical class: 19.2% of the wrongly POS-tagged subordinating conjunctions were assigned UPOS ‘PRON’. Therefore, it is likely that the system confused them with relative pronouns.

2.9.2 Stuttgart-Tübingen Tagset (STTS)

As for the tests with the STTS, I first had to face a problem concerning the selection of the training set. In fact, in the UD German GSD treebank 2.0, STTS stands in the XPOS field to encode the fine-grained POSes, while the UTS stands in the UPOS field to encode the coarse-grained POSes. Therefore, I could have used this treebank as training set to train the candidate tools on the STTS. But the XPOS had been automatically assigned in this version of the treebank, without any manual checking. Therefore, I could not consider this data as gold standard to train a model (since I am applying a supervised methodology). Conversely, the STTS was implemented in Tiger Corpus after manual revision (Brants et al. 2002). At the same time, some POS-tagging models pre-trained on the Tiger Corpus were available. Since I aimed to test pre-trained models first, I chose two different candidate POS-taggers for which pre-trained models were available: the Mate Tools Anna 3.6 POS-tagger and the Stanford Tagger (Toutanova and Manning 2000). Both POS-taggers were run in their standard configuration, using the provided pre-trained models, respectively. The experimental design and the overall baseline accuracy are reported in Table 23.

³⁵ http://bionlp-www.utu.fi/dep_search/, last access on 28rd September 2019.

Training Set	Test Set	Tool	Accuracy %
Tiger Corpus (Pre-Trained)	frag1_lem.conll09	Anna 3.6	97.3
		(Mate Tools) Stanford 3.7.0	92.9

Table 14 Overall accuracy attained by two different POS-taggers (STTS) on the test set of Fragments.

As shown in Table 14, Anna 3.6 outperformed Stanford Tagger by 4.6% on the target variety. Consequently, as done for the other previously-tested tools, I also tested it on a test set of the source variety through a 10-fold validation. I repeated the same process described above for testing the same POS-tagger on a test set from the UD German GSD treebank 2.0. In this case, though, the source variety was the Tiger Corpus, since the pre-trained model I used tested on the Fragments had been trained on this corpus. I thus performed a 10-fold-validation on this corpus. The average accuracy attained by Anna 3.6 on the Tiger Corpus is reported in Table 15, while the accuracy gap shown by Anna 3.6 in assigning the STTS to the source variety and the target variety is highlighted in Table 16.

Training Set	Test Set	Tool	Accuracy %
90% Tiger Corpus	10% Tiger Corpus	Mate Tools Anna 3.6	97.6

Table 15 Overall average accuracy attained by the Anna 3.6 POS-tagger (STTS) on a test set of the source variety.

As shown in Table 16, the accuracy gap between the two varieties tagged with the STTS is very low (0.3%). Moreover, it is remarkably lower than the gap shown by the same tool in assigning the UPOS to the two varieties (3.6 %). Furthermore, the Anna 3.6 POS-tagger attained a considerably higher accuracy on the Fragments by implementing the STTS rather than the UTS, as shown in Table 17.

Tool	Accuracy % on Source Variety	Accuracy % on Target Variety	Accuracy Gap %
Mate Tools Anna 3.6	97.6	97.3	0.3

Table 16 Accuracy gap shown by the Anna 3.6 POS-tagger between the source variety and the target variety.

Tool	Accuracy % UTS	Accuracy % STTS	Accuracy Gap %
Mate Tools Anna 3.6	90.0	97.3	7.3

Table 17 Accuracy gap shown by the Anna 3.6 on the test set of Fragments with two different tag sets.

Apparently, this result clashes with what stated above about the relation between the tag-set granularity and the POS-tagging accuracy. In general, accuracy is expected to be higher when using a coarse-grained tag set with respect to a fine-grained tag set. Intuitively, I thought that this result could be influenced by the different size of the two training sets. In fact, the size of the Tiger Corpus is more than 3 times the size of the UD German GSD 2.0, as highlighted in Table 18. To test this hypothesis, I tested the Anna 3.6 POS-tagger on the test set of Fragments again, but I varied the size of the training set from the Tiger Corpus. I opted for reducing the training set to 283,743 tokens, i.e. the same size of the training file of the UD German GSD 2.0 that I had used to train the model for the UTS. Results are reported in Table 19.

Corpus	Size (tokens)
German UD Treebank 2.0	283.743
Tiger Corpus 2.2	938.709

Table 18 Size of the training sets used to POS-tag the test set of Fragments.

Training Set	Test Set	Tool	Accuracy %
Tiger Corpus (283.743 tokens)	frag1_lem_conll09	Mate Tools Anna 3.6	94.1

Table 19 Overall accuracy attained by the Anna 3.6 POS-tagger on the test set of Fragments (STTS) after changing the training-set size.

As shown in Table 19, the accuracy by the Anna 3.6 POS-tagger with STTS on the target data decreases by 2.9% by reducing the size of the training set to the same size of the UD training set. Nevertheless, the gap with respect to the results attained by the same tool on the same test set with UTS is still considerable (+4.1% with STTS). Therefore, the difference in the sizes of the training sets has a role in the final results, but it is not the sole reason. The different accuracy shown by the two tag sets could be attributed to the tag-set design.³⁶ It seems that the STTS granularity let the tagger disambiguate better the input token by looking at the context. For a more detailed overview of the performance by the Anna 3.6 POS-tagger on

³⁶ The different accuracy could be due to a poor quality in the annotation of the UD training set as well.

the target test set with the pre-trained model and the STTS, I provide the scores on single POS. They are reported in Table 20 (open classes) and Table 21 (closed classes).

Tool	XPOS	Accuracy %	Accuracy %
		on Source Variety	on Target Variety
Anna 3.6 (Mate Tools)	VVFIN	93.3	94.5
	VVINP	93.4	96.1
	VVPP	95.8	96.0
	VVIZU	93.0	100
	ADJA	98.3	97.7
	ADJD	94.0	95.5
	ADV	97.2	88.4
	NN	98.7	99.2
NE	92.1	95.5	

Table 20 Accuracy attained by the Anna 3.6 POS- with the STTS, on both the source and the target variety (open classes).

Tool	XPOS	Accuracy %	Accuracy %
		on Source Variety	on Target Variety
Anna 3.6 (Mate Tools)	VMFIN	98.6	100
	VMINP	75	88
	VAFIN	98.4	100
	VAINP	94	94.5
	KON	-	100
	KOUS	97.7	100
	KOKOM	-	96.5

Table 21 Accuracy attained by the Anna 3.6 POS-tagger with the STTS, on both the source and the target variety (closed classes).

As shown in Table 29, the problems concerning both proper nouns (NE) and adverbs still remain at a certain extent, since the POS tagger faces the same issues described above. However, the overall accuracy on these classes remarkably increases with respect to that attained with the UTS. Furthermore, scores on the those tags concerning verbs are remarkably high in general. As for auxiliaries, the problems showed with the UTS seems to be solved with STTS, since the accuracy on both finite and non-finite forms is very high, attaining a 100% accuracy on the finite forms. Among the verbal POS-tags, the accuracy decreases on non-finite modal verbs only. Among conjunctions, accuracy attains 100% on both coordinating conjunctions (KON) and subordinating conjunctions (KOUJ), and it is high on comparative conjunctions (KOKOM) as well. In this last case, polysemy could have caused the tagger's accuracy to decrease. In fact, there are two lexical items playing the role of comparative conjunctions in German: *wie* and *als*, as shown in examples 1) and 2). These words often occur with different syntactic functions as well. Most of the times, they play the role of subordinate markers, as shown in 3) and 4) for *wie*, and in 5) for *als*.

- 1) Sie ist so schön *wie* ihre Freundin.
- 2) Sie ist schöner *als* ihre Freundin.
- 3) *Wie* gehet es dir?
- 4) Das Photo zeigt, *wie* sie sich verändert sind.
- 5) Als ich ein Kind war, es gab keine Laptops.

Given the remarkable accuracy gap between UTS and STTS, I chose the STTS as the reference tag set for POS-tagging. Therefore, I later used this tag set to process other Fragments from the source corpus during the treebank development. The STTS was assigned as POS tag in the CoNLL2009 format.³⁷ At the same time, UPOS are necessarily required by the UD guidelines, therefore they cannot be omitted. In fact, I opted for adding UPOS later in the treebank production, once all XPOS will have been automatically assigned. Since the accuracy between the two tag sets is significant, I chose to automatically obtain UPOS from XPOS, rather than assigning automatically UPOS through POS-tagging. Such an approach has already been applied to build other UD treebanks.

³⁷ Even if it encodes fine-grained POS-tags, I did not use it in the field FEATS, which usually hosts fine grained POS-tags that occur after the tag 'SubPOS=?'.

2.10 Tests on Dependency Parsing

In this section, I will describe experiments concerning syntactic dependency parsing. Notoriously, dependency parsing is a more complex task with respect to lemmatization and POS-tagging. Therefore, I opted for testing more candidate tools. In particular, I tested four parsers based on different parsing systems. Three of them implements the transition-based parsing, while one of them implements the graph-based parsing. For an overview on these systems, see (Jurafsky and Martin 2014). In two cases, these systems are integrated with other approaches that are emerged over the last few years in the NLP research. In the Joint Parser, the transition-based parsing is integrated with the beam search and an integrated POS-tagger. Whereas, in the *Parsito* parser (UD pipe 1.1), the transition-based parsing is integrated with a neural-network classifier. Candidate tools and their parsing systems are highlighted in Table 22.

Tool	System
Malt Parser 1.9.0 (Nivre et al. 2008)	Transition-Based
Mate Tools Anna 3.6 (Bohnet 2010)	Graph-Based
Joint Parser 1.30 (Bohnet and Nivre, n.d.)	Transition-Based + Beam Search + Integrated POS-Tagger
UD Pipe 1.1 - Parsito (Straka and Straková 2017)	Transition-Based + Neural Network Classifier

Table 22 Candidate Dependency Parsers.

After removing all the morphological features³⁸, all the parsers were trained on the whole training file of the UD German GSD treebank 2.0. A copy of the file of the UD German Treebank 2.0 was converted into CoNLL 2009 format, since some candidate parsers require the files to be in this format. Unlike the CoNLL-U format, the CoNLL 2009³⁹ format does not have any specific field for fine-grained POS-tags, because they are usually hosted as first item in FEATS field, preceded by the tag ‘SubPOS=’. Therefore, the XPOS was automatically put in the converted treebank file as ‘SubPOS=XPOS’. I applied the same solution to the POS-tagged test set of Fragments (“frag1”), before feeding it into the parser. In this case,

³⁸ I opted for this solution since morphological features were not annotated in the treebank, thus I wanted to leave them out from the parsing model. Furthermore, morphological features can even cause parsing accuracy to drop, since models trained on them are more complex.

³⁹ Cf. <https://ufal.mff.cuni.cz/conll2009-st/task-description.html>.

XPOS was assigned by the reference POS-tagger through the pre-trained model first, and it was then automatically brought into FEATS field.

As for POS-tagging, Anna 3.6 was fed with the automatically lemmatized input, and it was run in two separate sessions, one for UPOS and one for XPOS. Then, the automatically assigned XPOS and UPOS were merged together in one file, as described above for the UD training file. This file was the input file that I fed into all the parsers. For those parsers requiring files in CoNLL-U format, the input file was converted into this format through a python script. The pipeline adopted to conduct these first parsing experiment is highlighted in Figure 5. As for the configuration, each tool was set and run as described below. The overall parsing accuracy attained by each candidate tool on the test set of Fragments is reported in Table 26.



Figure 4 Pipeline adopted for the first test of the candidate parsers on the Fragments.

- **Malt Parser 1.9.0.** Initially, it was run in its standard configuration, i.e. any optional parameter was added in the command line. The initial baseline accuracy (LAS) is reported in Table 23. Malt Parser should enhance parsing accuracy after an automatic optimization through Malt Optimizer (Ballesteros and Nivre, 2012). This system is designed to automatically analyse the training data and suggest the best configuration for Malt Parser to parse similar data. In this case, since training data and target data belongs to two different varieties, I could not directly run the optimization to parse the target data without testing the tool on them in its standard configuration first. In fact, setting parser’s parameters on training data can cause overfitting problems. Therefore, I could not take the increase in accuracy on the target data for granted, since the optimization is calculated on UD, while the optimized configuration has to be run on the Fragments. That is the reason why I ran the tool in its standard configuration first, as described above. Then, I ran Malt Optimizer feeding it with the whole training file of the UD German treebank 2.0. Then I set Malt Parser with the suggested configuration, which is reported in Table 24, and I ran it on the test set from the target variety again. In fact, accuracy slightly increased after the optimization. This result is reported in Table 25 (the same results is reported in Table 26 as well).

Training Set	Test Set	Tool	LAS % (baseline)
de-ud2.0_train	frag1	Malt Parser 1.9.0	61.3

Table 23 Accuracy by Malt Parser 1.9.0 run in its standard configuration on the test set from the target variety.

Parameter	Configuration
Feature Model	addMergPOSTAGI0FORMLookahead0
Algorithm	stackproj

Table 24 Malt Parser configuration set after the automatic optimization through Malt Optimizer.

Training Set	Test Set	Tool	LAS % (optimization)
de-ud2.0_train	frag1	Malt Parser 1.9.0 + Malt Optimizer	63.4

Table 25 Accuracy by Malt Parser 1.9.0 run after an optimization with Malt Optimizer on the test set of Fragments.

- **Mate Tools Anna 3.6.** It was run in its standard configuration. Training iterations were set at 9 and projectivity threshold was set at 0.3, which is optimized for German.
- **Joint Parser 1.30.** It was run in its standard configuration. Training iterations were set at 10, while beam search was set at 40, as suggested for German (Bohnet and Nivre 2012). In this tool, POS-tagging is jointly performed with parsing. A model for POS-tagging was therefore automatically built during the training phase, and then tested in testing phase.

- **UD Pipe 1.1 – Parsito.** It was run in its standard configuration, without specifying any additional parameter.

Training Set	Test Set	Tool	Average Accuracy % (baseline, LAS)
de-ud2.0_train	frag1	Malt Parser 1.9.0	63.4
		Mate Tools Anna 3.6	66.4
		Joint Parser 1.30	64.2
		Ud Pipe 1.1 Parsito	60.6

Table 26 Overall accuracy by candidate dependency parsers on a test set of the target variety.

The graph-based parser Anna 3.6 (Mate Tools) turned out to be the best system on the target data, attaining the highest baseline accuracy. I opted for not running other tests trying a different set up for each candidate parser. Conversely, I focused on the best-performing one only, which I considered as reference tool for syntactic annotation. As I did for the previous tasks, I tested the baseline accuracy by the reference tool on the source variety too, in order to compare the accuracy attained on the target variety with that attained on the source variety. Therefore, I run a ten-fold cross validation on the UD German Treebank 2.0 training file. The average baseline accuracy (LAS) is reported in Table 27, while the accuracy gap between the two varieties is highlighted in Table 28.

Training Set	Test Set	Tool	LAS %
90% de-ud2.0-train	10% de-ud2.0-train	Mate Tools Anna 3.6	84.6

Table 27 Accuracy by the graph-based parser Anna 3.6 on a test set from the source variety.

Tool	LAS % on Source Variety	LAS % on Target Variety	Accuracy Gap %
Mate Tools Anna 3.6	84.6	66.4	18.2

Table 28 Accuracy gap shown by the graph-based parser Anna 3.6 between the source variety and the target variety.

As shown in Table 37, the gap between the contemporary variety (UD) and the Fragments is significant (18.2%). I therefore conducted an in-depth evaluation on the output by Anna 3.6, in order to highlight what causes the accuracy to drop on Fragments.

2.10.1 In-Depth Evaluation

To set this compared analysis, I build a UD test set, which I named *control set*. I assigned it the ID ‘UD_cont’. This test set has the same size of the test set of Fragments, i.e. 7,292 tokens. The test set was randomly taken from the UD 2.0 development set, and it was processed through the Anna 3.6 pipeline, applying the same methodology adopted to process the Fragments. For each task, i.e. lemmatization, POS-tagging and graph-based dependency parsing, I used the same models applied to process the Fragments. After performing lemmatization and POS-tagging, the output of the POS-tagger was fed into the parser. The overall parsing accuracy on this data set is shown in Table 29. The overall LAS attained on this UD data set was very similar to the average LAS showed by the same parser in the ten-fold validation on UD. Therefore, this made the ‘UD_cont’ test set suitable for a compared error analysis of parsing.

Tool	Training set	Test set	LAS % (baseline)
Anna 3.6	de-ud2.0	UD_cont	86.5
graph-based parser		(7292 tokens)	

Table 29 Overall LAS attained by the anna 3.6 graph-based parser on the control data set from UD 2.0.

I evaluated the parsing accuracy on single dependency relations. F-score concerning most of the dependency relations on Fragments is reported in Chart 1, where RF of each relation in the data set is

reported as well. The syntactic relations are ranked by decreasing RF in Chart 1. This illustration has a twofold purpose. On the one hand, it shows how each dependency relation affects the overall parsing accuracy. On the other hand, it can help draw a connection, if any, between the distribution of dependency relations and their variation of accuracy. For instance, it allows to check whether the most frequent relations in the data set are those attaining the highest parsing accuracy.

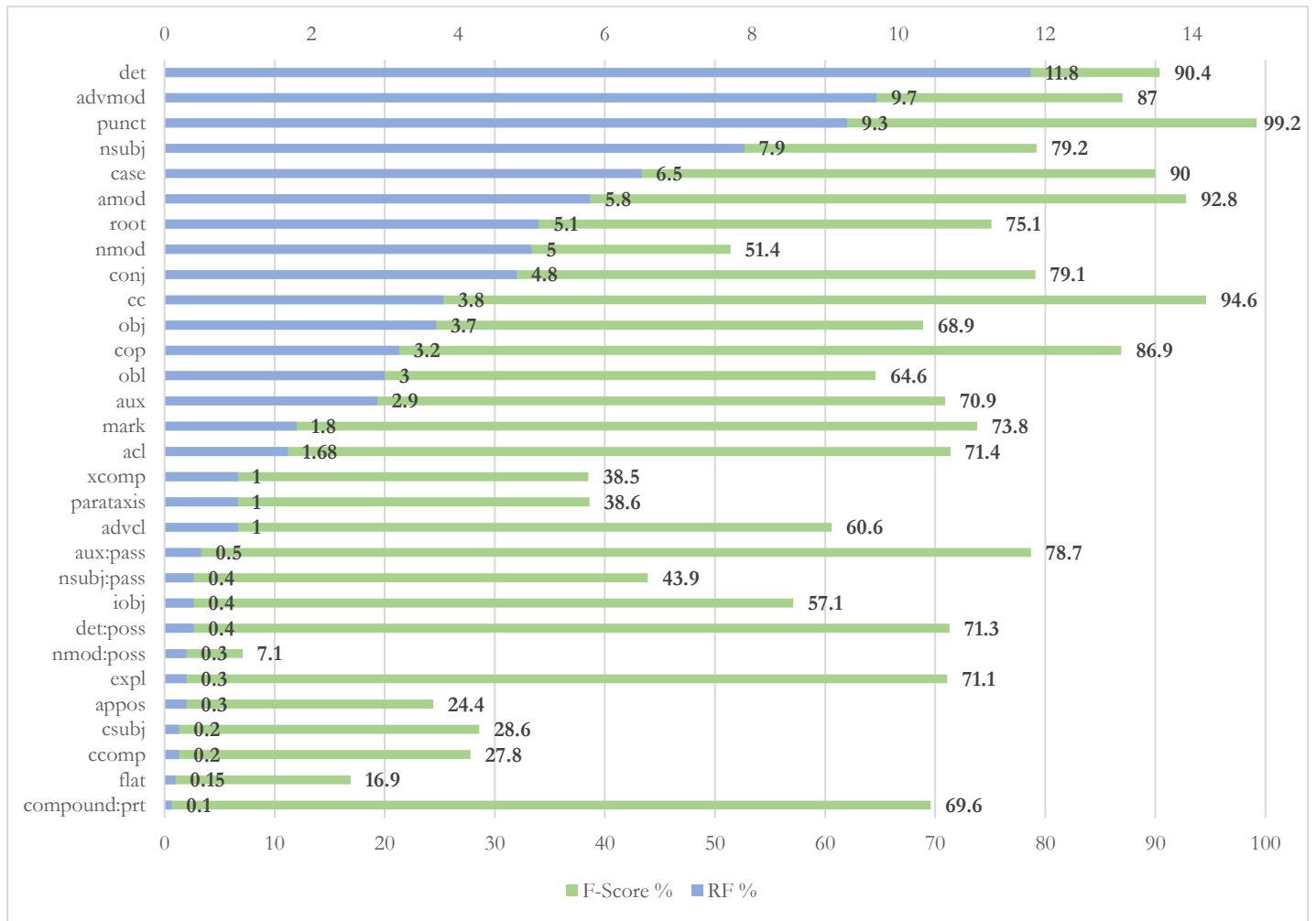


Chart 1 F-score and relative frequency (RF) of each single dependency relation in the test set of Fragments. RF lies on the axis above, while F-score lies on the axis beneath. Relations are sorted per decreasing RF.

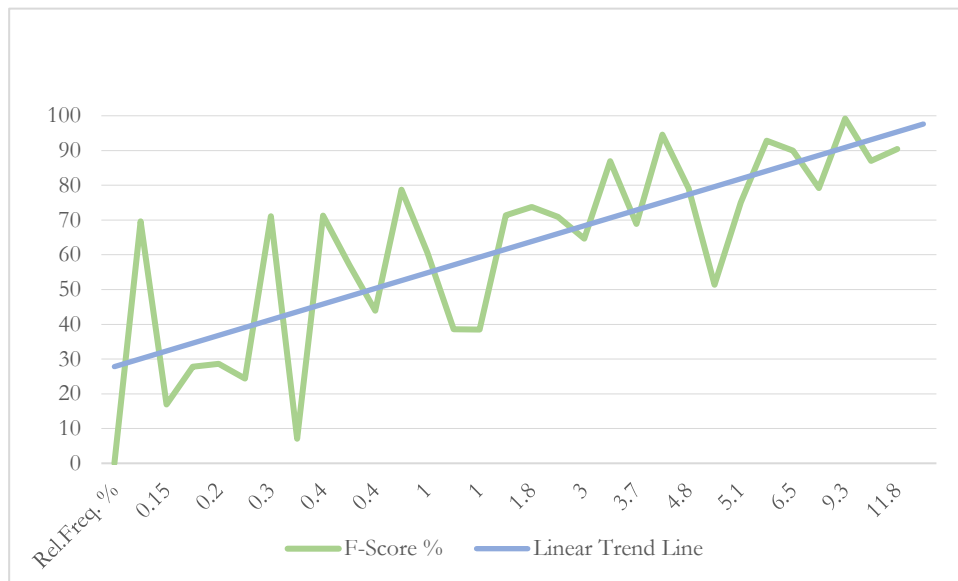


Chart 2 F-score by Anna 3.6 graph-based parser on the test set of Fragments, according to the variation in RF by each dependency relation. Accuracy lies on the y axis, while RF lies on the x axis.

As shown in Chart 1, the accuracy on some relations clearly diverges from the average LAS (66.4%). Considering the relation between RF and F-Score, it seems that the system attained a good accuracy, on average, on those relations occupying the highest positions in the RF ranking. On the contrary, despite isolated peaks, such as that regarding the relation *nmod*, it seems that accuracy tends to drop on average on those relations lying in the lowest part of the RF ranking. To better evaluate the relation between RF and accuracy, I reported the relation between RF and F-score in Chart 2, where the variation in accuracy is shown according to the variation in RF. The linear trendline seems to confirm this hypothesis about the relationship between parsing accuracy and the distribution of the dependency relation. In fact, on average, the higher the frequency of a dependency relation in the test set, the higher the parsing accuracy. Furthermore, such an assumption seems to be corroborated by the results displayed in Chart 3, in which all the dependency relations that were reported in Chart 1 are sorted per decreasing F-score. As shown, the highest positions in the accuracy ranking are occupied by those relations with a high RF, while RF tends to decrease on average on those relations lying lower in the F-score ranking.

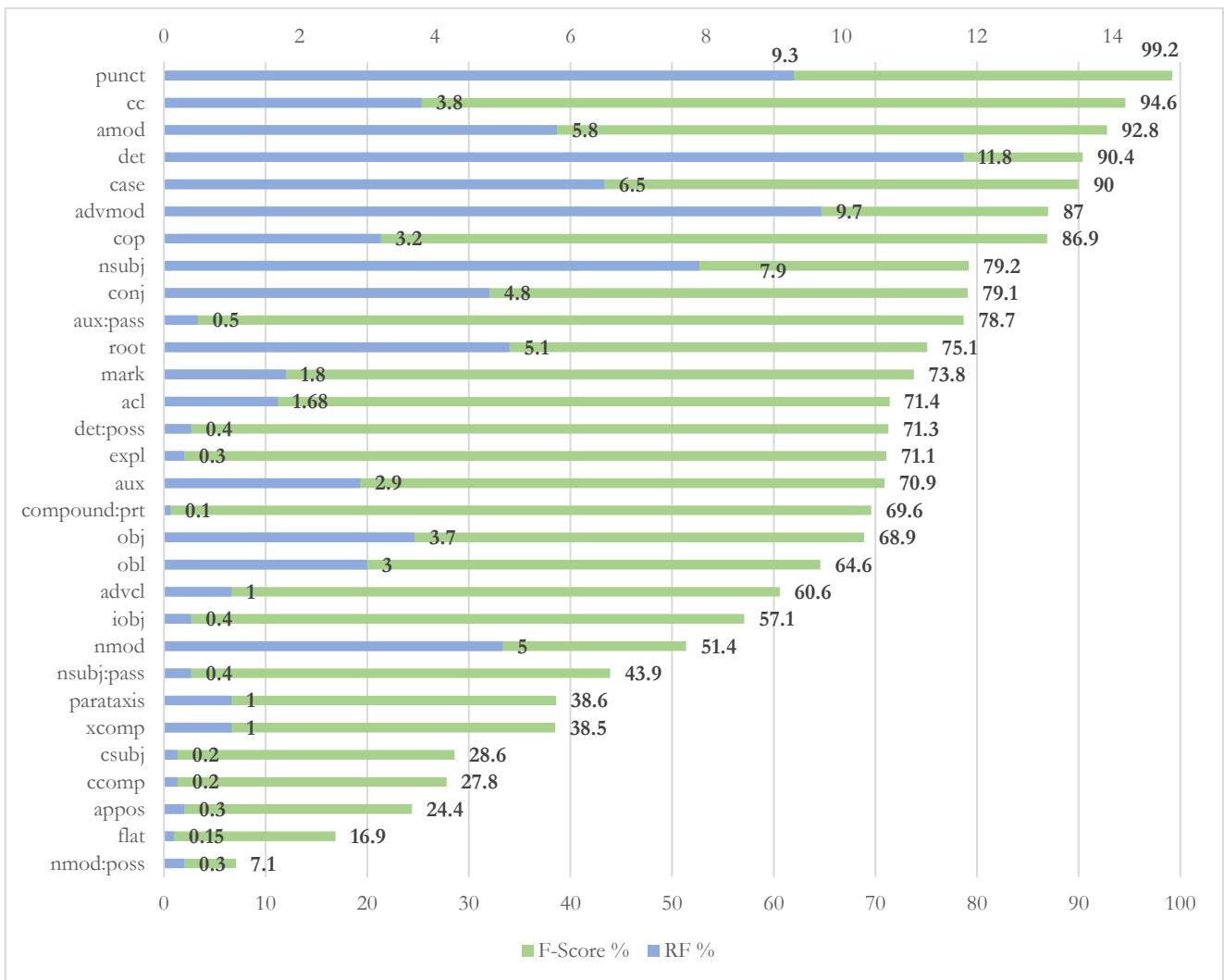


Chart 3 Accuracy by Anna 3.6 graph-based parser on each single dependency relation in the test set of Fragments. RF lies on the axis above, while F-score lies on the axis beneath. Relations are sorted per decreasing F-Score.

I then compared the accuracy by single dependency relations attained by the system on the Fragments with that attained by the same system on the 'UD_cont'. F-score is reported in Chart 4.

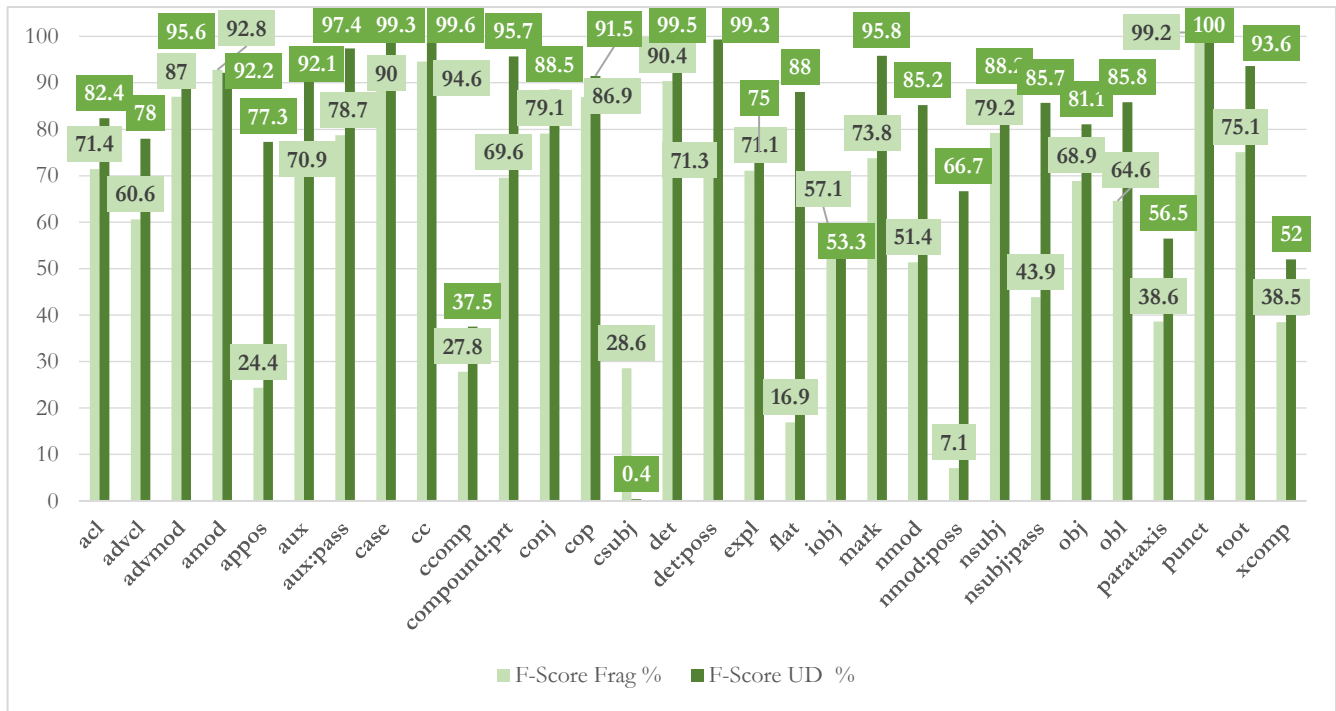


Chart 4 F-score by Anna 3.6 graph-based parser on Fragments and UD.

In addition, I provide an overall compared analysis of RF of each dependency relation in both the test sets. This helps analyse the different accuracy attained by the graph-based parser on the two different varieties. The distribution is reported in Table 30 and illustrated in Chart 5.

DEPREL	RF UD	RF FRAG
acl	0.9	1.68
advcl	0.4	1
advmod	5.6	9.7
amod	5.8	5.8
appos	2.7	0.3
aux	0.6	2.9
aux:pass	1.3	0.5

case	11.1	6.5
cc	3.3	3.8
ccomp	0.1	0.2
compound	0.2	0.0
compound:prt	0.4	0.1
conj	4.1	4.8
cop	1.5	3.2
csubj	0.0	0.2
csubj:pass	0.0	0.0
dep	0.1	0.0
det	12.9	11.8
det:poss	1.0	0.4
expl	0.1	0.3
fixed	0.0	0.0
flat	3.0	0.15
iobj	0.2	0.4
mark	0.8	1.8
nmod	8.1	5.0
nmod:poss	0.1	0.3
nsubj	5.7	7.9
nsubj:pass	1.3	0.4
nummod	1.1	0.0
obj	2.8	3.7
obl	6.5	3.0
parataxis	0.3	1.0

punct	7.4	9.3
root	5.1	5.1
xcomp	0.3	1.0

Table 30 RF of single dependency relations in UD and Fragments.

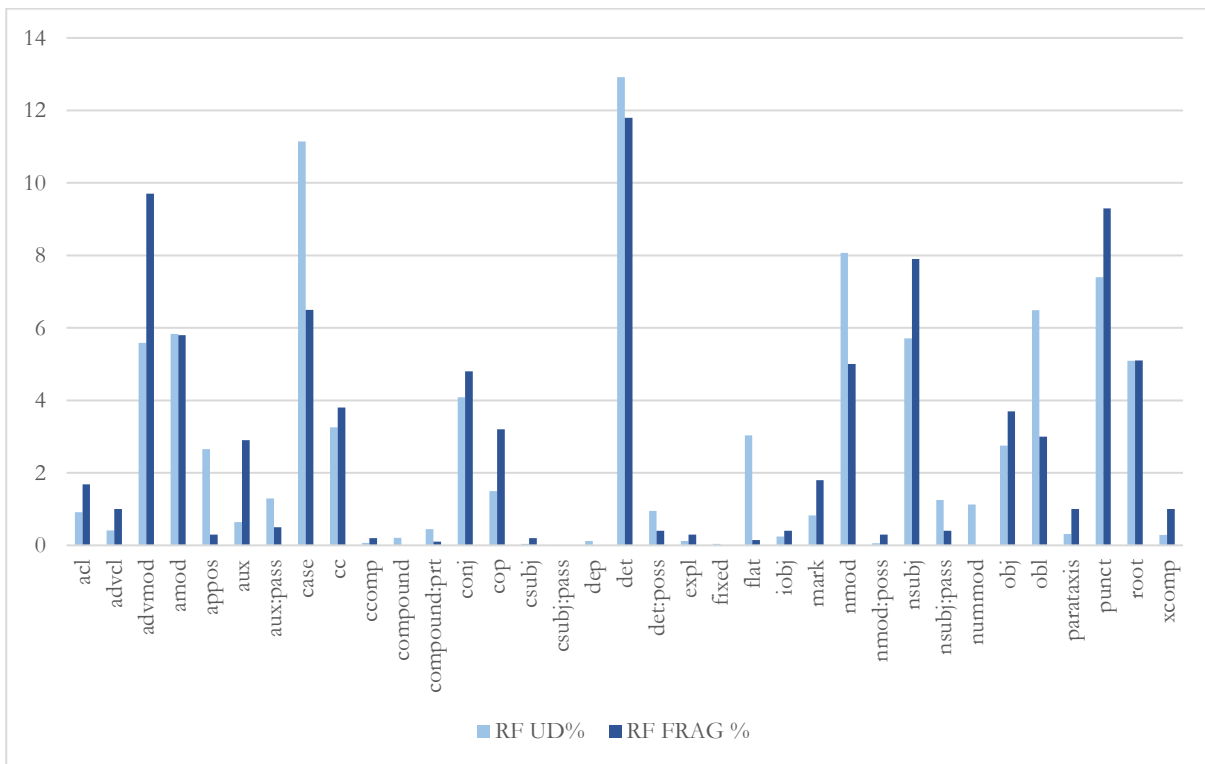


Chart 5 RF of each dependency relation in UD and Fragments.

Chart 7 compares the results on those relations governing core arguments, including the *root* relation. Both the accuracy and the RF of each single deprel in the test set are reported. As for the Fragments, the three most frequent relations attained an accuracy which is remarkably above the average LAS. In particular, *nsubj*, which is the most frequent relation as well (7.9%), attained the highest accuracy (79.2%); *root* relation, which is the second most frequent one (5.1%) (it occurs once for each sentence in the treebank), attained a good accuracy as well (75.1%). The third most frequent relation of this group (3.7%)

is the *obj* relation, i.e. the direct object, whose accuracy was slightly above the overall average LAS (68.2 %).

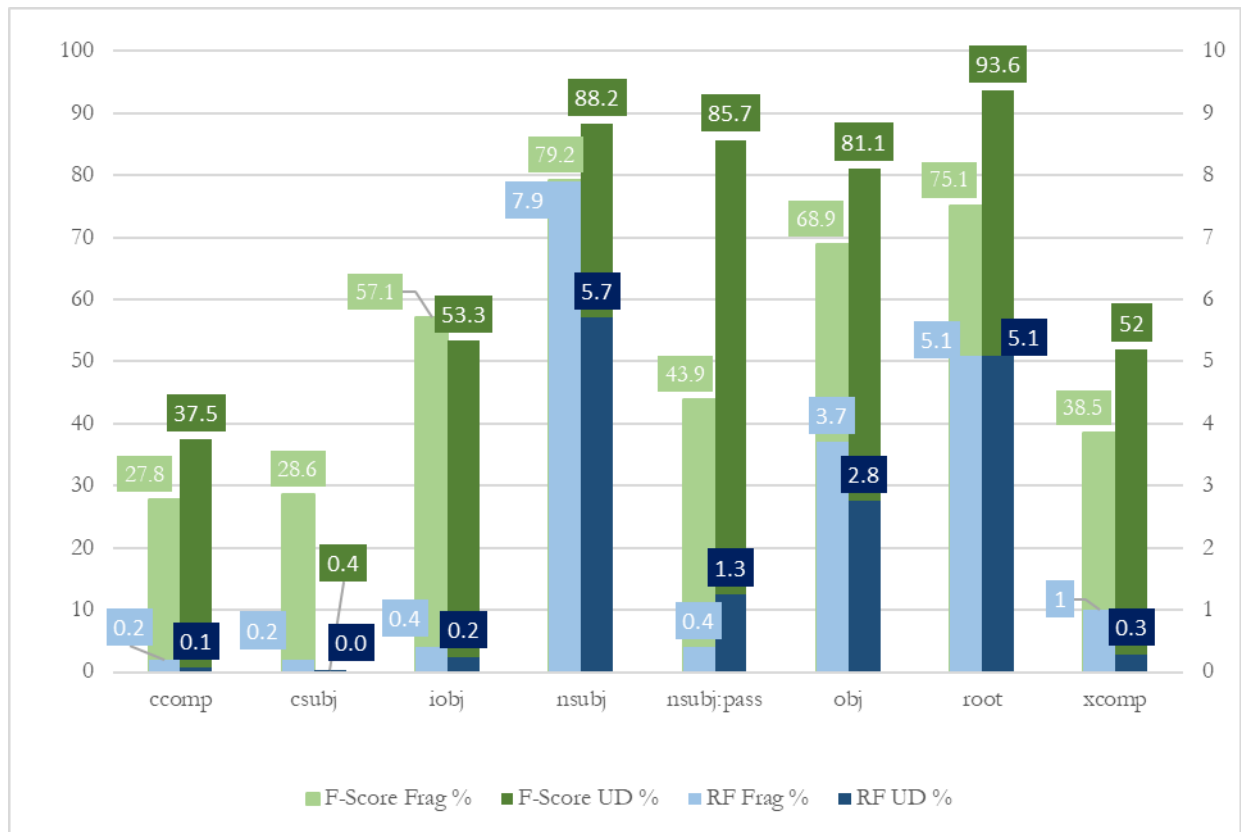


Chart 6 F-Score (values lie on y axis on the left) by Anna 3.6 graph-based parser on single dependency relations involving core arguments, on both the test sets. RF for each relation is reported as well (values lie on y axis on the right, the highest value was arbitrarily set on 10). Relation 'root' was included in this group.

As for UD, the system attained high accuracy on *nsubj*, *root* and *obj* as well. Since these relations are fundamental for the whole basic sentence structure, I will focus on the accuracy gap concerning these relations, trying to highlight the main parsing problems on the test set of Fragments. As for *obj*, i.e. the relation involving the direct object of the verb, F-score almost decreased by 12% with respect to the UD test set. Looking at the incorrectly assigned dependency relations for this class, 15.7 % of *obj* relations were assigned *nsubj* relation. It appears that the position of both direct objects and nominal subjects with respect to the verb could have a role in the degradation of accuracy on *obj* relation. In fact, the nominal subject and the direct object can both precede or follows the verb in declarative clauses in German. Also, it turned out that 6.5% of the wrongly parsed nominal subjects were actually assigned the *obj* relation, confirming that the accuracy on *nsubj* and *obj* and their mutual position in the sentence are correlated. In other words, it seems that the system could be influenced by differences in the distribution of postverbal

nominal subjects and direct objects between the training set (UD) and the test set (Fragments). In this respect, I checked the distribution of postverbal nominal subjects and direct objects in the UD training set, and in the test set of Fragments. It turned out that 26% of nominal subjects in the UD training set have a post verbal position in declarative clauses, while only the 10% of nominal subjects occupy a postverbal position in the same clauses in the test set of Fragments. Moreover, 32% of direct objects in the test set of Fragments occur in postverbal position, while 52% of direct objects in the training set occur in postverbal position. Therefore, the distribution of the position of these core-arguments with respect to the verb clearly varies between training set (UD) and test set (Fragments). This could have influenced the accuracy by the parsing model on the test set of Fragments. An example of a nominal subject occurring in post-verbal position that was mistakenly assigned the relation *obj* is reported in Figure 6.

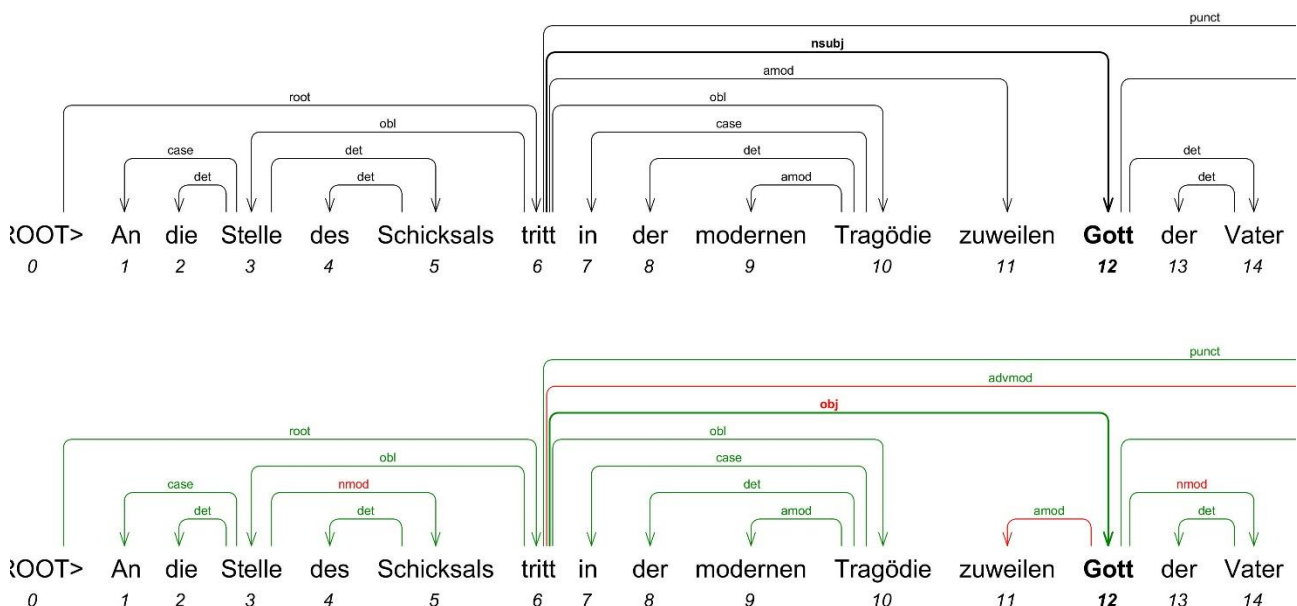


Figure 5 A sub-tree in linear form from the test of Fragments where the dependency relation *nsubj* in post-verbal position was incorrectly parsed as *obj*.

In Figure 6, the main verb is the finite verb 'tritt' (token 6), while the subject is the noun *Gott* occurring in post-verbal position (token 12). In addition, a noun phrase occurs at the beginning of the sentence, where the head is the noun *Stelle*, and another noun phrase, whose head is *Tragödie* occurs between the main verb and the nominal subject. In this case, the system mistakenly parsed this sub-tree assigning the relation *obj* to the nominal subject of its advanced position in the sentence. Moreover, there is no direct object in this sentence indeed, since the verb *treten* is intransitive and requires an indirect object, which was supposed to be the noun *Vater* in this case. But the indirect object was wrongly parsed either, since it was assigned *nmod* relation.

As for other errors involving *nsubj* relation, 2.4 % of *nsubj* were assigned *det* relation, i.e. they were classified as determiners, especially when the role of subject is played by a personal pronoun such as *der* or *die*. Furthermore, 1.8 % of them was assigned the label *root*, therefore they were classified as main verbs of the sentence rather than as subjects. As for *root* relation, the two varieties show a gap of 18% in F-score, which is significantly higher with respect to the gap shown on the other two previously considered relations. In the test set of Fragments, the confusion concerning this relation was more various, compared with that shown on *nsubj* and *obj* relation. Just to name the most frequent errors, *root* was incorrectly parsed as *nsubj* in 3.4 % of cases, as *cop* in 2.6 % of cases, as *advmod* in 2.3 % of cases., as *xcomp* in 1.8% of cases, as *obj* in 1.5 % of cases.

In the UD test set, the relation *nsubj:pass* attained high accuracy as well, and the accuracy gap between the two varieties concerning this relation was very high (41.8%). In the test set of Fragments, *nsubj:pass* was mismatched with *nsubj* in 32.3% of cases in Fragments. Since the confusion rate between *nsubj* and *nsubj:pass* is remarkable, this parsing problem could be caused by troubles in correctly parsing the forms of the passive auxiliary *werden* occurring with the past participle from of the passive verbs. But this is not the case, since almost 96% of the occurrences of *werden* as passive auxiliary in the test set of Fragments were correctly parsed by the system as ‘aux:pass’. Therefore, the troubles concerning passive verbs in the test set of Fragments could be due to the really low frequency of this sub-class of the *nsubj* relation. In fact, RF in Fragments is really low, while RF of this relation in the UD test set is more than double. Therefore, the low number of occurrences of observation to assign to these class could have played a role in the high error rate. As for *iobj*, the system performed slightly better on the test set of Fragments rather than on the UD test set, but accuracy on this relation was low in both cases. As for RF of this relation, it is really low in both test sets. Such problem affects the training set as well, as shown in Table 31. Therefore, troubles in parsing indirect objects could be due to *underfitting*, i.e. the observations in the training set are not enough to build an accurate model which is able to predict the same classes observed in the training data on unseen data.

Training Set	Deprel	RF%
de-ud2.0-train	csubj	0.06
	ccomp	0.23
	xcomp	0.34
	iobj	0.43

Table 31 RF of some core-arguments relations in the training set.

As for the relations involving clausal core arguments, i.e. *csubj*, *ccomp* and *xcomp*, parsing accuracy remarkably drops for all the relations belonging to this group in both the test sets. As for the target variety, the system attained a particularly low F-score: *csubj* 28.6%, *ccomp* 27.8 %, *xcomp* 38.5%. There are different possible explanations behind this result. First of all, such relations tend to show much lower RF in both the test sets with respect to the other relations of this group. This tendency affects the distribution of these relation in the training set as well, as shown in Table 31. Therefore, like in the case of *iobj*, the strong decrease in accuracy can be caused by *underfitting* again. Moreover, all these relations are likely to generate long-distance dependencies in German, i.e. they are relations where the number of tokens occurring between the head and the dependent can be conspicuous. Long-distance relations are notoriously more difficult to parse, (McDonald and Nivre 2007), (Salomoni 2017b). As for the subordinating relations involving completive clauses, subjective clauses, and adverbial clauses, i.e. *ccomp*, *csubj*, *advcl* respectively, verbs usually cause high dependency length in German, since they necessarily occupy the last position in the subordinate clause while depending back to the main verb of the main clause. As for open clausal complements, i.e. *xcomp*, the taxonomy of errors is more variable, since this class includes a wide range of different syntactic roles. Analysing the error rate of this relation, it is worth noting that 16.2 % of the relations that were supposed to be assigned *xcomp* were actually assigned *acl*, i.e. the class for adjectival clauses, such as relative clauses. An example is reported in Figure 7. In this case, the non-finite verb *vernehmen* was supposed to depend back on the verb of the higher clause *hören* through *xcomp* relation, since it is a non-finite verb in a non-finite clause introduced by the subordinating conjunction *ohne*. On the contrary, the system parsed the verb as if it was a finite verb modifying the element *ohne*, building a totally incorrect sub-tree for this clause.

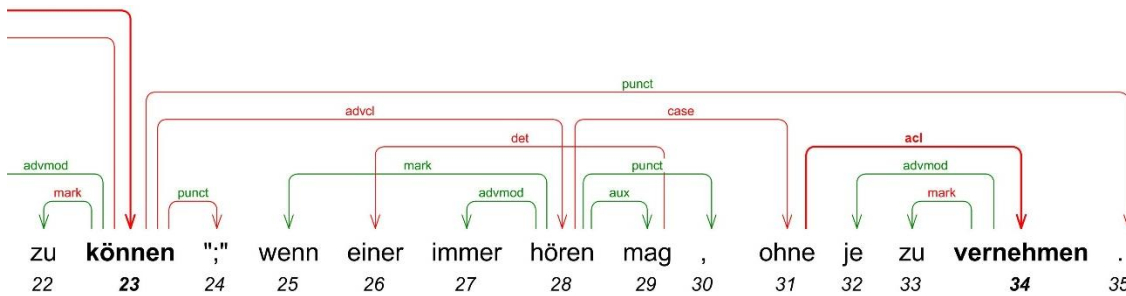
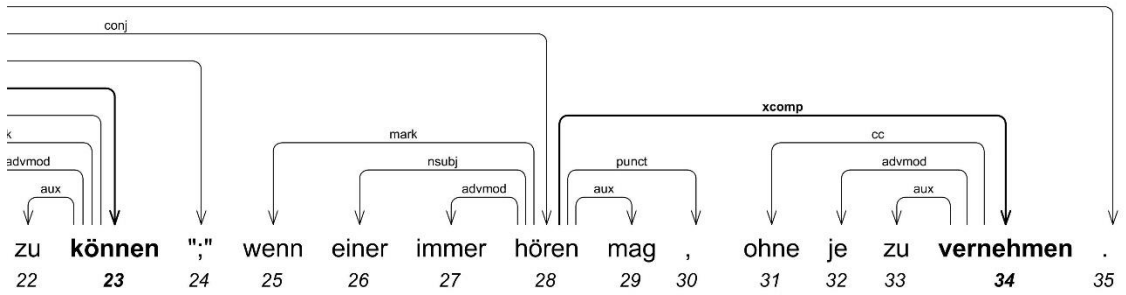


Figure 6 A sub-tree in linear form from the test set of Fragments where the dependency relation 'xcomp' was mistakenly parsed as 'acl'.

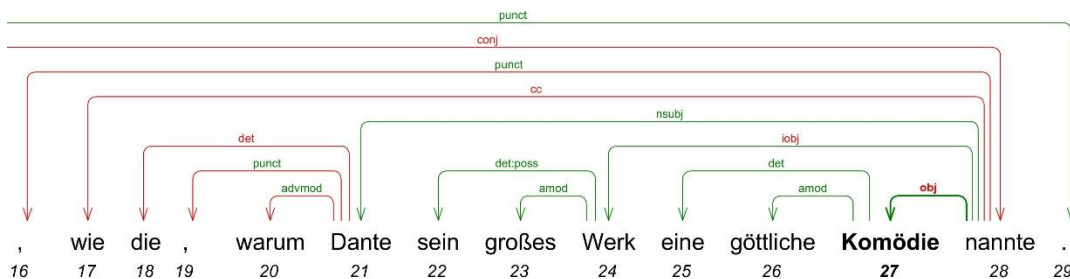
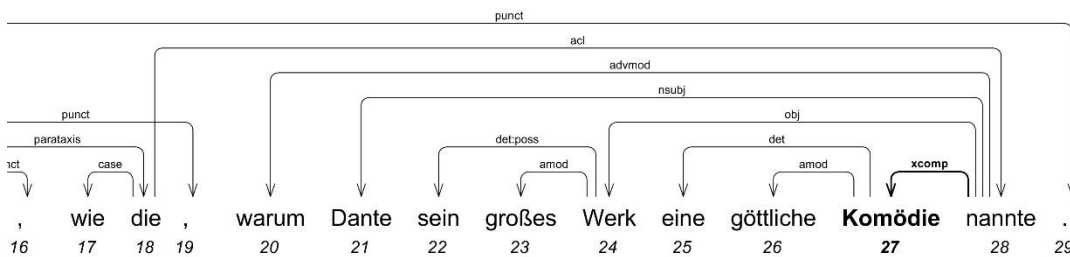


Figure 7 A sub-tree in linear form from the test set of Fragments where the dependency relation 'xcomp' was mistakenly parsed as 'obj'.

Furthermore, 10.8 % of the *xcomp* relations were mistakenly assigned *obj* relation. This error especially occurred in all those cases where *xcomp* were supposed to connect an adjective or a noun playing the role

of predicative part of a verb to the verb they refer to, such as in the example reported in Figure 45. In Figure 8, the relation spanning from the subordinate verb *nannte* to the noun *Komödie* was mistakenly parsed as *obj*. Indeed, the noun *Komödie* plays a predicative role for *nannte*, which is one of the functions covered by the class *xcomp*.

As for the clausal complements playing the role of clausal subject and clausal complement, i.e. *csbj* and *ccomp* respectively, their RF was really low in both the test sets and in the training set as well. In particular, *csbj* seems not to occur at all in UD_cont, since RF is equal to 0. The taxonomy of the relations with which they were mistakenly parsed is very sparse, therefore it is almost impossible to detect very precise parsing problems behind the error rate. As I said above, these two relations are very likely to generate very long-distance dependencies. This surely causes the system having many troubles in predicting a correctly parsed structure. In addition, overfitting could play a significant role as well, because of the very low RF in the training set.

2.11 First Release of the Treebank

Author	Work	Metadata	Tokens	ASC ⁴⁰
Friedrich Schlegel	Kritische Fragmente [entire collection]			
Friedrich Schlegel et al.	Athenäums-Fragmente [Fragments from 1 to 421]	LEMMA, UPOS, XPOS, HEAD, DEPREL	40,000	80%
Novalis	Blüthenstaub [entire collection]			

Table 32 Data and metadata in the first official version of the treebank (UD 2.4).

The process of semi-automatic annotation⁴¹ went on until when two of the three main collections of Fragments, i.e. *Kritische Fragmente* and *Blüthenstaub*, were completely annotated, while almost the entire biggest collection of *Athenäums-Fragmente* was almost entirely annotated (93% of the work, i.e. 421 Fragments out of 451). Overall, 80% of the data originally collected in the source corpus were annotated. Once the annotation was terminated, the file was brought into the original CoNLL-U format, and UPOS

⁴⁰ ASC stands for Annotation Source Corpus. I introduced this index in order to measure the percentage of annotated data with respect to the amount of raw data of the target genre collected in the source corpus.

⁴¹ The output of the Mate Tool pipeline was converted into a simplified version of the CoNLL-U format, in which all the comments, as well as all the double IDs of multiword tokens were removed. The correction of the output was then performed through Dependency Viewer, a simple tool run in Windows environment, which allows to both edit and visualize dependency trees. It was developed by the NLP group at the Nanjing University, China.

were automatically assigned to the data by running a conversion script, i.e. they were automatically derived from the annotated XPOS. The portrait of the treebank at this final stage of the annotation process is summarized in Table 32. At this stage, the treebank was ready to undergo the procedure that is required by the UD guidelines to officially release new treebanks in the UD online infrastructure⁴². In particular, the treebank has to pass an official automatic validation test, which is automatically run online in the UD infrastructure. Therefore, I uploaded the treebank file in CoNLL-U format in the dev branch of the GitHub page of the literary treebank (which had been previously created). After launching the validation script, some inconsistencies and errors in the annotation were detected, especially concerning dependency relations. All the problems were manually fixed by the deadline that is established to freeze data for one of the two annual official UD release. I uploaded the corrected treebank file to be evaluated again. The treebank passed the validation test, therefore it was finally published in the 2.4 release under CC BY-NC-SA 4.0 license (Nivre et al. 2019).⁴³ Each UD treebank is assigned a univocal ID in the online infrastructure. The official ID of the literary treebank in the online repository, as well as the number of tokens are reported in Table 33. Currently, the treebank is available in the most recent UD release as well, i.e. UD 2.6 (Zeman et al. 2020).

Trebank	ID	Genre	Tokens	First Release
Literary Treebank	LIT (UD_German-LIT)	Fragments	40,440	UD 2.4 (May 2019)

Table 33 Portrait of the literary treebank published in the UD 2.4 release.

⁴² Cf. https://universaldependencies.org/release_checklist.html.

⁴³The official web page of the literary treebank is available at the following link: https://universaldependencies.org/treebanks/de_lit/index.html, while the treebank file can be downloaded for research purposes from the GitHub official repository at the following link: https://github.com/UniversalDependencies/UD_German-LIT.

3 Linguistic Annotation

3.1 Introduction to Dependency Grammar

Different dependency-oriented grammatical descriptions have been developed over the centuries within theoretical linguistics, from Antiquity up to the early 20th century. For a comprehensive overview, see (Imrényi and Mazziotta 2020). As far as the modern theory of dependency grammar is concerned, two main approaches have risen: that by Luciene Tesnière (Tesnière 1959), and that by Mel'čuk (Mel'čuk 1988), (Mel'čuk 2009). Both of them agree about the core aspect of the dependency grammar, i.e. the fact that the syntax essentially consists of words linked by binary, asymmetrical relations called dependency relations, or dependencies for short. See e.g. (Kübler, McDonald, and Nivre 2009). These relations hold between a head and a dependent. The syntactic core of the whole sentence is the predicate, which is usually the root of all the other dependencies, which, in turn, can involve both direct and indirect dependents of the predicate. The relations can be illustrated both in linear form and in tree-like form. When they are displayed in linear form, relations are usually arches spanning from the head into the dependent. When they are displayed in tree-like form, they are represented as edges connecting the leaves (or nodes) of the tree, which, in turn, represents the lexical items of the sentence (even if, in some representations, the relations are extended to the punctuation too). In any case, dependencies encode grammatical functions within the sentence, which is usually the grammatical function played by the dependent with respect to the head. For instance, the dependent can be the nominal subject of the head, or a direct object, an oblique argument, or, it is a predicate, a clausal subject, and many others. This is the main difference with respect to those syntactic formalisms based on the constituency syntax (or phrase-structure grammar), which, on the contrary, describe the syntactic relations as a series of phrases, such as noun phrases, verbal phrases, or prepositional phrases, which combine with each other to progressively build larger structural units. See e.g. (Nivre 2005). Constituency syntax has a long-standing tradition in descriptive linguistics. Conversely, dependency syntax has gained a lot of ground over the last few years, especially in the field of computational linguistics and natural language processing (NLP). Such a rise in interest toward this form of syntactic representation is due to different (practical) reasons, ranging from the possibility of faster automatic syntactic parsing, to the usability of dependency annotations compared to constituency trees, and to the close parallelism between dependency relations and the predicate-argument relations, which are often the ultimate target of many NLP systems (Silveira 2016). Moreover, they were demonstrated very useful for multilingual NLP, which has been gaining a constantly increasing importance over the last few years (Zeman and Resnik 2008), (De Marneffe et al. 2014). In fact, as

explained below, the Universal Dependencies (UD) project was also born to facilitate multilingual NLP applications (Nivre et al. 2016).

Usually, dependency relations are typed, i.e. they are labelled according to the grammar function that is played by the dependent with respect to the head. A dependency representation of a sentence from the source corpus of the literary treebank is provided in the following example, both in linear form (Figure 9) and in tree-like form (Figure 10). In this case, the relations are those from the UD 2.0 scheme, which is introduced in the rest of this section.

[...] Die Tiefen unsers Geistes kennen wir nicht.⁴⁴

[...] The depths of our spirit are unknown to us.⁴⁵

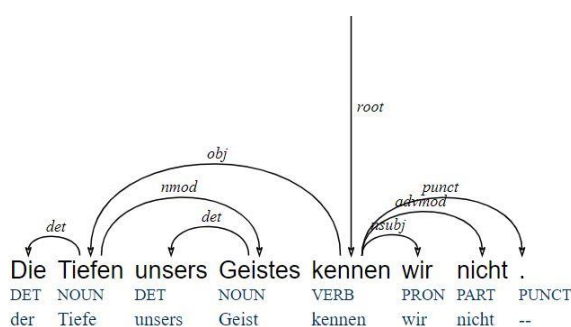


Figure 8 Dependency representation in linear form of the German sentence "Die Tiefen unsers Geistes kennen wir nicht.", according to the UD 2.0 scheme. The part of speech stands beneath each word, as well as the lemma.

As shown in both Figure 6 and Figure 7, syntax consists in hierarchical relations between lexical units (in this case, the punctuation is involved too). The root of the sentence is the main predicate, that is the verb *kennen* in this case, which is, therefore, the highest node of the tree (Figure 7). Each of the other lexical items of the sentence is edged, i.e. it is linked to a head through a syntactic relation.⁴⁶ In the linear representation (Figure 8), this means that each word has an incoming arc. According to the scheme applied here, which is illustrated in the following paragraph, each item of the sentence must have an incoming arch (or edge), i.e. a head, but it must have one incoming arc (or edge) at most. In addition, only one single root is allowed. In other words, each token must be single-headed, while each sentence must be single-rooted (Silveira 2016). The noun *Tiefen* is a direct dependent of *kennen*, as well as the pronouns *wir*. They depend on the predicate through the function of direct object (*obj*) and nominal subject (*nsubj*), respectively. The adverb *nicht* is also a direct dependent of the predicate, playing the role

⁴⁴ Novalis, *Bluetbestaub*, fragment 16.

⁴⁵ STOLJAR, Margaret Mahony, et al. (ed.). Novalis: Philosophical Writings. SUNY Press, 1997.

⁴⁶ Actually, in the representation in Figure 5, the main predicate also depends on a node. This is a fictional node indeed, which is required by the UD scheme. The reason is explained later in this chapter.

of negation marker (which is labelled as *advmod* in this scheme). Then, there are some indirect dependents of the predicate, such as the article *der*, which directly depends on the noun *Tiefen* as determiner (*det*).

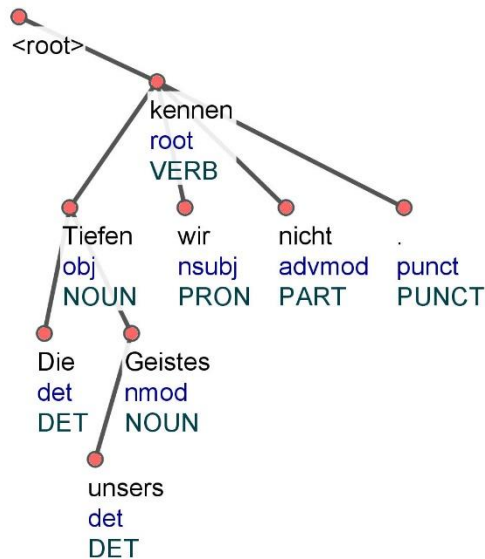


Figure 9 Dependency representation in tree-like form of the German sentence "Die Tiefen unsers Geistes kennen wir nicht.", according to the UD 2.0 scheme. The syntactic function encoded by each dependency is typed in blue; the dependency relations are the black edges connecting the red nodes, which, in turn, represents the words of the sentence. The words of the sentence are typed in black. The part of speech of each word is typed in green.

3.2 The Annotation Scheme: An Overview on Universal Dependencies

The Universal Dependencies (UD) scheme is a concrete application of the dependency grammar. Overall, UD aims at developing a cross-linguistically consistent standard for the linguistic annotation of textual data, in which the syntactic annotation is based on the dependency formalism. In fact, UD is a project that was born with the goal of facilitating multilingual-parser development, cross-lingual learning, research on syntactic parsing, and linguistic treebank-based analyses from a language typology perspective (Nivre et al. 2016).⁴⁷ The annotation scheme is based on an evolution of (universal) Stanford Dependencies (SD) (De Marneffe and Manning 2008), (De Marneffe et al. 2014), which were combined with the Google universal part-of-speech tags (Petrov, Das, and McDonald 2012) and the Intersect interlingua for morphosyntactic tag sets (Zeman 2008). The general philosophy is to provide a universal

⁴⁷ Cf. <https://universaldependencies.org/introduction.html>

inventory of categories and guidelines to facilitate consistent annotation of similar constructions across languages, while allowing language-specific extensions when necessary. From a wider perspective, UD is an open-community scientific effort, which is now counting hundreds of contributors that has been developing more than 150 treebanks in over 80 languages all over the world. In the history of computational linguistics, this is the first attempt ever to develop a shared, international framework to work with parsed corpora. In fact, historically, dependency representations for NLP have been developed for specific languages; this has often led to very significant disparities between the representations of the same linguistic phenomena across languages (Silveira 2016). By contrast, thanks to UD, treebanks are not designed according to subjective criteria adopted in local initiatives, but in compliance with official standard guidelines, which are accepted and constantly discussed by a worldwide research community. Without any standard annotation, results of both NLP tasks and treebank-based analysis are not comparable, especially in a cross-lingual perspective. Consequently, UD is a great attempt to meet the demand for a shared methodology by the CL and NLP communities. The UD treebanks are collected in an online infrastructure, and they are freely available for research purposes. As said above, UD was the culmination of a series of stages, each one aiming to provide a homogeneous annotation for a certain level of annotation. The first product of this progressive process was the first collection of treebanks annotated with an harmonized scheme called Universal Dependency, which was announced in 2013 (McDonald et al. 2013a). The official version (1.1) of an online repository of multilingual harmonized treebanks was presented in 2016 (Nivre et al. 2016). That version, which had been released in May 2015, counted 19 treebanks in 18 different languages. Over the last few years, the number of treebanks developed within the project has dramatically surged up. The most recent version, that is UD 2.4 (Nivre et al. 2019), was released on 15th May 2019, and counts 143 treebanks in 83 different languages. The standard data format in which the UD treebanks are encoded is the CoNLL-U format (Nivre et al. 2016) (see Chapter 2).

3.2.1 Parts of Speech

At the levels of parts of speech, the UD 2.4 scheme implements the universal tag set (UTS) (Petrov, Das, and McDonald 2012). It consists in a set of coarse lexical categories that exist across languages. In the UD data format, they are named UPOS, which stands for Universal Part of Speech. The tag set is summarized in Table 34, 35, and 36.

UPOS	Meaning
ADJ	Adjective
ADV	Adverb
INTJ	Interjection
NOUN	Noun
PROPN	Proper Noun
VERB	Verb

Table 34 Universal tag set for open lexical classes.

UPOS	Meaning
ADP	Adposition
AUX	Auxiliary
CCONJ	Coordinating conjunction
DET	Determiner
NUM	Numeral
PART	Particle
PRON	Pronoun
SCONJ	Subordinating conjunction

Table 35 Universal tag set for closed lexical classes.

UPOS	Meaning
PUNCT	Punctuation
SYM	Symbol
X	Other

Table 36 Universal tag set for lexical classes other than the previous ones.

As said above, UD allows language-specific metadata too. In this work, I used specific fine-grained POS-tags. The official tag set for language-specific parts of speech for German is the *Stuttgart-Tübingen-Tag Set* (STTS) (Schiller, Teufel, and Thielen 1995), which has been used in many important German corpora, such as the Tiger Treebank (Brants et al. 2002). In the CoNLL-U data format, the language-specific POS fills the XPOS field. STTS is reported in Table 37.

XPOS	Meaning	Examples
ADJA	Attributive adjective	[das] große [Haus]
ADJD	Adverbial or predicative adjective	[er fährt] schnell, [er ist] schnell
ADV	Adverb	schon, bald, doch
APPR	Preposition	in [der Stadt], ohne [mich]
APPRART	Preposition with article	im [Haus], zur [Sache]
APPO	Postposition	[ihm] zufolge, [der Sache] wegen
APZR	Right Circumposition	[von jetzt] an
ART	Definite and indefinite article	der, die, das, ein, eine
CARD	Cardinal number	zwei [Männer], [im Jahre] 1994
FM	Foreign material	[Er hat das mit ``] A big fish [" übersetzt]
ITJ	Interjection	mhm, ach, tja
KOUI	Subordinating conjunction with zu + non-finite form	um [zu leben], anstatt [zu fragen]
KOUS	Subordinating conjunction	weil, dass, damit, wenn, ob
KON	Coordinating conjunction	und, oder, aber
KOKOM	Comparative conjunction	als, wie
NN	Noun	Tisch, Herr, [das] Reisen
NE	Proper noun	Hans, Hamburg, HSV
PDS	Demonstrative pronoun in substitutive function	dieser, jener
PDAT	Demonstrative pronoun in attributive function	jener [Mensch]
PIS	Indefinite pronoun in substitutive function	keiner, viele, man, niemand
PIAT	Indefinite pronoun in attributive function without any determiner	kein [Mensch], irgendein [Glas]
PIDAT	Indefinite pronoun in attributive function with determiner	[ein] wenig [Wasser], [die] beiden [Brüder]
PPER	Personal pronoun	ich, er, ihm, mich, dir
PPOSS	Possessive pronoun in substitutive function	meins, deiner
PPOSAT	Possessive pronoun in attributive function	mein [Buch], deine [Mutter]
PRELS	Relative pronoun in substitutive function	[der Hund,] der
PRELAT	Relative pronoun in attributive function	[der Mann,] dessen [Hund]
PRF	Reflexive pronoun	sich, einander, dich, mir
PWS	Interrogative pronoun in substitutive function	wer, was
PWAT	Interrogative pronoun in attributive function	welche[Farbe], wessen [Hut]
PWAV	Adverbial relative or interrogative pronoun	warum, wo, wann, worüber, wobei
PAV	Pronominal adverb	dafür, dabei, deswegen, trotzdem
PTKZU	Particle zu	zu [gehen]
PTKNEG	Negation particle	nicht
PTKVZ	Separable particle of verbs	[er kommt] an, [er fährt] rad
PTKANT	Answer particle	ja, nein, danke, bitte
PTKA	Particle in adjectives or adverbs	am [schönsten], zu [schnell]
TRUNC	First member of a composition	An- [und Abreise]
VVFIN	Finite verb	[du] gehst, [wir] kommen [an]
VVIMP	Imperative verb	komm [!]
VVINFIN	Non-finite verb	gehen, ankommen
VVIZU	Non-finite verb with zu	anzukommen, loszulassen

VVPP	Past participle	gegangen, angekommen
VAFIN	Finite auxiliary verb	[du] bist, [wir] werden
VAIMP	Imperative auxiliary verb	sei [ruhig !]
VAINF	Non-finite auxiliary verb	werden, sein
VAPP	Auxiliary verb in past-participle form	gewesen
VMFIN	Finite modal verb	dürfen
VMINF	Non-finite modal verb	wollen
VMPP	Modal verb in past-participle form	gekonnt, [er hat gehen] können
XY	Non-word	3:7, H2O, D2XW3
\$,	Comma	,
\$.	End-of-sentence punctuation	. ? ! ; :
\$(Internal punctuation	- [] ()

Table 37 Stuttgart-Tübingen Tag Set (STTS).

3.2.2 Syntactic Relations

As for the syntactic relations, the UD scheme consists in typed dependency relations between words. In UD 2.4, there are two layers of dependency relations: basic dependencies and enhanced dependencies. The basic dependencies are mandatory, therefore they have necessarily to be annotated in all UD treebanks. They form a tree, in which only one word is the head of the whole sentence, and it depends on a fictional root node. At the same time, all the other words depend on another word in the sentence, as exemplified above in Figure 10. Enhanced dependencies are optional, and they pertain to a further level of annotation. In fact, they add (and in a few cases change) basic relations in order to give a more complete basis for the semantic interpretation of the sentence. They are especially used for treating some specific phenomena, such as ellipsis resolution, and they do not usually generate a tree, but a general graph structure. They were not considered in this work, therefore they are not analysed in this section. Since the UD standard is designed as a mixed functional-structural system, basic dependencies encode both the grammar function and the structural category of the dependent. By grammar function, as anticipated above, I mean the syntactic function played by the dependent with respect to the head. Conversely, by structural category, I mean the role of the dependent in the structure of the sentence. The structural category changes according to whether it generates a dependency within a clause or, on the contrary, whether it introduces a dependency that spans into a new clause. In fact, the sentence is assumed to consist of four main structural elements⁴⁸:

- **Noun phrases**, such as nominal subjects, or oblique arguments

⁴⁸ Cf. <https://universaldependencies.org/u/overview/syntax.html#a-mixed-functional-structural-system>.

- **Clauses headed by a predicate**, such as final clauses or adverbial clauses
- **Other miscellaneous modifiers**, such as adverbs or adjectives
- **Function words**⁴⁹

For instance, if the lexical item with the syntactic function of subject of a predicate is a noun phrase within a clause, it is typed as nominal subject. Therefore, the dependent has the grammar function of subject, while, structurally, it is a nominal. By contrast, if a lexical item with the role of subject of a predicate is the predicate of a completive clause that, in turn, play the role of subject of the main clause, such an item is typed as clausal subject. Therefore, the dependent has the grammar role of subject, but, structurally, it is a clause headed by a predicate. Similarly, if a lexical item modifies a verb within a clause, it can be typed in two ways: as adverb, if it belongs to the group of miscellaneous modifiers; as oblique, if it is a noun (i.e. part of a noun phrase) that specify some semantic information about the verb, for instance specification, or location, or others. Conversely, if this item modifies a verb as predicate of a subordinate clause, it is typed as adverbial clause.

The core issue of each application of dependency grammar is to find criteria to detect the head of dependency relations. In UD, there are some fundamental principles orientating this operation⁵⁰:

- The primacy of content words. In fact, dependency relations hold primarily between content words, rather than being indirect relations mediated by function words. Words are thus mostly headed by content words, apart from specific cases.
- Function words attach as direct dependents of the most closely related content word. Apart from a few specific, special cases⁵¹, function words are never heads, but only dependents. It means that multiple function words related to the same content word always appear as siblings, generating a flat annotation structure. A typical case is that of auxiliary verbs, which include modal verbs in UD, which never depend on each other. Therefore, if there is a copula occurring in a nonverbal predicate, which, in turn, is modified by a modal verb, both the copula and the modal verb depend on the nonverbal predicative element through two distinct relations in a flat structure. The same happens for a copula and an auxiliary for past tense. If they occur together, they both depend on the predicate through a flat structure. In the UD guidelines, these type of relations involving function words as dependents are defined as *functional relations* or *function word relations*⁵². They are therefore regarded as different from the *dependency relations* between content

⁴⁹ In the UD guidelines, this category is not mentioned, since it is included in *miscellaneous modifiers*. On the contrary, it was here kept separated, in order to maintain a parallelism with the official taxonomy of the UD relations.

⁵⁰ Cf. <https://universaldependencies.org/u/overview/syntax.html>.

⁵¹ Cf. <https://universaldependencies.org/u/overview/syntax.html#the-status-of-function-words>.

⁵² Cf. <https://universaldependencies.org/u/overview/syntax.html#the-status-of-function-words>.

words. Indeed, this view makes function words functionally (but not structurally) similar to morphological operations, and it is also compatible with Tesnière’s notion of the *nucleus* (Tesnière 1959) as the locus of syntactic dependencies.

- Coordination is treated asymmetrically. In fact, the head of the relation is the first conjunct and all the other conjuncts depend on it through the *conj* relation. Moreover, coordinating conjunctions and punctuation delimiting the conjuncts are attached through the *cc* and *punct* relations, respectively, to the immediately following conjunct.
- Punctuation attaches to the head of the clause or phrase to which they belong.

In UD 2.4, there are 37 universal syntactic relations, which are a revision of the SD dependency relations (De Marneffe et al. 2014). They are sorted into groups according to their functional category. In particular, the UD taxonomy of functional categories, and consequently of syntactic relations, is based upon the crucial distinction between core arguments and oblique dependents (or non-core dependents) (Thompson 1997), (Andrews 2007), (Zeman 2017)⁵³. On the contrary, the argument/adjunct distinction is totally discarded. The taxonomy is reported in Table 38. Rows correspond to functional categories in relation to the head, while columns correspond to structural categories of the dependents. A brief explanation of the functional categories follows. Those relation in bold were used in the treebank annotation. The application of the single dependency relations is described later.⁵⁴

FC	Nominals	Clauses	Modifiers	Function Words
Core Arguments	nsubj (nominal subject) obj (direct object) iboj (indirect object)	csubj (clausal subject) ccomp (clausal object) xcomp (open clausal complement)		
Non-Core Arguments	obl (oblique modifier) vocative expl (expletive element) dislocated (dislocated elements)	advcl (adverbial clause)	advmod ⁵⁵ (adverb)	aux (auxiliary) cop (copula) mark (marker)
Nominal Dependents	nmod (nominal modifier) appos (apposition) nummod (numerical modifier)	acl (adjectival clause, i.e. clausal modifier of a nominal)	amod ⁵⁵ (adjectival modifier)	det (determiner) clf (classifier) case (case marking)

⁵³ Cf. <https://universaldependencies.org/u/overview/syntax.html#core-arguments-vs-oblique-modifiers>.

⁵⁴ For further information about the usage of each single dependency relation, see: <https://universaldependencies.org/u/overview/syntax.html>.

⁵⁵ The *advmod* relation is used not only for modifiers of predicates but also for other modifiers.

Main Predicate	root (root node)
Coordination	cc (coordinating conjunction) conj (conjunct)
Multi-Word Expressions (MWE)	fixed (fixed multiword expression) flat (flat multiword expression) compound (compound nouns)
Loose	list (a list of items) parataxis (asyndetic coordination)
Special	orphan (ellipsis) goeswith (two words mistakenly written separately) reparandum (overridden disfluency)
Other	punct (punctuation) dep (unspecified dependency)

Table 38 Syntactic relations in UD 2.4. FC stands for functional category.

The root node is the fundamental element in the sentence in dependency grammar. Functionally, it corresponds to the main predicate of the sentence, which could be either a verb or a predicative part of a nominal predicate, or even a nominal in the nominal sentences. In UD, each token must be headed. Therefore, conventionally, the root node depends on a fictional node, whose ID in the CoNLL-U format is always 0.

As said above, the distinction between **core arguments** and non-core dependents is fundamental in this scheme. The criterion adopted in UD to detect the core arguments follows (Andrews 2007), and it is based on a semantic principle. Let us therefore introduce the following definitions. If a noun phrase is serving as an argument of a two-argument verb, and receiving a morphological and syntactic treatment normally accorded to an agent of a primary transitive verb, it has the grammatical function **A**. Analogically, an argument receiving treatment normally accorded to a patient of a primary transitive verb has the grammatical function **P**. In addition, if a verb takes only a single argument, the verb is called intransitive, and its argument has the grammatical function **S**. In UD, those arguments that have one of the S, A, or P functions are considered core arguments (Zeman 2017). As a consequence, nominals whose grammatical function is A or S are called subjects, and their dependency relation to the verb is *nsubj*. Nominals whose grammatical function is P are called (direct) objects instead, and their dependency relation to the verb is *obj*. When the same functions are played by predicates of completive clauses, their dependency relations to the main predicate are *csbj* and *ccomp*, respectively. As shown in Table 9, indirect objects are currently regarded as core-arguments in the UD taxonomy. The status of indirect objects in the predicate-argument structure is notoriously debated (Andrews 2007), (Dryer 2007). In the UD 2.4 version, all those arguments that are bare nominals in dative are considered core arguments, as suggested

by Zeman (2017). Therefore, I considered indirect objects in ditransitive verbs as core-arguments, and they depend on the predicate through the *ibj* relation. However, this status of indirect objects is open to debate. As for **non-core arguments (or non-core dependents)**, the criteria to define them are notoriously highly debated in literature (Andrews 2007), (Dixon 2012). In the UD 2.4, all those arguments that are marked by coding strategies that are different from the strategies used by core arguments described above are considered non-core arguments. In German, in general, as in English, when a noun phrase has the role of argument of a predicate and it is introduced by a prepositional phrase, the noun of such phrase depends on the predicate through the *obl* relation (Zeman 2017). No distinction between adjuncts and oblique arguments is done (and only the label *obl* is used, regardless of the role of the noun phrase introduced by the preposition), since criteria to make this distinction are still very unclear in literature.⁵⁶

The main issues concerning the functional categories in UD were highlighted. For their application, see the next section, in which the application of the dependencies from functional categories other than core-argument and non-core dependent is also described.

3.3 The Linguistic Annotation of the Fragments

In this paragraph, I illustrate the linguistic annotation of the Fragments in detail. I directly base the explanations upon the syntactic trees, whose design is in fact the core of the annotation process. All the dependency trees displayed in the following paragraph were obtained through CoNLL-U Viewer⁵⁷, an online free-access tool hosted in the UD website, which depicts sentences in tree-like form, if fed with CoNLL-U files. This tool is not able to show either the lemma nor the POS lying in the field XPOS. Therefore, besides the syntactic relations, only the tag lying in the field UPOS of the file is reported. In this case, I opted to put the STTS in the field UPOS, therefore XPOS, i.e. the fine-grained POS-tag, is displayed in the following trees rather than UPOS. The reason of this choice is due to processing reasons, which were addressed in Chapter 2.

The examples of annotation are grouped in five macro categories: those relations occurring within a single clause; those occurring between clauses, i.e. between predicates, which, in turn, are grouped in coordination and subordination; those involving comparative constructions; those involving ellipsis. For each of these categories, I report and discuss a series of dependency trees from the gold-standard final version of the literary treebank (UD 2.4). In the titles of each example, I reported the names of those syntactic functions or syntactic phenomena that are discussed in that example. The relations mentioned

⁵⁶ Cf. <https://universaldependencies.org/u/overview/syntax.html#avoiding-an-argumentadjunct-distinction>

⁵⁷ http://universaldependencies.org/conllu_viewer.html

in the title are explained in detail, while those already discussed in previous examples are only mentioned or even skipped. When a syntactic relation is mentioned in the title between parenthesis, it means that it should have been discussed in a different macro category, but it was dealt with in that section for the matter of convenience. As for the parts of speech, each XPOS appearing for the first time is explained as well. However, for further explanations concerning the STTS, I redirect to the official webpage of STTS⁵⁸.

Before each tree, I reported the whole fragment in linear form from which the sentence was extracted, in order to provide the context. I believe that this choice is twofold. On the one hand, the relation with the original text is preserved, therefore the sentence is not dealt with as an isolated item. In fact, as I stated above, maintaining a parallelism between the original unannotated text and the treebanked one is one of the principles of the literary treebank. On the other hand, the context can help understand better the decisions made in the annotation. The reported fragments are taken from the raw texts, therefore the numeration follows that adopted in the edition from which the digital text was obtained. In addition, each fragment is followed by an English translation from a critical English edition of the collection to which it belongs. For both *Kritische Fragmente* and *Athenaeums Fragmente*, I reported the translation by Peter Firchow (Schlegel 1971). Whereas, for those fragments from *Bliithenstaub*, I reported the translation by Margaret Mahony Stoljar (Stoljar 1997). Both in the original fragment and in the English translation, the sentence which is then represented in the tree is highlighted in bold. In addition, I also reported the sentence in linear form, with an English gloss of each token beneath, when possible. Since both official English translations are not literal, and since the contents of fragments is often cryptic, this second translation exclusively aims to clarify the syntactic role of each single token in the sentence. In the captions of the reported trees, an ID is displayed, which is the univocal ID that the sentence is assigned in the treebank file. Finally, I also provide a brief colour key of the labels:

- Black: tokens.
- Blue: dependency relations between tokens.
- Green: XPOS.

⁵⁸ <https://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-table.html>.

3.3.1 Relations within a Single Clause

3.3.1.1 Main Verb in Simple Form, Nominal Subject, Direct Object, Interrogative Adverb, Nominal Modifier, Adjectival Modifier, Preposition, Determiner

[13] Wenn junge Personen beiderlei Geschlechts nach einer lustigen Musik zu tanzen wissen, so fällt es ihnen gar nicht ein, deshalb über die Tonkunst urteilen zu wollen. **Warum haben die Leute weniger Respekt vor der Poesie?**⁵⁹

13. When young people of both sexes know how to dance to a lively tune, it doesn't in the least occur to them to try to make a critical judgment about music just for that reason. **Why do people have less respect for poetry?**⁶⁰

Warum	haben	die	Leute	weniger	Respekt	vor	der	Poesie ?
Why	have	the	people	less	respect	for	the	poetry

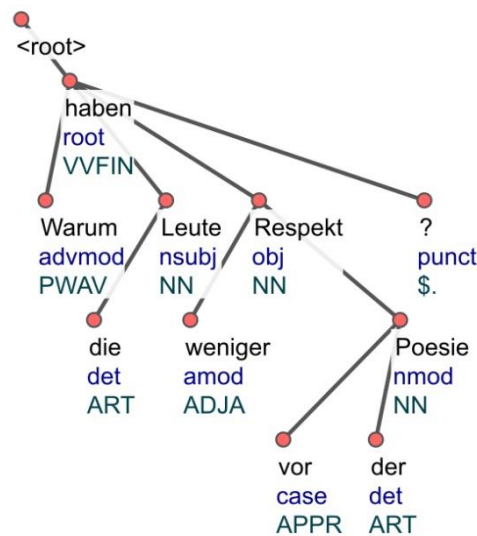


Figure 10 Dependency representation in tree-like form of the German sentence "Warum haben die Leute weniger Respekt vor der Poesie?" from *Athenäum Fragmente* by F. Schlegel, according to the UD 2.0 scheme.

sent_id = athenaeum-f13-s2

⁵⁹ F. Schlegel, *Athenäum Fragmente*, fragment 13.

⁶⁰ Friedrich Schlegel's *Lucinde and the Fragments*, University of Minnesota Press, 1971. ProQuest Ebook Central, <http://ebookcentral.proquest.com/lib/unibg-ebooks/detail.action?docID=345421>.

In Figure 11, there is a simple interrogative sentence, where the main predicate is the finite verb *haben* (VVFİN, where VV = verb, FİN = finite), used in a present tense, therefore in simple form (i.e. without any auxiliary). It depends on the fictional node through the *root* relation.

I consider now the core-arguments of *haben*. The noun *Leute* (NN, where NN = noun) is the nominal subject of the verb *haben*, therefore it depends on it through the *nsubj* relation (*nsubj* stands for nominal subject). The noun *Respekt* (NN) has the role of direct object, therefore it depends on *haben* through the *obj* relation.

As for the non-core dependents of the main predicate, *Warum* has the role of interrogative adverb (PWAV, where PAV = adverbial pronoun, W= interrogative), therefore it depends on *haben* through the *advmod* relation (*advmod* stands for adverbial modifier), which is the relation used in UD for all the adverbs.

I shift now to the nominal modifiers. First, the determinate article *die* (ART, where ART = article), which depends on the noun *Leute* through the *det* relation (*det* stands for determiner). This relation is used for all the articles, either definite or indefinite. There is no distinction between definite and indefinite in the STTS either, therefore the tag ART is always used for all the types of articles. The article *der* depends on the noun *Poesie* alike. In the noun phrase whose head is *Poesie*, *vor* works as preposition (APPR = preposition), therefore it depends on the noun through the *case* relation (*case* stands for case-marking element). This relation is used for all the kinds of propositions. In the same noun phrase, the noun *Poesie* has the role of modifier of the noun *Respekt*. Apparently, this noun could be regarded as a non-core dependent of the verb *haben*, therefore as a verbal modifier that specify the location where the action takes place. But, in this case, I evaluated it differently, since the syntactic construction [vor + noun in dativ] is inherently required by the noun *Respekt* to specify the recipient, regardless of the verb with which this noun occurs. In fact, if we, for instance, consider the nominal sentence *Respekt vor der Poesie!*, the construction [vor + noun in dativ] is triggered by the noun itself, without any verb. In terms of valency, the noun in dative case, in this construction, is not required by any semantic or logical valency of the verb. Rather, it seems to be required by the morpho-syntactic valency (Pittner and Berman 2015) of the noun *Respekt*. Therefore, I opted for considering the noun in dative case *Poesie* as modifier of the noun *Respekt* rather than non-core nominal dependent of the verb *haben*. Therefore, *Poesie* depends on *Respekt* through the *nmod* relation (*nmod* stands for nominal modifier). This relation is used for all those nouns modifying other nouns. Finally, *weniger* modifies the noun *Respekt* as an adjective (ADJA, where ADJ = adjective, A= attributive), therefore it depends on *Respekt* through the *amod* relation (*amod* stands for adjectival modifier), which is always used for all the kinds of those dependents of nouns that work as adjectives.

Finally, I will move to the relations concerning punctuation. In this case, there is a final mark involved. As for all the final marks, they always depend on the highest node possible of the main clause, if non-projectivity⁶¹ (Marcus 1965), (Robinson 1970) is avoided. In other words, the branch of the relations occurring between the head of the final mark and the final mark itself must not cross any other branch of the sentence. All final marks, regardless of their type (full stops, question marks, exclamation marks) depend on the highest node possible of the sentence, usually the main predicate, through the *punct* relation. Therefore, here, the question mark ? depends on *haben* through the *punct* relation. As for the XPOS for final marks, it is always the symbol \$ in STTS.

3.3.1.2 Main Verb in Complex Form, Auxiliary for Past Form, Oblique Argument

[26] Die Romane sind die sokratischen Dialoge unserer Zeit. **In diese liberale Form hat sich die Lebensweisheit vor der Schulweisheit geflüchtet.**⁶²

26. Novels are the Socratic dialogues of our time. **And this free form has become the refuge of common sense in its flight from pedantry.**⁶³

In	diese	liberale	Form	hat	sich	die	Lebensweisheit	vor	der	Schulweisheit	geflüchtet.
In	this	free	form	has	itself	the	common sense	from	its	refuge	run away

⁶¹ This phenomenon is generated by crossing branches.

⁶² *Kritische Fragmente*, fragment 26.

⁶³ Friedrich Schlegel's *Lucinde and the Fragments*, University of Minnesota Press, 1971. ProQuest Ebook Central, <http://ebookcentral.proquest.com/lib/unibg-ebooks/detail.action?docID=345421>.

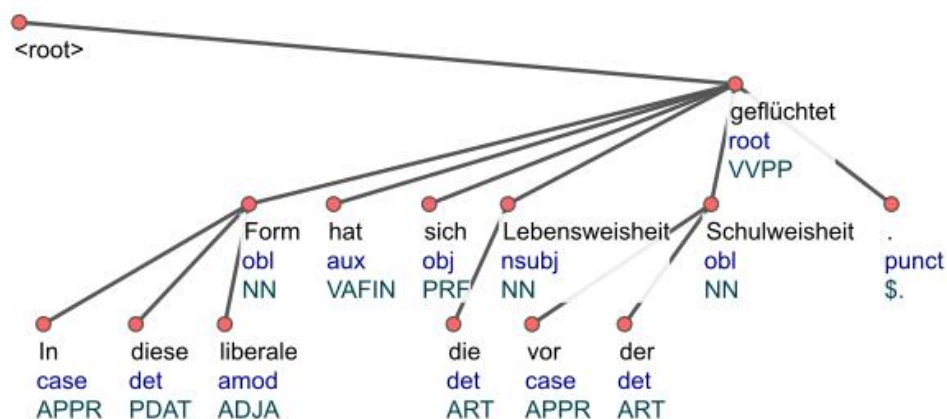


Figure 11 Dependency representation in tree-like form of the sentence “In diese liberale Form hat sich die Lebensweisheit vor der Schulweisheit geflüchtet” from Kritische Fragmente by F. Schlegel, according to the UD 2.0 scheme.

sent_id = lyceum-f26-s2

In Figure 12, there is a simple sentence, in which the main predicate is a verbal predicate in past form, precisely in *Partizip II* form, which therefore requires a complex construction. In fact, it is made up of the auxiliary verb *hat* (VAFIN, where VA = auxiliary verb, FIN = finite) and the past participle *geflüchtet* (VVPP, where VV = verb, PP = past participle). In these kind of verb phrases, the auxiliary depends on the verb through the *aux* relation (*aux* stands for auxiliary), while the verbal part in past participle is the head of the verbal predicate. Complex-verb constructions in the declarative clauses are usually discontinuous, since the verbal part occupies the final position, i.e. it lies at the very the end of the clause, while the auxiliary verb normally occupies the second position. In this case, the past participle *geflüchtet* is the verbal part of the main predicate also, therefore it stands at the end of the sentence, and it depends on the fictional node through the *root* relation.

As for the core-arguments of the predicate, the subject is the noun *Lebensweisheit* (NN), while the direct object is the reflexive pronoun *sich* (PRF, where P = pronoun, RF = reflexive). In UD 2.0, there is no specific dependency relation to deal with the reflexive pronouns, therefore they are assigned the relation of the syntactic function that they have in the clause.

As for the non-core dependents, I focus on the noun phrase *in diese liberale Form* first. The head of this phrase is the noun *Form* (NN), which is the governor of all the modifiers. In particular, *in* is a preposition (APPR = preposition) depending on the head through the *case* relation. Then, *diese* is a determiner, more precisely a demonstrative pronoun in attributive function, since it does not replace any noun, but it works as an adjective (PDAT, where PD = demonstrative pronoun, AT = attributive). It depends on *Form* through as *det*, which is the same relation used for articles. Therefore, for this kind of determiners, the disambiguation between attributive or substitutive function is done through the XPOS only. On the

contrary, the adjective *liberale* (ADJA, where ADJ = adjective, A = attributive⁶⁴) depends on *Form* through the *amod* relation, as usual. In turn, the noun *Form* depends on the main predicate through the *obl* relation (*obl* stands for oblique argument). Such relation is used which for all those nouns (or pronouns) working as non-core dependents of verbs. By non-core dependents, I mean items that do not play the role either of subject or object. Mostly, they specify additional circumstances such as location, time and manner (Zeman 2017). In this sentence, the same kind of relation also occurs between the noun *Schulweisheit* and the main predicate.

As for punctuation, the full stop should depend on the main predicate, if non-projectivity is avoided. Therefore, it depends on *geflüchtet* through the *punct* relation.

3.3.1.3 Nonverbal Predicate, Nominal Modifier of a Nonverbal Predicate

In Figure 13, there is a simple sentence with a nonverbal predicate, where the predicative role is played by the noun *Feindin*, while the role of copula is played by the finite auxiliary *ist* (VAFIN, where VA = auxiliary verb, FIN = finite). In UD, the predicative part of a nonverbal predicate, either a noun or an adjective, is considered as the head of the predicate. Consequently, in this case, the noun *Feindin* depends on the fictional node through the *root* relation, while *ist* depends on *Feindin* through the *cop* relation (*cop* stands for copula). The noun *Besitzungen* has the function of noun modifier of the nominal part of the predicate, therefore it depends on the noun *Feindin* through the *nmod* relation.

13. **Die Natur ist Feindin ewiger Besitzungen.** Sie zerstört nach festen Gesetzen alle Zeichen des Eigenthums, vertilgt alle Merkmale der Formazion. Allen Geschlechtern gehört die Erde; jeder hat Anspruch auf alles. Die Früheren dürfen diesem Primogeniturzufalle keinen Vorzug verdanken. – Das Eigenthumsrecht erlischt zu bestimmten Zeiten. [...] ⁶⁵.

14. **Nature is the enemy of eternal possession.** It destroys all signs of property according to the fixed laws, it eradicates all marks of formation. The earth belongs to all generations – each person has a claim

⁶⁴ In STTS, adjectives playing a predicative role, such as in the adjectival part of a nominal predicate, are tagged with a different tag (ADJD). More on this later in this paragraph.

⁶⁵ Novalis, *Bluetbestaub*, fragment 13.

to everything. Those born earlier may owe no advantage to the chance of process. The right to property is extinguished at certain times⁶⁶. [...]

Die	Natur	ist	Feindin	ewiger	Besitzungen.
The	nature	is	enemy	of eternal	possession.

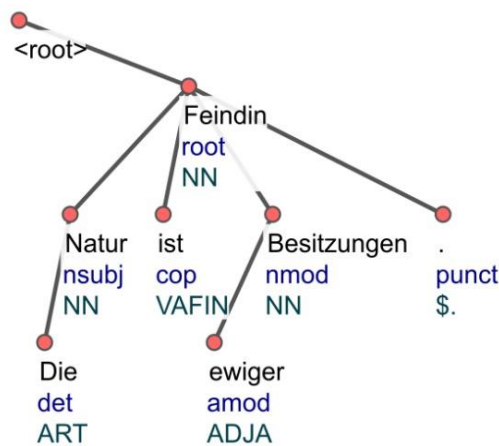


Figure 12 Dependency representation in tree-like form of the sentence "Die Natur ist Feindin ewiger Besitzungen" from *Blüthenstaub* by Novalis, according to the UD 2.0 scheme.

https://doi.org/10.1017/S0022268319000141

I reported this example in order to highlight the difference between this particular use of the *nmod* relation and the use of the use of the *obl* relation illustrated in Figure 12. In fact, in this case, *Besitzungen* is a noun which has a function of specification of the predicate, therefore, according to what explained above, it should be considered as an oblique argument. On the contrary, in the nominal predicates, nouns playing this role actually modify other nouns (or pronouns), or adjectives, never verbs. Therefore, the *obl* relation is never used for any modifier of the predicative part of the nominal predicates, since it exclusively pertains to those relations involving the direct non-core dependents of verbal nodes.

3.3.1.4 Indirect Object, Possessive Determiner

⁶⁶ STOLJAR, Margaret Mahony, et al. (ed.). *Novalis: Philosophical Writings*. SUNY Press, 1997. In this edition, there is a mismatch about the number of this fragment with respect to the edition from which the digital raw text comes from.

9. **Unser sämtliches Wahrnehmungsvermögen gleicht dem Auge.** Die Objekte müssen durch entgegengesetzte Media durch, um richtig auf der Pupille zu erscheinen⁶⁷.

9. **Our entire faculty of perception is like the eye.** Objects must pass through opposite mediums in order to appear correctly in the pupil.⁶⁸

Unser	sämtliches	Wahrnehmungsvermögen	gleicht	dem	Auge.
Our	complete	Faculty of perception	is like	the	Eye.

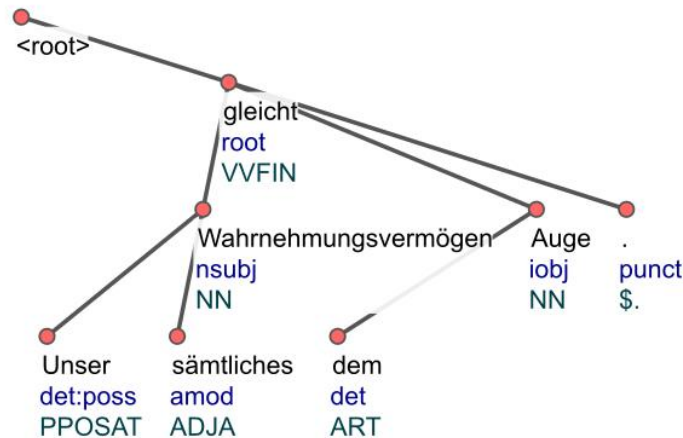


Figure 13 Dependency representation in tree-like form of the sentence " Unser sämtliches Wahrnehmungsvermögen gleicht dem Auge." from Blütenstaub by Novalis, according to the UD 2.0 scheme.

sent_id = bluethenstaub-f9-s1

In Figure 14, the main predicate is the intransitive finite verb *gleicht*. The role of subject is played by the compound noun *Wahrnehmungsvermögen*. This noun is modified by the possessive pronoun in attributive function *Unser* (PPOSAT, where PPOS = possessive pronoun, AT = attributive), which plays the role of determiner. Therefore, *Unser* depends on the noun it refers to through a sub-type of the *det* relation, i.e. the *det:poss* relation (in which *poss* stands for possessive), which is a specific relation for possessive determiners. The verb *gleichen* necessarily requires an indirect object in the dative case, which is the noun *Auge* in this case. Therefore, *Auge* depends on *gleicht* through the *iobj* relation. In German, the indirect object can be marked with the dative case without any preposition. When it happens, the indirect object depends on the predicate through the 'iobj' relation. Otherwise, if preceded by any preposition, it depends on the predicate through the *obl* relation, even when the object is in the dative case.

⁶⁷ Novalis, Blütenstaub, fragment 9.

⁶⁸ STOLJAR, Margaret Mahony, et al. (ed.). Novalis: Philosophical Writings. SUNY Press, 1997.

3.3.1.5 Modal Verb

[103] [...]So mächtig ist aber der Trieb nach Einheit im Menschen, daß der Urheber selbst, was er durchaus nicht vollenden oder vereinigen kann, oft gleich bei der Bildung doch wenigstens ergänzt; oft sehr sinnreich und dennoch ganz widernatürlich. Das Schlimmste dabei ist, daß alles, was man den gediegenen Stücken, die wirklich da sind, so drüber aufhängt, um einen Schein von Ganzheit zu erkünsteln, meistens nur aus gefärbten Lumpen besteht. Sind diese nun auch gut und täuschend geschminkt, und mit Verstand drappiert: so ist's eigentlich um desto schlimmer. Dann wird anfänglich auch der Auserwählte getäuscht, welcher tiefen Sinn hat für das wenige tüchtig Gute und Schöne, was noch in Schriften wie in Handlungen sparsam hie und da gefunden wird. **Er muß nun erst durch Urteil zur richtigen Empfindung gelangen!** Geschieht die Scheidung auch noch so schnell: so ist doch der erste frische Eindruck einmal weg⁶⁹.

103. But so powerful is the instinct for unity in mankind that the author himself will often bring something to a kind of completion which simply can't be made a whole or a unit; often quite imaginatively and yet completely unnaturally. The worst thing about it is that whatever is draped about the solid, really existent fragments in the attempt to mug up a semblance of unity consists largely of dyed rags. And if these are touched up cleverly and deceptively, and tastefully displayed, then that's all the worse. For then he deceives even the exceptional reader at first, who has a deep feeling for what little real goodness and beauty is still to be found here and there in life and letters. **That reader is then forced to make a critical judgment to get at the right perception of it!** And no matter how quickly the dissociation takes place, still the first fresh impression is lost.⁷⁰

In Figure 15, there is a simple sentence with a verbal predicate modified by a modal verb. In German, modal verbs, exactly like auxiliaries, occupy the second position in the clause, while the verb they refer to stands at the very end of the clause in non-finite form. Therefore, the finite modal verb *muß* (VMFIN, where VM = modal verb, while FIN = finite) depends on the non-finite verb *gelangen* through the *aux* relation, which is the same relation used for auxiliary verbs. In UD, no distinction is made in terms of syntactic functions between modals and auxiliaries. *gelangen* (VVINF, where VV = verb, while INF = non-finite) is the main predicate of the sentence, therefore it depends on the fictional node through root relation. Among the modifiers of *gelangen*, there are two adverbs, i.e. *nun* and *erst*. When more adverbs modify a verb at the same time, they depend on the verb through the *advmod* relation separately, generating a flat structure.

⁶⁹ F. Schlegel, *Kritische Fragmente*, fragment 103.

⁷⁰ Friedrich Schlegel's *Lucinde and the Fragments*, University of Minnesota Press, 1971. ProQuest Ebook Central, <http://ebookcentral.proquest.com/lib/unibg-ebooks/detail.action?docID=345421>.

Er	muß	nun	erst	durch	Urteil	zu	der	Richtigen	Empfindung	gelangen!
He	must	now	first	through	Judgement	to	the	right	perception	get

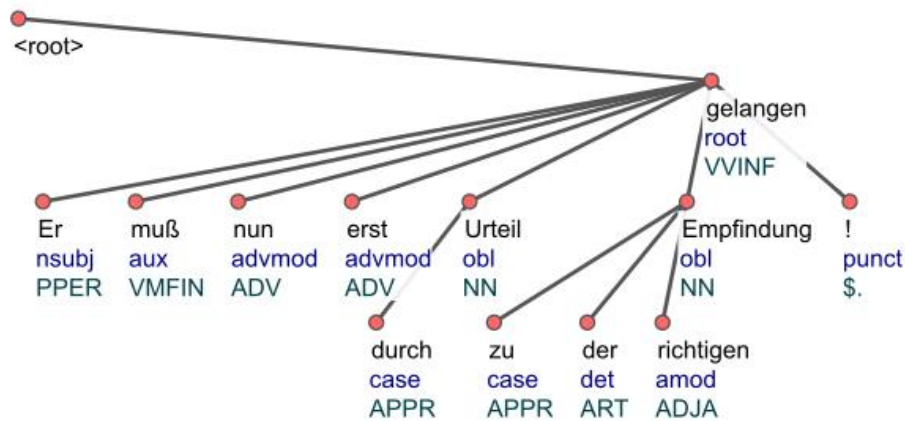


Figure 14 Dependency representation of the sentence "Er muß nun erst durch Urteil zu der richtigen Empfindung gelangen !" from *Kritische Fragmente* by F. Schlegel, according to the UD 2.0 scheme.

sent_id = lyceum-fl03-s10

3.3.1.6 Main Verb in Passive Voice

14. Leben ist der Anfang des Todes. Das Leben ist um des Todes willen. Der Tod ist Endigung und Anfang zugleich, Scheidung und nähere Selbstverbindung zugleich. **Durch den Tod wird die Reduktion vollendet**⁷¹.

15. Life is the beginning of death. Life is for the sake of death. Death is at once the end and the beginning - at once separation and closer union of the self. **Through death the reduction is complete.**⁷²

⁷¹ Novalis, *Blüthenstaub*, fragment 14

⁷² STOLJAR, Margaret Mahony, et al. (ed.). *Novalis: Philosophical Writings*. SUNY Press, 1997.

Durch	den	Tod	wird	die	Reduktion	vollendet.
Through	The	Death	is	the	Reduction	Finished.

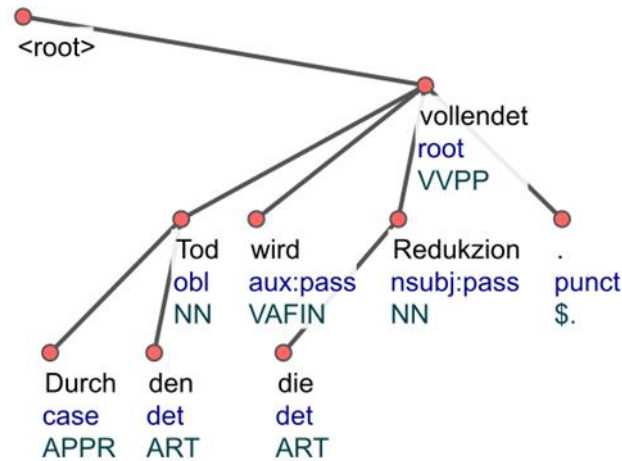


Figure 15 Dependency representation in tree-like form of the sentence "Durch den Tod wird die Reduktion vollendet." from *Blüthenstaub* by Novalis, according to the UD 2.0 scheme.

sent_id = bluetenstaub-f14-s4

In Figure 16, we have a simple sentence where the predicate is in passive form. In German, the passive form is built through the verb *werden* used as auxiliary verb, and the past participle of the verb. As in the case of complex verbs in declarative sentences, the auxiliary occupies the second position in the sentence, after the first element, while the verbal part in past participle stands at the very end of the clause. Therefore, in this sentence, the finite auxiliary *wird* depends on the past participle *vollendet* through the *aux:pass* relation, which is a subtype of the *aux* relation (where *pass* stands for passive). This relation is used for all those auxiliaries involved in passive constructions. The past participle *vollendet* is the main predicate, therefore it depends on the fake node through the *root* relation. When a passive predicate occurs, the nominal subject depends on the verb through a specific subtype of the *nsubj* relation, which is the *nsubj:pass* relation. Such a choice is made to mark the fact that the subject is not actually the agent of the action, but the patient indeed. Here, therefore, the noun *Reduktion* depends on *vollendet* through the *nsubj:pass* relation.

3.3.1.7 Modal Verb Depending on a Main Verb in Passive Voice

[117] **Poesie kann nur durch Poesie kritisiert werden.** Ein Kunsturteil, welches nicht selbst ein Kunstwerk ist, entweder im Stoff, als Darstellung des notwendigen Eindrucks in seinem Werden, oder durch eine schöne Form, und einen im Geist der alten römischen Satire liberalen Ton, hat gar kein Bürgerrecht im Reiche der Kunst⁷³.

117. **Poetry can only be criticized by way of poetry.** A critical judgment of an artistic production has no civil rights in the realm of art if isn't itself a work of art, either in its substance, as a representation of a necessary impression in the state of becoming, or in the beauty of its form and open tone, like that of the old Roman satires.⁷⁴

Poesie	kann	nur	durch	Poesie	kritisiert	werden.
Poetry	can	only	by way of	poetry	criticized	be.

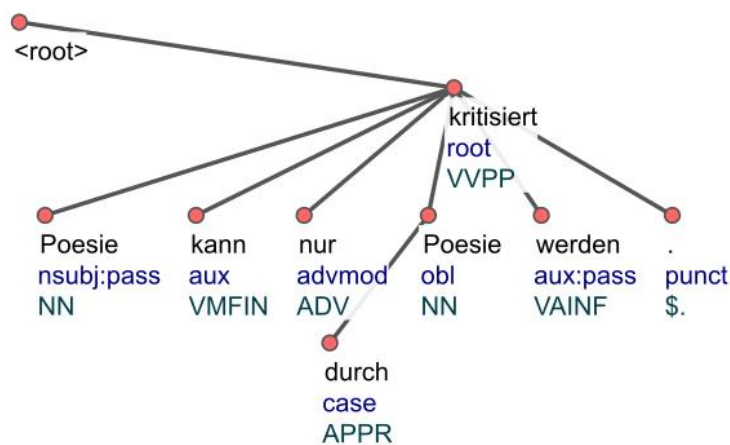


Figure 16 Dependency representation in tree-like form of the sentence "Poesie kann nur durch Poesie kritisiert werden." from *Kritische Fragmente* by F. Schlegel, according to the UD 2.0 scheme.

sent_id = lyceum-fl17-s1

In Figure 17, there is the modal verb *kann* in second position, while the main predicate is in passive voice, therefore it is made up of the past participle *kritisiert* and the non-finite auxiliary *werden*. In German, as shown in Figure 15, a modal verb requires the verb at the end of the clause to be in non-finite form. When this verb is in passive voice, the past participle occupies the penultimate position, while the auxiliary occupies the last position in the clause. The auxiliary occurrences in non-finite form. As illustrated in

⁷³ F. Schlegel, *Kritische Fragmente*, fragment 117.

⁷⁴ Friedrich Schlegel's *Lucinde* and the *Fragments*, University of Minnesota Press, 1971. ProQuest Ebook Central, <http://ebookcentral.proquest.com/lib/unibg-ebooks/detail.action?docID=345421>.

Figure 15, modal verbs always depend on the main predicate through the *aux* relation. Therefore, here, *kann* depends on the past participle *kritisiert* through the *aux* relation, while *werden* depends on the same verb through the *aux:pass* relation.

3.3.1.8 Modal Verb Modifying Another Modal Verb, with the Main Verb in Passive Voice

[20] **Eine klassische Schrift muß nie ganz verstanden werden können.** Aber die, welche gebildet sind und sich bilden, müssen immer mehr draus lernen wollen⁷⁵.

20. **A classical text must never be entirely comprehensible.** But those who are cultivated and who cultivate themselves must always want to learn more from it.⁷⁶

Eine	klassische	Schrift	muß	nie	ganz	verstanden	werden	können.
A	classical	text	must	never	entirely	understood	be	can

In Figure 18, there is a simple sentence, where the predicate has a passive voice. Here, the modal verb *must* actually refers to the other modal verb *können*, which stands at the end of the sentence. In turn, *können* refers to the verb in past participle *verstanden*.

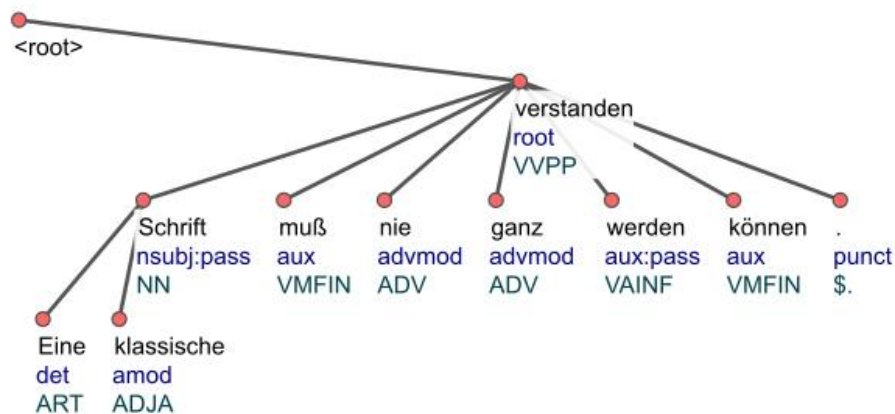


Figure 17 Dependency representation in tree-like form of the sentence "Eine klassische Schrift muß nie ganz verstanden werden können." from *Kritische Fragmente* by F. Schlegel, according to the UD 2.0 scheme.

sent_id = lyceum-f20-s1

⁷⁵ F. Schlegel, *Kritische Fragmente*, fragment 20.

⁷⁶ Friedrich Schlegel's *Lucinde and the Fragments*, University of Minnesota Press, 1971.

Intuitively, the modal verb *must* should depend on *können*, since it is the verb it refers to. However, modal verbs are considered auxiliaries in *UD*, i.e. function words, and one of the main principles of the UD scheme is that function words can never be heads. Therefore, when more auxiliaries occur together for the same predicate, they depend on the predicate through a flat structure. I therefore applied this principle to this situation, where both *and* and *must* and *können* depend on the predicate *verstanden* (Cf. 3.2).

3.3.1.9 Expletive Element, Coordination between Nominals

[114] **Es gibt so viele kritische Zeitschriften von verschiedener Natur und mancherlei Absichten!** Wenn sich doch auch einmal eine Gesellschaft der Art verbinden wollte, welche bloß den Zweck hätte, die Kritik selbst, die doch auch notwendig ist, allmählich zu realisieren.⁷⁷

114. **There are so many critical journals of varying sorts and differing intentions!** If only a society might be formed sometime with the sole purpose of gradually making criticism — since criticism is, after all, necessary — a real thing.⁷⁸

Es		gibt		so		viele		kritische		Zeitschriften		von		verschiedener		Natur
there		are		so		many		critical		journals		of		varying		sorts

und		mancherlei		Absichten!
and		different		intentions

⁷⁷ F. Schlegel, *Kritische Fragmente*, fragment 114.

⁷⁸ Friedrich Schlegel's *Lucinde and the Fragments*, University of Minnesota Press, 1971. ProQuest Ebook Central, <http://ebookcentral.proquest.com/lib/unibg-ebooks/detail.action?docID=345421>.

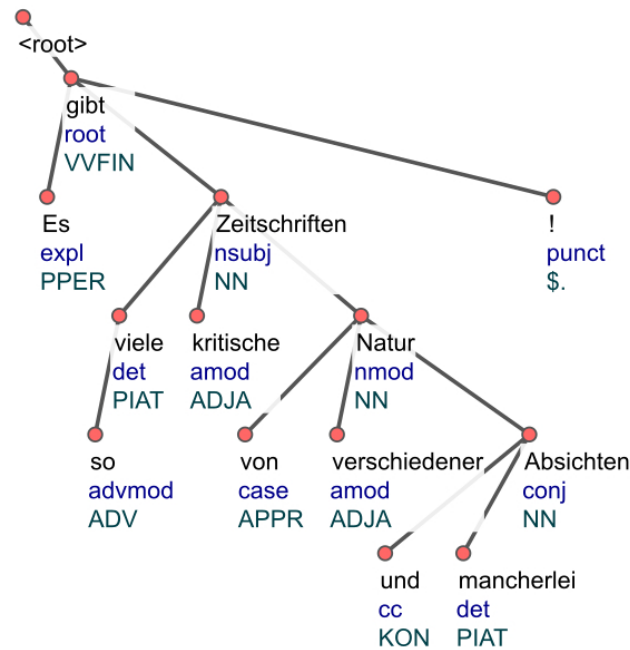


Figure 18 Dependency representation of the sentence "Es gibt so viele kritische Zeitschriften von verschiedener Natur und mancherlei Absichten!" from *Kritische Fragmente* by F. Schlegel, according to the UD 2.0 scheme.

sent_id = lyceum-fl14-s1

In Figure 19, there is a simple sentence where the main predicate is the finite verb *gibt*. In the first position, there is the personal pronoun *Es* (PPER, where P = pronoun, PER = personal), which plays the role of expletive element. By expletive elements, I mean nouns or pronouns that take a core-argument position, because they are syntactically required, but do not actually play any semantic role. For instance, let us consider the German verb *geben*, when used at the third person in the following impersonal construction:

Es		gibt		etwas/jemand [AKK]
There		is		Something/someone

In this construction, it always requires the neutral pronoun *Es* in the first position in the nominative case. But *es* is not the subject actually, exactly like *there* in the English construction *there is something/someone*. Conversely, the real subject follows the verb in third position in the accusative case, which is usually the case of the direct object. Therefore, in this verbal phrase, *es* fills the gap of a subject in nominative case, which is necessarily required by the German verbal syntax, but it does not any meaning on a logical and semantic level. In terms of valency, it is triggered by the morpho-syntactic valency of the verb rather than

by the logical or the semantic valency (Pittner and Berman 2015). Therefore, in this sentence, *Es* depends on the verb *gibt* through the *expl* relation (*expl* stands for expletive), while the real subject of *gibt* is the noun in accusative case *Zeitschriften*, which, therefore, depends on the predicate through the *nsubj* relation.

I focus now on the noun phrase involving the noun *Absichten*. This noun is connected through coordination to the previous noun phrase whose head is *Natur*, which, in turn, plays the role of noun modifier of the subject *Zeitschriften*. The noun phrase of *Absichten* is introduced by the coordinating conjunction *und* (KON, where KON = [coordinating] conjunction), then an adjectival modifier follows. When two lexical items are coordinated in this way, the coordinating conjunction depends forward on the second coordinated item, while, in turn, the second coordinated item depends back on the first item of the coordination. Therefore, here, *und* depends on *Absichten* through the *cc* relation (*cc* stands for coordinating conjunction), while the noun *Absichten* depends back on the first noun involved in the coordination *Natur* through the *conj* relation. The reason of this choice is that, in UD, coordination is dealt with asymmetrically, since there is hierarchy between the first conjunct and the successive conjuncts. This is clearly shown in the next example as well. The modifiers of the second coordinate item, if any, depend on this element regularly instead. Therefore, in this case, the indefinite pronoun in attributive function *mancherlei* (PIAT, where PI = indefinite pronoun, AT = attributive)⁷⁹ depends on the noun *Absichten* through the *det* relation.

3.3.1.10 Series of Coordinate Items

[4] Es gibt so viel Poesie, und doch ist nichts seltner als ein Poem! **Das macht die Menge von poetischen Skizzen, Studien, Fragmenten, Tendenzen, Ruinen, und Materialien**⁸⁰.

4. There is so much poetry and yet there is nothing more rare than a poem! **This is due to the vast quantity of poetical sketches, studies, fragments, tendencies, ruins, and raw materials**⁸¹.

⁷⁹ In STTS tag set, indefinite pronouns are classified according to two functions: attributive and substitutive function. They play an attributive function when they are used as noun modifiers, like in this case. They play a substitutive function when they actually replace nouns. The same for demonstrative pronouns. Other examples are reported in the successive dependency trees.

⁸⁰ F. Schlegel, *Kritische Fragmente*, fragment 4.

⁸¹ Friedrich Schlegel's *Lucinde and the Fragments*, University of Minnesota Press, 1971. ProQuest Ebook Central, <http://ebookcentral.proquest.com/lib/unibg-ebooks/detail.action?docID=345421>.

3.3.1.11 Copula-Like Verb

24. Selbstentäußerung ist die Quelle aller Erniedrigung, so wie im Gegentheil der Grund aller ächten Erhebung. **Der erste Schritt wird Blick nach Innen, absondernde Beschauung unsers Selbst.** Wer hier stehn bleibt, geräth nur halb. Der zweyte Schritt muß wirksamer Blick nach Außen, selbstthätige, gehaltne Beobachtung der Außenwelt seyn.⁸²

26. Sacrifice of the self is the source of all humiliation, as also on the contrary it is the foundation of all true exaltation. **The first step will be an inward gaze am isolating contemplation of ourselves.** Whoever stops here has come only halfway. The second step must be an active outward gaze-autonomous, constant observation as an artist who cannot depict anything other than his own experience.⁸³

Der	erste	Schritt	wird	Blick	nach	Innen,	absondernde
The	first	step	will be	gaze	toward	inward	isolating

Bschauung	unsers	Selbst.
contemplation	of our	self.

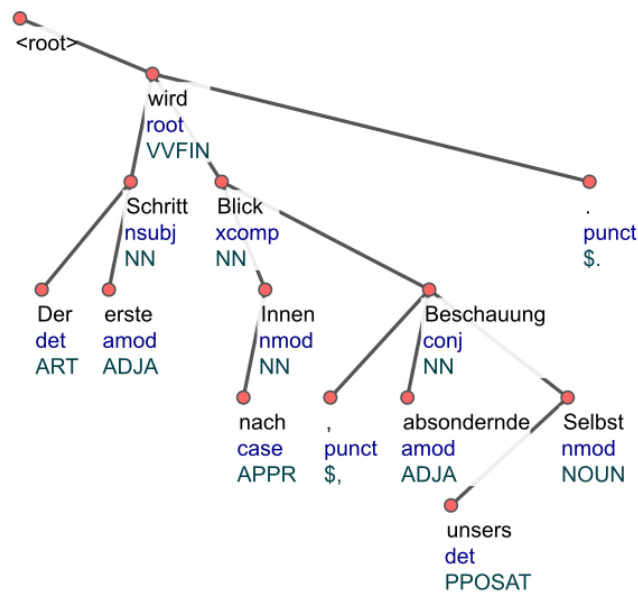


Figure 20 Dependency representation in tree-like form of the sentence “Der erste Schritt wird Blick nach Innen, absondernde Beschauung unsers Selbst.” from *Blüthenstaub* by Novalis, according to the UD 2.0 scheme.

sent id = bluethenstaub-f24-s2

⁸² Novalis, *Blüthenstaub*, fragment 24.

⁸³ STOLJAR, Margaret Mahony, et al. (ed.). Novalis: Philosophical Writings. SUNY Press, 1997.

In Figure 21, there is a simple sentence, where the predicate is the intransitive verb *werden*, ('to become'). Verbs like this cannot have a direct object as core argument, but they inherently require a predicative element, which could be either a noun, as in this case, or an adjective. For instance, in the following fragment, the same verb *werden* requires an adjective playing the predicative role:

Daher erscheint das Schöne so ruhig. Alles Schöne ist ein selbsterleuchtetes, vollendetes Individuum. Jede Menschengestalt belebt einen individuellen Keim in dem Betrachtenden. **Dadurch wird diese Anschauung unendlich**, sie ist mit dem Gefühl einer unerschöpflichen Kraft verbunden, und darum so absolut belebend .

Indeed, these types of verbs work like copulas, since they require a predicative element, either a noun or an adjective, to be fulfilled. This is considered a form of secondary predication, where the copula-like verb is the first predicate, while the predicative element is the second predicate. In these cases, the second predicate depends on the first predicate through the *xcomp* relation. For other instances of secondary predication, 3.3.2 and 3.3.3. Therefore, here, the noun *Blick* depends on *wird* through the *xcomp* relation. Then, the second predicate is followed by a noun phrase with *Innen*. I chose to let it depend on *Blick* as noun modifier, because such a phrase is triggered by the noun *Blick*, not by the verb *werden*. Therefore, I adopted the same criterion applied in see above.

As for the noun *Beschauung*, it is linked to the noun *Blick* through asyndetic coordination, i.e. without any explicit coordinating conjunction. Therefore, it depends on *Blick* through the *conj* relation. In this case, even if there is no explicit conjunction, we cannot assign the relation *parataxis*, since coordination does not take place between two clauses, but between elements inside the same clause. Finally, the possessive pronoun in attributive function (PPOSAT, where PPOS = possessive pronoun, AT = attributive) modify the noun *Selbst*, therefore it depends on it as determiner. In turn, *Selbst* is a noun modifier of the noun *Beschauung*.

3.3.1.12 Nominal Sentence, Adverb Coordinated to a Verb

12. Wunder stehn mit naturgesetzlichen Wirkungen in Wechsel: sie beschränken einander gegenseitig, und machen zusammen ein Ganzes aus. Sie sind vereinigt, indem sie sich gegenseitig aufheben. Kein Wunder ohne Naturbegebenheit und umgekehrt.

13. Miracles alternate with the effects of natural laws – they each limit the other, and together they constitute a whole. They are united in that they complement each other. There is no miracle without a natural event and viceversa.

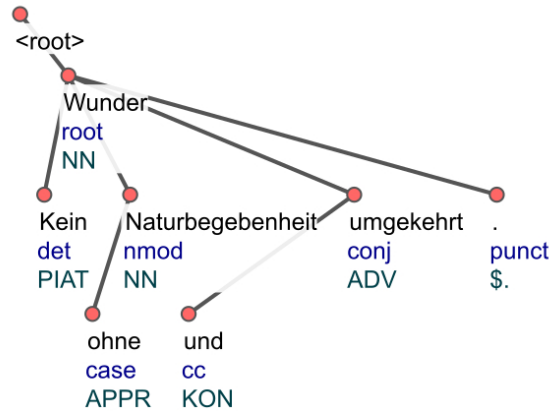


Figure 21 Dependency representation in tree-like form of the sentence "Kein Wunder ohne Naturbegebenheit, und umgekehrt." from *Blüthenstaub* by Novalis, according to UD 2.0.

sent_id = bluetenstaub-fl2-s3

In Figure 22, there is a simple nominal sentence, where the predicate is missing. In particular, it consists of two noun phrases and one adverbial phrase. To assign the root node to the sentence, I reasoned as follows. The noun *Naturbegebenheit* modifies the noun *Wunder* with a role of specification, therefore it has the function of nominal modifier, and it cannot be the root node. As for the adverbial phrase of *umgekehrt*, it is introduced by a coordinating conjunction, therefore it must necessarily be linked to a higher parent node other than the root one. Consequently, I promoted the noun *Wunder* as main node of the sentence. It is modified by the indefinite pronoun in attributive function *Kein*, which plays the role of negative determiner. Therefore, it depends on *Wunder* through the *det* relation.

As for the adverbial phrase, finding the head of the adverb does not look trivial. Usually, coordinate items should share the same POS. In this case, there is an adverbial phrase following two noun phrases. Furthermore, adverbs usually modify verbs. Therefore, I considered the adverb *umgekehrt* as if the head of a new clause consisting in a single adverbial phrase. Indeed, this could be considered a form of ellipsis of the predicate in the second clause. The only element that can be promoted is therefore the adverb itself. In doing so, the adverb depends on the predicate of the main clause through the *conj* relation. Since the main predicate is actually missing and the main node is a noun, the adverb *umgekehrt* actually depends back on *Wunder* through the *conj* relation. This was an operational solution adopted to solve this

annotation problem. In any case, the theoretical status of this sentence as a complex sentence consisting in a nominal clause and an adverbial clause with a predicate ellipsis is open to discussion.

3.3.2 Relations between Clauses: Coordination

3.3.2.1 Syndetic Coordination Between Verbs

1. Wir suchen überall das Unbedingte, und finden immer nur Dinge⁸⁴.

1. We seek the absolute everywhere, and only ever find things⁸⁵.

Wir	suchen	überall	das	Unbedingte,	und	wir	finden	immer	nur	Dinge.
We	seek	everywhere	the	absolut,	and	we	find	ever	only	things.

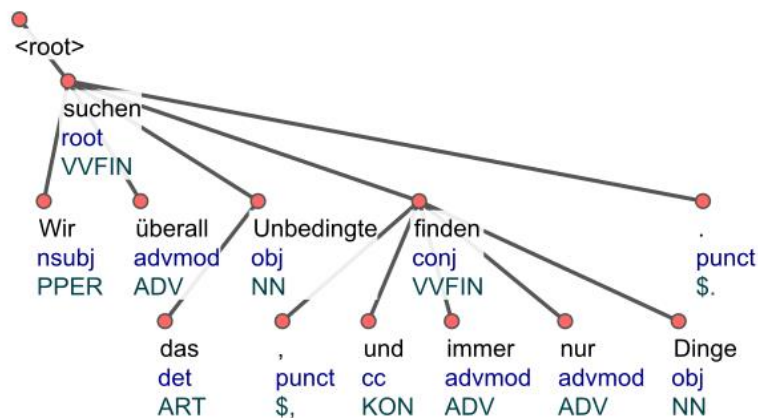


Figure 22 Dependency representation of the German sentence "Wir suchen überall das Unbedingte, und finden immer nur Dinge." from *Blüthenstaub* by Novalis, according to the UD 2.0 scheme.

sent_id = bluethenstaub-fl-s1

In Figure 23, there are two coordinate clauses. In this case, they are linked through an explicit coordinating conjunction, therefore the verb of the coordinate clause depends back on the verb of the main clause through the *conj* relation, while the coordinating conjunction *und* (KON) depends on the verb *finden*, i.e. the second element in the coordination, through the *cc* relation. As shown, the same

⁸⁴ Novalis, *Blüthenstaub*, fragment 1.

⁸⁵ STOLJAR, Margaret Mahony, et al. (ed.). Novalis: Philosophical Writings. SUNY Press, 1997.

fundamental rules for coordination between lexical items within the clause are applied to coordination between clauses as well. The rules are applied to the higher nodes of the two clauses, i.e. the heads of the two predicates. In this case, both predicates are verbal, therefore the relation goes from the subordinate verb to the main verb. In case of nominal predicates, the child node of the *advcl* relation is the predicative element. In this case, in the second clause, there are also two adverbs (ADV, where ADV = adverb): *immer* and *nur*. They depend on the verb through a flat structure, as shown above.

3.3.2.2 Asyndetic Coordination Between Verbs

13. Die Natur ist Feindin ewiger Besitzungen. Sie zerstört nach festen Gesetzen alle Zeichen des Eigenthums, vertilgt alle Merkmale der Formazion. **Allen Geschlechtern gehört die Erde; jeder hat Anspruch auf alles.** Die Früheren dürfen diesem Primogeniturzufalle keinen Vorzug verdanken. ⁸⁶[...]

13. Nature is the enemy of eternal possession. It destroys all signs of property according to the fixed laws, it eradicates all marks of formation. **The earth belongs to all generations – each person has a claim to everything.** Those born earlier may owe no advantage to the chance of process. The right to property is extinguished at certain times.

Allen	Geschlechtern	gehört	die	Erde;	jeder	hat	Anspruch	auf	Alles.
To all	genders	belongs	the	earth	everyone	has	claim	from	Everything.
the									

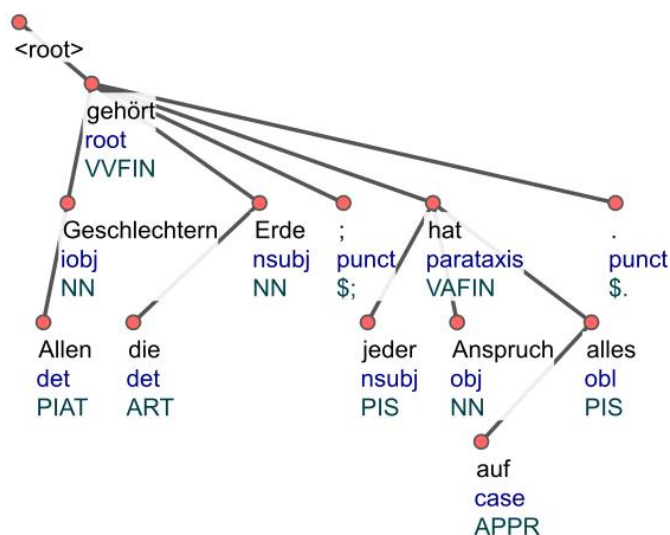


Figure 23 Dependency representation in tree-like form of the sentence “Allen Geschlechtern gehört die Erde; jeder hat Anspruch auf alles”, from *Blüthenstaub* by Novalis, according to UD 2.0 scheme.

sent_id = blüthenstaub-f13-s3

⁸⁶ Novalis, *Blüthenstaub*, fragment 13.

In Figure 24, there are two coordinate clauses. Unlike the sentence in Figure 23, the clauses are linked through asyndetic coordination, i.e. there is no explicit coordinating item between them, such as the coordinating conjunctions *und*, *aber* or *oder*. In fact, the two clauses simply stand next to each other, linked by the presence of the punctuation mark. In this case, unlike in the sentence in Figure 23, the verb of the coordinate clause depends back on the main verb through the *parataxis* relation⁸⁷. The nominal subject of the main verb *gehört* is the noun *Erde*, which occupies the post-verbal position. The noun *Geschlechtern*, which is the head of the first noun phrase of the main clause, is in dative case, and plays the role of indirect object of the main verb. Therefore, it depends on *gehört* through the *iobj* relation.

3.3.3 Relations between Clauses: Subordination

3.3.3.1 Relative Clause (with Secondary Predication in the Main Clause)

[1] **Man nennt viele Künstler, die eigentlich Kunstwerke der Natur sind.**⁸⁸

1. **Many so-called artists are really products of nature's art.**⁸⁹

Man (We)	nennt call	viele many	Künstler, artists,	die who	eigentlich really	Kunstwerke products	der of the	Natur nature	sind. are.
-------------	---------------	---------------	-----------------------	------------	----------------------	------------------------	---------------	-----------------	---------------

⁸⁷ This relation is used for other cases of coordination not mediated by any explicit coordinating element, such as in parenthetical clauses, or in dealing with the direct speech. More on these uses later in this paragraph..

⁸⁸ F. Schlegel, *Kritische Fragmente*, fragment 1.

⁸⁹ Friedrich Schlegel's *Lucinde and the Fragments*, University of Minnesota Press, 1971. ProQuest Ebook Central, <http://ebookcentral.proquest.com/lib/unibg-ebooks/detail.action?docID=345421>.

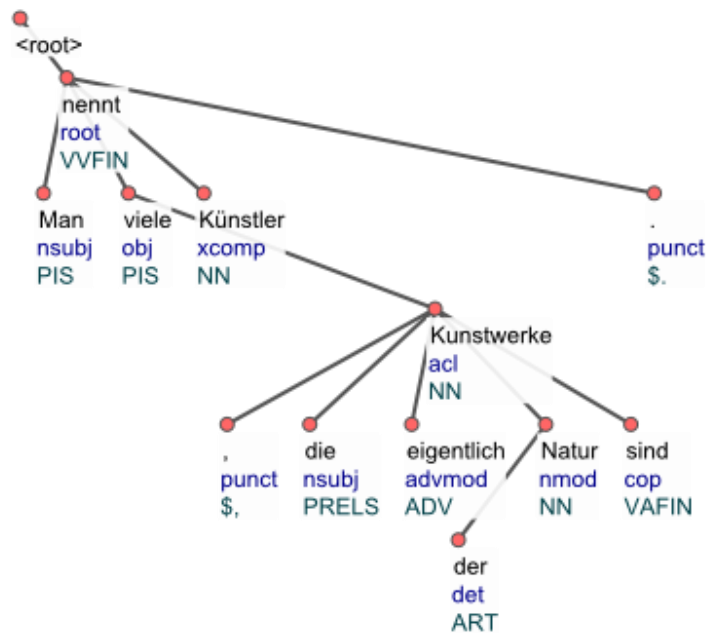


Figure 24 Dependency representation in tree-like form of the sentence "Man nennt viele Künstler, die eigentlich Kunstwerke der Natur sind." from *Kritische Fragmente* by F. Schlegel, according to the UD 2.0 scheme.

sent_id = lyceum-fl-s1

In Figure 25, there is a complex sentence consisting of a main clause and a subordinate clause. I first focus on the main clause. The main verb is the finite verb *nennt*. The nominal subject of *nennt* is the substitutive pronoun *Man* (PIS, where P = pronoun, I = indefinite, S = substitutive, that is it plays the role normally played by nouns), which, therefore, is the child of *nennt* through the *nsubj* relation. The direct object is the indefinite pronoun in substitutive function *viele*, which, therefore, is governed by *nennt* through the *obj* relation. Apparently, *viele* could be an indefinite adjective modifying the successive noun *Künstler*. This would be the correct syntactic construction of the sentence, if the noun *Künstler* were the direct object of the verb *nennen*. On the contrary, in this case, the noun *Künstler* has a predicative role. In fact, the best translation in English for the verb *nennen* in this particular context should be ‘to name’, meaning ‘we give this name [kueslter] to (many) people. Therefore, *viele* is a here substitutive pronoun replacing the noun *Lente* (‘people’). *Künstler* is therefore a core argument of the verb, since it is necessary to fulfill the semantic valency the verb, but it has a predicative function. Such a phenomenon is usually referred to as secondary predication (Rothstein 2013). When this phenomenon takes place, the predicative element depends on the verb through the *xcomp* relation. Therefore, *Künstler* depends on *nennt*.

through the *xcomp* relation (*xcomp* stands for open clausal complement)⁹⁰. In UD, this relation is used for all those predicative or clausal complement without their own subject.⁹¹

I now move to the subordinate clause. Specifically, it is a relative clause, whose nominal predicate modifies the pronoun *viele* in the main clause. It is introduced by the relative plural pronoun *die* (PRELS, where P = pronoun, RELS = relative), which refers back to *viele* in the main clause and plays the role of subject of the subordinate predicate. Since the predicate is not a verbal predicate but a nominal predicate, the pronoun *viele* depends on the predicative part of the subordinate predicate, therefore it depends on the noun *Kunstwerke* through the *nsubj* relation. Subordinate predicates modifying nouns or pronouns in the higher clause, i.e. the regent clause they refers to,⁹² depend on the element of the higher clause that they modify through the *acl* relation (*acl* stands for adjectival clause). The role of this syntactic function is clear: indeed, these verbs, which can be either finite or non-finite, modify nominals, i.e. both nouns and pronouns, which is a role prototypically played by adjectives. In this case, the predicate of the relative clause is a nominal predicate, where the nominal part is the noun *Kunstwerke*. Consequently, *Kunstwerke* depends back on *viele* in the main clause through the *acl* relation.

Finally, I deal with the punctuation preceding the subordinate clause. In German, relative clauses are always separated from the main clauses through a comma. According to the UD scheme, a comma preceding and following a subordinate clause should depend on the predicate of the subordinate clause. Therefore, in this case, the comma (,) preceding the relative clause depends on the noun *Kunstwerke*.

3.3.3.2 Subordinating Conjunction, Adverbial Clause

12. Wunder stehn mit naturgesetzlichen Wirkungen in Wechsel: sie beschränken einander gegenseitig, und machen zusammen ein Ganzes aus. **Sie sind vereinigt, indem sie sich gegenseitig aufheben.** Kein Wunder ohne Naturbegebenheit und umgekehrt.⁹³

13. Miracles alternate with the effects of the natural laws – they each limit the other, and together they constitute a whole. **They are united in that they complement each other.** There is no miracle without a natural event and vice versa.⁹⁴

⁹⁰ The label *xcomp* originally comes from the Lexical Functional Grammar.

⁹¹ For a detailed explanation of the usage of this relation, see <https://universaldependencies.org/it/dep/xcomp.html>.

⁹² We cannot define it as a *main clause*, since it could be a subordinate clause modified by another subordinate clause. An example is reported in 2.9.15.

⁹³ Novalis, *Blüthenstaub*, fragment 12.

⁹⁴ STOLJAR, Margaret Mahony, et al. (ed.). Novalis: Philosophical Writings. SUNY Press, 1997.

Sie They	sind are	vereinigt, united,	indem in that	sie they	sich (themselves)	gegenseitig each other	aufheben. complement.
-------------	-------------	-----------------------	------------------	-------------	----------------------	---------------------------	--------------------------

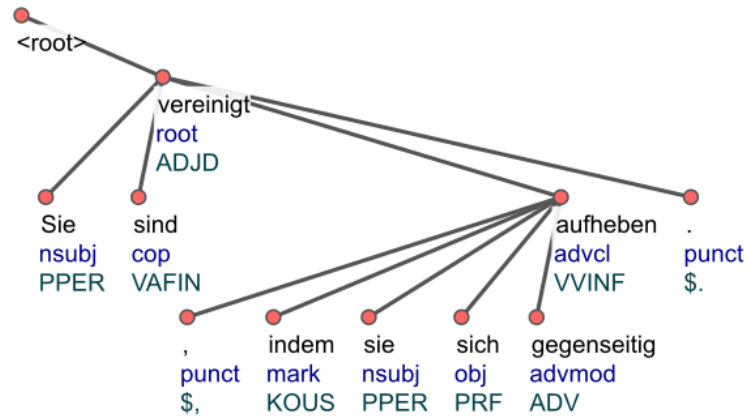


Figure 25 Dependency representation of the sentence "Sie sind vereinigt, indem sie sich gegenseitig aufheben" from *Blütenstaub* by Novalis, according to the UD 2.0 scheme.

sent_id = bluetenstaub-f12-s2

In Figure 26, there is a main clause followed by a subordinate clause. The main clause has a nominal predicate, where the predicative role is played by the adjective *vereinigt* (ADJD, where ADJ = adjective, D = predicative), which, therefore, is the root node of the whole clause as well as head of the copula *sind* (VAFIN). The subordinate clause is an adverbial clause. Unlike adjectival clauses, adverbial clauses modify predicates, not nominals within the highest clause. That is the reason why they are said adverbial, because they play the role which is usually played by adverbs in clauses. In German, they are always introduced by subordinating markers that have the role of subordinating conjunctions, such as *indem* (KOUS, where KOU = conjunction, S = subordinating) in this case. In addition, they are always finite and the verb always occupy the last position of the clause. In this kind of clauses, the subordinate verb depends on the verb of the higher clause through the *advcl* relation, where *advcl* stands for adverbial clause), while the subordinating conjunction depends on the subordinate verb through the *mark* relation (mark stands for marker, that is an element signalling the presence of a finite subordinate verb in the clause). An adverbial clause can also modify a verb of another subordinate clause. Let us consider the following sentence:

Darwin macht die Bemerkung, daß wir weniger vom Lichte beym Erwachen geblendet werden, wenn wir von sichtbaren Gegenständen geträumt haben.⁹⁵

⁹⁵ Novalis, *Blütenstaub*, fragment 17.

Darwin makes the observation that we are less dazzled by the light on walking – if we have been dreaming of visible objects.⁹⁶

In this example, we have a series of two adverbial clauses, where the second one modifies the first one. In a dependency representation, the verb of the second subordinate clause *geträumt* would depend back on the predicate of the second clause, i.e. *geblendet*, through the *advcl* relation. In turn, *geblendet* would be governed by the predicate of the main clause, i.e. *macht*, through the *advcl* relation too. Let us now go back to the example in Figure 26. The subordinate finite verb *aufheben* depends on *vereinigt* through the *advcl* relation, while the subordinating conjunction *indem* depends on *aufheben* through the *mark* relation. The subject of the adverbial clause is the personal pronoun *sie* (PPER, where PP = pronoun, PER = personal). Finally, the comma introducing (or following) the subordinate clause should always depend on the predicate of the subordinate clause. Therefore, it depends on *aufheben*.

3.3.3.3 Non-Finite Subordinate Clause

[16] Genie ist zwar nicht Sache der Willkür aber doch der Freiheit, wie Witz, Liebe und Glauben, die einst Künste und Wissenschaften werden müssen. **Man soll von jedermann Genie fordern, aber ohne es zu erwarten.** Ein Kantianer würde dies den kategorischen Imperativ der Genialität nennen.⁹⁷

16. Through genius isn't something that can be produced arbitrarily, it is freely willed — like wit, love, and faith, which one day will have to become arts and sciences. **You should demand genius from everyone, but not expect it.** A Kantian would call this the categorical imperative of genius.⁹⁸

Man	soll	von	jedermann	Genie	fordern,	aber	ohne	es	zu	erwarten.
(you)	should	from	everyone	Genius	demand,	but	without	it	to	expect.

⁹⁶ STOLJAR, Margaret Mahony, et al. (ed.). Novalis: Philosophical Writings. SUNY Press, 1997.

⁹⁷ F. Schlegel, *Kritische Fragmente*, fragment 16.

⁹⁸ Friedrich Schlegel's Lucinde and the Fragments, University of Minnesota Press, 1971. ProQuest Ebook Central, <http://ebookcentral.proquest.com/lib/unibg-ebooks/detail.action?docID=345421>.

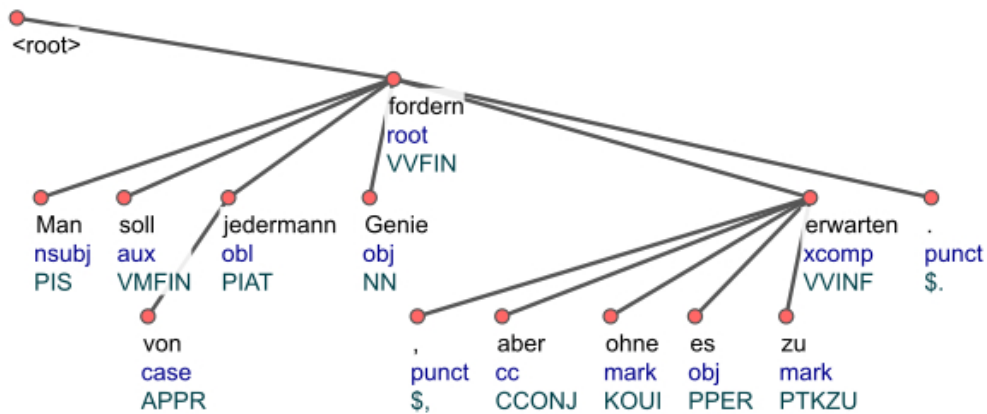


Figure 26 Dependency representation in tree-like form of the sentence “Man soll von jedermann Genie fordern, aber ohne es zu erwarten” from *Kritische Fragmente* by F. Schlegel, according to the UD 2.0 scheme.

sent_id = lyceum-fl6-s2

In Figure 27, the main clause is followed non-finite subordinate clause. I now focus on the subordinate clause, since the structure of the main clause is clear and all the relations occurring there have already been explained in previous subparagraphs.

The subordinate non-finite verb is *erwarten*, which occupies the last position of the clause. Non-finite subordinate verbs depend on the verb of the higher clause through the *xcomp* relation, since these verbs do not have their own subject in the subordinate clause. In fact, they can be considered as open clausal complements of a higher verbal node. In this case, *erwarten* depends on the main predicate of the sentence *fordern* through the *xcomp* relation. The subordinate clause is introduced by the subordinating conjunction *ohne*, which therefore depends on *erwarten* through the *mark* relation. The personal pronoun *es* plays the role of direct object of the non-finite clause. The comma preceding the subordinate clause, which is necessary in the written German language to introduce a non-finite clause, depends on the verb of the subordinate clause through the *punct* relation. Finally, it is worth focusing on *aber*. It is usually used as a coordinating conjunction with adversative function. It can link two nominals, as in the following example:

[25] Die griechische Mythologie ist zu dem Theil eine solche Übersetzung einer Nationalreligion . Auch die moderne Madonna ist ein solcher Mythos . Grammatische Übersetzungen sind die Übersetzungen in dem gewöhnlichen Sinn. **Sie erfordern sehr viel Gelehrsamkeit, aber nur diskursive Fähigkeiten** ⁹⁹.

⁹⁹ F. Schlegel, *Athenäums-Fragmente*, fragment 25.

In this case, *aber* would depend on the second item of the coordination, i.e. the noun *Fähigkeiten*, which, in turn, would depend back on the noun *Gelehrsamkeit*. Alternatively, *aber* can link two coordinate predicates, as in the following example:

[...]Die Verwornen haben in dem Anfang mit mächtigen Hindernissen zu kämpfen, sie dringennur langsam ein, sie lernen mit Mühe arbeiten : dann aber sind sie auch Herrn und Meister auf immer . Der Geordnete kommt geschwind hinein, aber auch geschwind heraus . **Er erreicht bald die zweyte Stufe : aber da bleibt er auch gewöhnlich stehn** .[...]¹⁰⁰

In this second case, *aber* would depend on the verb *bleibt* through the *cc* relation, while, in turn, the verb *bleibt* would be the child node of the verb *erreicht* through the relation *conj*. The situation in Figure 26 is different though, since the coordinating conjunction depends on a verb, which, on the contrary, depends back on the verb of the main clause as a subordinate verb, not as a coordinate verb. At the same time, there are no alternatives to annotate with a function other than *cc*, since it cannot here be considered nor an adverb nor any other POS indeed. Therefore, even if it is not followed by any coordinate clause, I opted to let it depend on the non-finite verb as coordinating conjunction through the *cc* relation.

3.3.3.4 Non-Finite Subordinate Clause with Auxiliary

[37] **Um über einen Gegenstand gut schreiben zu können, muß man sich nicht mehr für ihn interessieren;** der Gedanke, den man mit Besonnenheit ausdrücken soll, muß schon gänzlich vorbei sein, einen nicht mehr eigentlich beschäftigen. [...] ¹⁰¹

37. **In order to write well about something, one shouldn't be interested in it any longer.** To express an idea with due circumspection, one must have relegated it wholly to one's past; one must no longer be preoccupied with it. ¹⁰²

¹⁰⁰ Novalis, *Blütenstaub*, fragment 54.

¹⁰¹ F. Schlegel, *Kritische Fragmente*, fragment 37.

¹⁰² Friedrich Schlegel's *Lucinde and the Fragments*, University of Minnesota Press, 1971. ProQuest Ebook Central, <http://ebookcentral.proquest.com/lib/unibg-ebooks/detail.action?docID=345421>.

Um	über	einen	Gegenstand	gut	schreiben	zu	können,	muss	man
to	on	an	object	well	write	to	can,	should	one
sich	nicht	mehr	für	ihn	interessieren.				
himself	not	anymore	for	that	be interested				

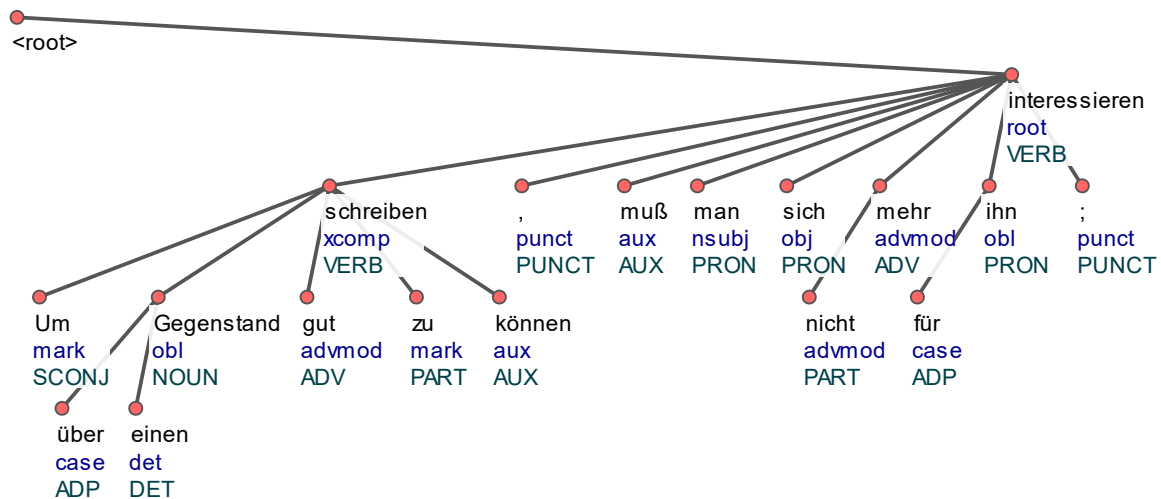


Figure 27 Dependency representation in tree-like form of the sentence “Um über einen Gegenstand gut schreiben zu können, muß man sich nicht mehr für ihn interessieren.” from *Kritische Fragmente* by F. Schlegel, according to the UD 2.0 scheme.

sent_d = lyceum-f37-s1

In Figure 28, there is a non-finite subordinate clause preceding the main clause. The subordinate clause is a final clause, introduced by the subordinating conjunction *um*, which is used as marker for final clauses in German. Therefore, *um* depends on the subordinate non-finite verb *schreiben* through the *mark* relation. In turn, *schreiben* depends forward on the predicate of the main clause, which is the non-finite verb *interessieren*, which occurs at the end of the clause, because of the presence of the auxiliary verb *muss* in the canonical second position. In addition, in German, non-finite subordinate verbs are always introduced by the particle *zu*, which, in this case, has the function of marker of subordination. Therefore, it depends on the non-finite verb through the *mark* relation, together with *um*. In the sentence reported in Figure 11, the final clause has a modal verb modifying the non-finite verb. When this type of construction occurs, the modal verb stands at the end of the clause, since it is the non-finite verb of the final clause actually. Conversely, the verb that the modal verb modifies occupies the penultimate position in the clause. The particle *zu* (PTKZU, where PTK = particle, ZU = *zu* for infinitive), which has the function of marker of subordination, depends on the non-finite verb *schreiben*. Intuitively, it should depend on the

non-finite modal verb *können*. However, one of the fundamental principles is that auxiliaries can never be heads. Therefore, the marker *zu* depends on the predicate to which the non-finite modal verb refers. The verb *schreiben* depends forward on the main predicate *interessieren* through the *xcomp* relation, since it is the real predicate of the final clause. In the main clause, it is worth noting the role of the particle *nicht* (PTKNEG, where PTK = particle, NEG = negation), which plays the role of negation marker, exactly like *not* in English. In this case, it modifies the adverb *mehr*, therefore it depends on it through the *advmod* relation. This relation is always assigned to the occurrences of *nicht*, since it is a particle whose syntactic function here is to modify the meaning of both verbs and adverbs.

3.3.3.5 Clausal Subject, Verb Followed by Bare Infinitive

24. Selbstentäußerung ist die Quelle aller Erniedrigung, so wie im Gegentheil der Grund aller ächten Erhebung. Der erste Schritt wird Blick nach Innen, absondernde Beschauung unsers Selbst. **Wer hier stehn bleibt, geräth nur halb.** Der zweyte Schritt muß wirksamer Blick nach Außen, selbstthätige, gehaltne Beobachtung der Außenwelt seyn.¹⁰³

26. Sacrifice of the self is the source of all humiliation, as also on the contrary it is the foundation of all true exaltation. The first step will be an inward gaze an isolating contemplation of ourselves. **Whoever stops here has come only halfway.** The second step must be an active outward gaze- autonomous, constant observation as an artist who cannot depict anything other than his own experience.¹⁰⁴

Wer		hier		stehn		bleibt,		geräth		nur		halb.
Whoever		here				stops		has come		only		halfway

¹⁰³ Novalis, *Blütenstaub*, fragment 24.

¹⁰⁴ STOLJAR, Margaret Mahony, et al. (ed.). Novalis: Philosophical Writings. SUNY Press, 1997.

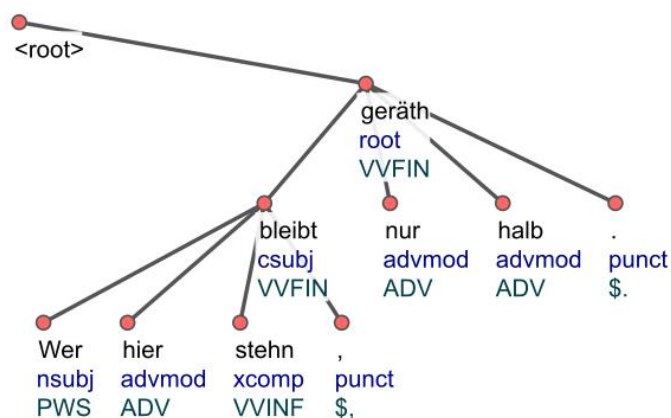


Figure 28 Dependency representation in tree-like form of the sentence “Wer hier stehn bleibt, geräth nur halb.” from Blütenstaub by Novalis, according to UD 2.0.

sent_id = bluethenstaub-f24-s3

In Figure 29, there is a complex sentence, where a subordinate clause precedes the main clause. The main clause has no nominal subject. Indeed, the whole subordinate clause plays the role of subject, therefore is a clausal subject. Consequently, the subordinate predicate *bleibt* depends forward on the main predicate, i.e. the verb *geräth*, through the *csubj* relation. Within the subordinate clause, the interrogative pronoun in substitutive function *Wer* plays the role of subject (PWS, where PW = interrogative pronoun, S = substitutive)¹⁰⁵. Then, there is the bare-infinitive verb *stehen* preceding the real verb of the clause *bleibt*. In these verbal phrases, which occur in other frequent verbal constructions in German such as *kennen lernen*, or *etwas tun lassen*, the bare-infinitive verb depends on the other verb through the *xcomp* relation. Therefore, here, *stehen* depends on *bleibt* through the *xcomp* relation.

3.3.3.6 Verb Followed by Infinitive with *Zu*

[49] Eins der wichtigsten Moyens der dramatischen und romantischen Kunst bei den Engländern sind die Guineen. **Besonders in der Schlußcadence werden sie stark gebraucht, wenn die Bässe anfangen recht voll zu arbeiten.**¹⁰⁶

¹⁰⁵ In STTS, it is also assigned to this kind of pronouns when they do not introduce any interrogative clause, as in this case.

¹⁰⁶ F Schlegel, *Kritische Fragmente*, fragment 49.

49. One of the most important techniques of the English drama and novel is guineas. **They're used a great deal especially in the final cadenza when the bass instruments begin to have hard work of it.**¹⁰⁷

In Figure 30, there is a complex sentence consisting of a main clause in passive voice which precedes a subordinate adverbial clause. The adverbial clause is introduced by the coordinating conjunction *wenn*, while the subordinate verb is the finite verb *anfangen*, therefore it depends on the main predicate through the *advcl* relation.

Besonders	in	der	Schlußcadence	werden	sie	stark	gebraucht,
A great	in	the	final cadenza	are	they	frequently	used
deal							
wenn	die	Bässe	anfangen	recht	voll	zu	arbeiten.
when	the	bass instruments	begin	right	hard	to	work.

¹⁰⁷ Friedrich Schlegel's *Lucinde and the Fragments*, University of Minnesota Press, 1971. ProQuest Ebook Central, <http://ebookcentral.proquest.com/lib/unibg-ebooks/detail.action?docID=345421>.

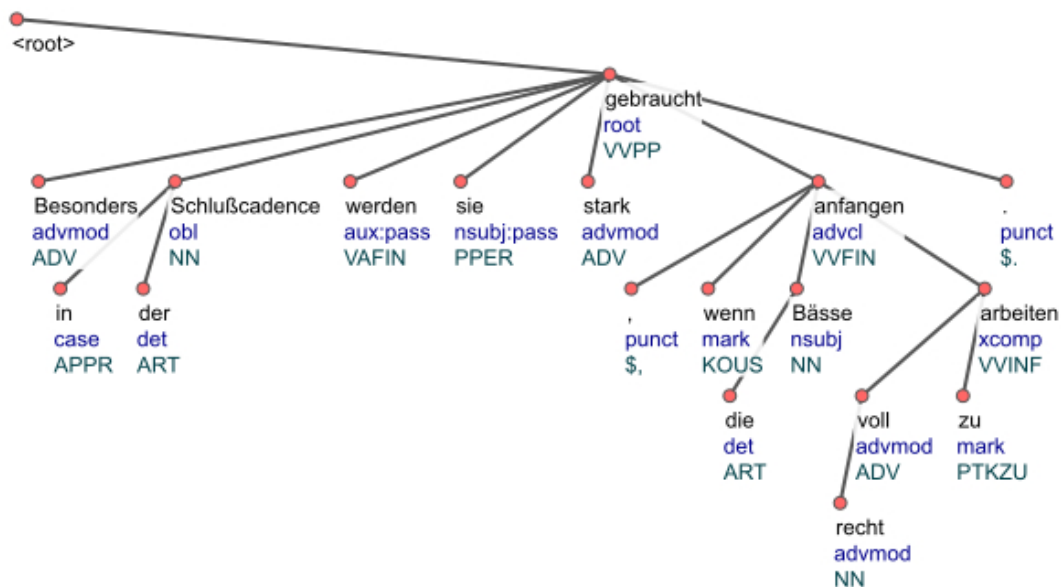


Figure 29 Dependency representation in tree-like form of the sentence "Besonders in der Schlußcadence werden sie stark gebraucht, wenn die Bässe anfangen recht voll zu arbeiten." from *Kritische Fragmente* by F. Schlegel, according to the UD 2.0 scheme.

sent_id = lyceum-f49-s2

The verb *anfangen* triggers a non-finite subordinate clause, in which the non-finite verb *anfangen* is preceded by the particle *zu*. The finite clause with the verb *anfangen* and the non-finite clause with the verb *arbeiten* share the same subject, which is the plural noun *Bässe*. When this construction occurs, the subject depends on the finite verb, while the non-finite verb depends on the finite verb through the *xcomp* relation. Actually, the clause whose predicate is *arbeiten* has a completive function for the verb *anfangen*, since it should be the clausal object. But, in this case, we cannot consider this clause as a clausal object, since it does not have its own subject. Conversely, the subject is controlled, it is the same of the higher clause with no other possible interpretation, since both *anfangen* and *arbeiten* share the same subject, i.e. *Bässe*.¹⁰⁸ Therefore, in this case, *Bässe* depends on *anfangen* through the *nsubj* relation, while *arbeiten* depends on *anfangen* through the *xcomp* relation. Moreover, the adverb *voll* modifies the non-finite verb, therefore it depends on it through the *advmod* relation.

¹⁰⁸ Cf. <https://universaldependencies.org/u/dep/ccomp.html>.

3.3.3.7 Clause as Predicative Part of a Nonverbal Predicate, (Parenthetical Clause)

[86] **Der Zweck der Kritik, sagt man, sei, Leser zu bilden!** – Wer gebildet sein will, mag sich doch selbst bilden. Dies ist unhöflich: es steht aber nicht zu ändern.¹⁰⁹

86. **The function of criticism, people say, is to educate one's readers!** Whoever wants to be educated, let him educate himself. This is rude: but it can't be helped.¹¹⁰

Der	Zweck	der	Kritik,	sagt	man,	sei	Leser	zu	bilden!
The	function	of	criticism,	say	(people)	is	readers	to	educate

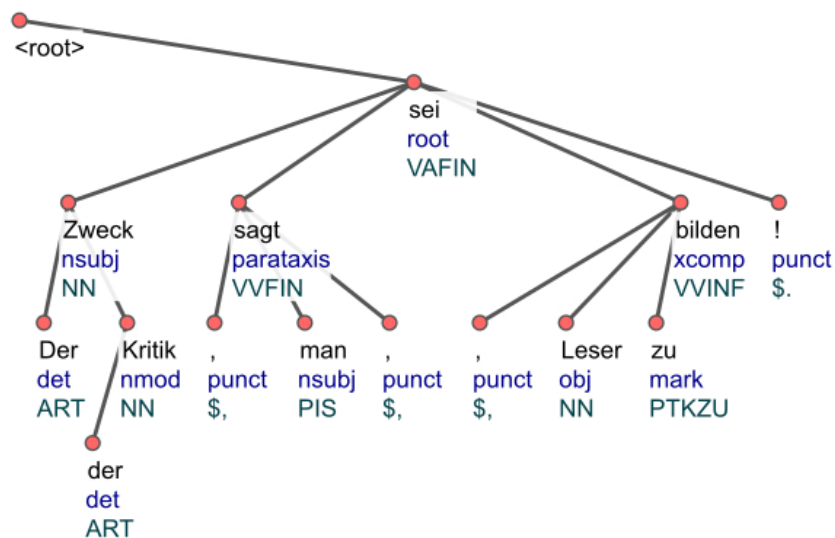


Figure 30 Dependency representation in tree-like form of the sentence "Der Zweck der Kritik, sagt man, sei, Leser zu bilden!" from *Kritische Fragmente* by F. Schlegel, according to UD 2.0.

sent_id = lyceum-f86-s1

In Figure 31, there is a complex sentence made up of three clauses. The main clause has a nominal predicate, in which the copula is *sei*, while the predicative part consists in a non-finite clause whose predicate is the verb *bilden*. Apparently, the predicative part of the nominal predicate should be considered as the main predicate of the sentence. Conversely, the copula verb is promoted as main node of the tree

¹⁰⁹ F. Schlegel, *Kritische Fragmente*, fragment 86.

¹¹⁰ Friedrich Schlegel's *Lucinde and the Fragments*, University of Minnesota Press, 1971. ProQuest Ebook Central, <http://ebookcentral.proquest.com/lib/unibg-ebooks/detail.action?docID=345421>.

rather than considering the subordinate predicate as main node. Such a solution is adopted for all those cases where the head of a copula should be a verb, as in this case. On the one hand, this is done to avoid the presence of false dependents with false functions in the surface structure of the sentence. For instance, in this case, if the verb *bilden* was promoted, the noun *Zweck* would play the role of nominal subject of the verb *bilden*. Indeed, this is unacceptable, since the verb pertains to a different clause. On the other hand, this is done to avoid the presence of two different subjects governed by the same verb. This problem arises in all those sentences where the subordinate predicate that is the predicative part of a copula in the main clause has its own subject.¹¹¹ Therefore, here, the copula *sei* depends on the fictional node through the root relation, while the non-finite verb *bilden* depends on *sei* through the *xcomp* relation. In fact, it cannot be considered a completive clause because the subject is missing. Finally, there is a third clause, which is the parenthetical clause, in which the verb *sagt* is the predicate, while the indefinite pronoun *man* is the subject. Predicates in parenthetical clauses depend on the predicate of the higher clause through the *parataxis* relation. Commas preceding and following the parenthetical clause depend both on the predicate of this clause.

3.3.3.8 Clausal Subject of a Nonverbal Predicate

[92] Auch der Geist kann, wie das Tier, nur in einer aus reiner Lebensluft und Azote gemischten Atmosphäre atmen. **Dies nicht ertragen und begreifen zu können, ist das Wesen der Torheit;** es schlechthin nicht zu wollen, der Anfang der Narrheit.¹¹²

92. Like animals, the spirit can only breathe in an atmosphere made up of life-giving oxygen mixed with nitrogen. **To be unable to tolerate and understand this fact is the essence of foolishness;** to simply not want to do so, is the beginning of madness.¹¹³

Dies	nicht	ertragen	und	begreifen	zu	können,	ist	das	Wesen	der
this	not	tolerate	and	understand	to	can	is	the	essence	of

Torheit.
Foolishness.

¹¹¹ Cf. <https://universaldependencies.org/u/dep/ccomp.html#ccomp-clausal-complement>

¹¹² F. Schlegel, *Kritische Frgamente*, fragment 92.

¹¹³ Friedrich Schlegel's *Lucinde and the Fragments*, University of Minnesota Press, 1971. ProQuest Ebook Central, <http://ebookcentral.proquest.com/lib/unibg-ebooks/detail.action?docID=345421>.

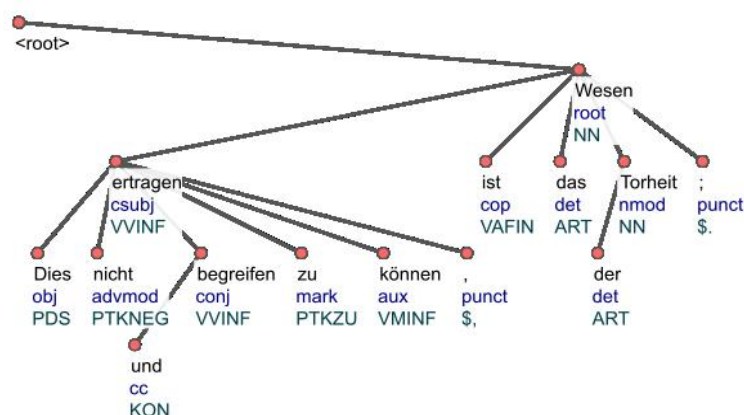


Figure 31 Dependency representation in tree-like form of the sentence “Dies nicht ertragen und begreifen zu können, ist das Wesen der Torheit;” from *Kritische Fragmente* by F. Schlegel, according to the UD 2.0 scheme.

sent_id = lyceum-f92-s2

In Figure 32, there is a complex sentence consisting of a subordinate clause preceding the main clause. The main predicate is a nominal predicate, whose predicative part is the noun *Wesen*. In the main clause, there is no nominal subject, since the role of subject is played by the whole non-finite subordinate clause actually. It is worth noting the difference with the sentence previously illustrated in Figure 31, in which the situation was reverse. In fact, in that case, the clause played the predicative role, which, on the contrary, is here played by *Wesen*. Therefore, in this case, the predicate of the subordinate clause plays the role of subject of the main clause. In particular the clause working functioning as clausal subject is a non-finite clause, where the predicate is the verb *ergrreifen*. Actually, the non-finite verb triggering the main clause is the modal verb *können*, which is preceded by the particle *zu*, and occupies the last position in the clause. Despite this, modal verbs (and auxiliary verbs in general) must always depend on the non-finite verb they refer to, therefore the subordinate predicate is *ertragen* in truth. For the same reason, the subordinating marker *zu* depends on the actual predicate, not on the modal verb. At the same time, there is a verb coordinated to the subordinate predicate – *begrreifen* – which depends on *ertragen* through the *conj* relation. In turn, *ertragen* depends on the main predicate through the *csubj* relation (*csubj* stands for clausal subject), because of its role of clausal subject, as explained above.

3.3.3.9 Clausal Complement, Adjectival Clause Modifying a Pronoun, (*Etwas* Followed by a Substantive)

[47] **Wer etwas Unendliches will, der weiß nicht was er will.** Aber umkehren läßt sich dieser Satz nicht.¹¹⁴

47. **Whoever desires the infinite doesn't know what he desires.** But one can't turn this sentence around.¹¹⁵

Wer	etwas	Unendliches	will,	der	weiss	nicht	was	er	will.
Who	something	eternal	desire,	he	know	not	what	he	desire.

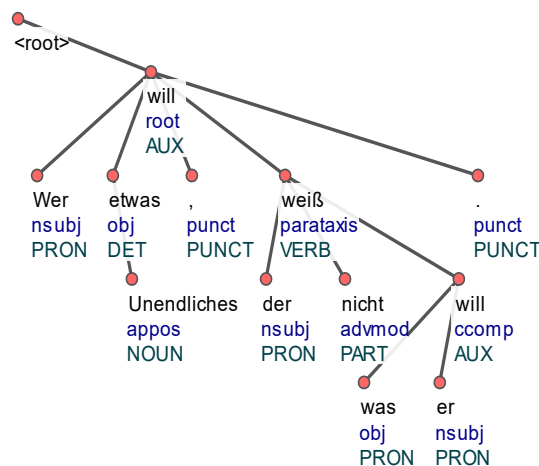


Figure 32 Dependency representation in tree-like form of the sentence "Wer etwas Unendliches will, der weiß nicht was er will." from *Kritische Fragmente* by F. Schlegel, according to the UD 2.0 scheme.

sent_id = lyceum-f47-s1

In Figure 33, there is a complex sentence made up of three clauses. In the first clause, the interrogative pronoun *Wer* plays the role of subject, while the indefinite pronoun in substitutive function *etwas* is the direct object. In this case, this pronoun is used to mark indefiniteness, but in substitutive function (PIS) rather than in attributive function. In fact, it plays the role of direct object of the verb *will*, and it is

¹¹⁴ F. Schlegel, *Kritische Fragmente*, fragment 47.

¹¹⁵ Friedrich Schlegel's *Lucinde and the Fragments*, University of Minnesota Press, 1971. ProQuest Ebook Central, <http://ebookcentral.proquest.com/lib/unibg-ebooks/detail.action?docID=345421>.

modified by the substantive *Unendliches*, which has a function of specification. Such construction is common in German. It resembles the English construction [*something*] + [adjective], which occurs, for instance, in the following phrases: *something wrong*, *something great*, *something spectacular*, *something eternal*. In German, the items modifying the value of indefiniteness expressed by the indefinite pronoun become substantives, as in the following examples: *etwas Schlechtes* ('something wrong'), *etwas Grosses* ('something big'), *etwas Wunderbares* ('something wonderful'). Consequently, I opted for annotating *etwas* as direct object, whereas I let *Unedliches* depend on *etwas* through the *nmod* relation.

After the clause whose predicate is *will*, a second clause follows, introduced by the demonstrative pronouns *der*, whose predicate is the finite verb *weiss*. Apparently, the first clause could be the clausal subject of the second clause. Conversely, the second clause has its own nominal subject, which is the personal pronoun *der* indeed. In this case, it is a back reference to the preceding clause, as if it encapsulates the whole meaning of the first clause. It is not an expletive element, since it is the actual subject, both syntactically and semantically.

Finally, there is a third clause, whose predicate is *will*, used as non-modal verb again. For pure stylistic reasons, there is no comma preceding this clause, which is usually required in German before subordinate clauses. In any case, the interrogative pronoun *was* plays the role of direct object, while the personal pronoun *er* is the subject. In the main clause, there is no object, which is played by the entire clause, which is therefore a completive clause. Therefore, the subordinate predicate *will* depends back on the higher predicate, which is *weiss*, through the *ccomp* relation (*ccomp* stands for clausal complement).

3.3.4 Comparative

3.3.4.1 Comparative: *Als* as Subordination Marker

103. **Manche Bücher sind länger als sie scheinen.** Sie haben in der That kein Ende. Die Langeweile die sie erregen, ist wahrhaft absolut und unendlich. [...] ¹¹⁶

103. **Many books are longer than they seem.** They have indeed no end. The boredom that they cause is truly absolute and infinite. ¹¹⁷[...]

¹¹⁶ Novalis, *Blütenstaub*, fragment 103.

¹¹⁷ STOLJAR, Margaret Mahony, et al. (ed.). Novalis: Philosophical Writings. SUNY Press, 1997.

Manche	Bücher	sind	länger	als	sie	scheinen.
some	books	are	longer	than	they	seem

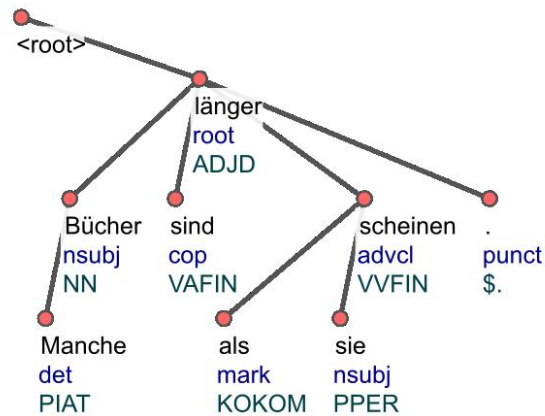


Figure 33 Dependency representation in tree-like form of the German sentence "Manche Bücher sind länger als sie scheinen." from *Blüthenstaub* by Novalis, according to the UD 2.0 scheme.

sent_id = bluethenstaub-f103-s1

In Figure 34, there is a complex sentence, made up of a main clause and a subordinate clause. In the main clause, the predicative adjective *länger* is morphologically marked (ä + -er) to introduce an inequality comparison, specifically a scalar increasing comparison, where the first adjective conveys a higher degree in quality compared with a second element, which, in turn, plays the role the standard of comparison. This second element can be a noun phrase, an adjective, an adverb or a clause. In this case, it is a clause introduced by *als*, which therefore has the function of comparative conjunction (KOKOM, where KO = conjunction, KOM = comparative). In particular, it occurs as comparative conjunction in subordinating role. Therefore, it depends on the subordinate verb *scheinen* through the *mark* relation. When this kind of comparison takes place, in which the standard of comparison is encoded in a subordinate clause, the clause is treated as an adverbial clause modifying the adjective (or the adverb) introducing the comparison in the higher clause. In fact, it is this adjective or adverb which triggers the comparative clause, both syntactically and semantically, while none of the other elements of the first clause has any syntactic relation with the comparative clause.

3.3.4.2 Comparative: *Als* as Comparative Conjunction Introducing a Noun Phrase

[4] **Es gibt so viel Poesie, und doch ist nichts seltner als ein Poem!** Das macht die Menge von poetischen Skizzen, Studien, Fragmenten, Tendenzen, Ruinen, und Materialien.¹¹⁸

4. **There is so much poetry and yet there is nothing more rare than a poem!** This is due to the vast quantity of poetical sketches, studies, fragments, tendencies, ruins, and raw materials.¹¹⁹

Es	gibt	so	viel	Poesie,	und	doch	ist	nichts	seltner
there	is	so	much	Poetry,	and	yet	is	nothing	More
									rare

als	Ein	Poem!
than	a	Poem

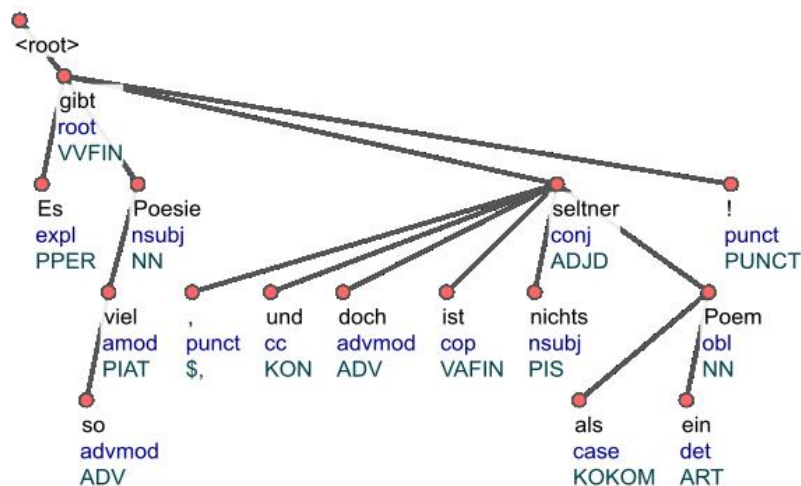


Figure 34 Dependency representation in tree-like form of the sentence “Es gibt so viel Poesie, und doch ist nichts seltner als ein Poem!” from *Kritische Fragmente* by F. Schlegel, according to the UD 2.0 scheme.

sent_id = lyceum-f4-s1

In Figure 35, the second clause shows a comparative construction triggered by the degree adjective *seltner*, which is morphologically marked (-er) to introduce an inequality comparison, in particular a scalar increasing comparison. In this case, the standard of comparison is a noun phrase introduced by *als*.

¹¹⁸ F. Schlegel, *Kritische Fragmente*, fragment 4.

¹¹⁹ Friedrich Schlegel’s *Lucinde and the Fragments*, University of Minnesota Press, 1971. ProQuest Ebook Central, <http://ebookcentral.proquest.com/lib/unibg-ebooks/detail.action?docID=345421>.

To be consistent with what explained in the previous example, the standard of comparison should depend on the element that triggers the comparison, which is the adjective *seltner* in this case. Therefore, the noun *Poem* depends back on *seltner*. However, the kind of relations is different with respect to the case analysed above. In fact, when *als* introduces a comparative noun phrase, i.e. a nominal comparative and not a clausal one, the head of this phrase depends on the first element of the comparison through the *obl* relation. In this case, the *obl* relation is extraordinarily used for a non-verbal dependent. since the comparative phrase is seen as a contracted clause.

3.3.4.3 Comparative: *Wie* Introducing an Oblique Argument

[54] **Es gibt Schriftsteller die Unbedingtes trinken wie Wasser;** und Bücher, wo selbst die Hunde sich aufs Unendliche beziehen.¹²⁰

54. **There are writers who drink the absolute like water;** and books in which even the dogs refer to the infinite.¹²¹

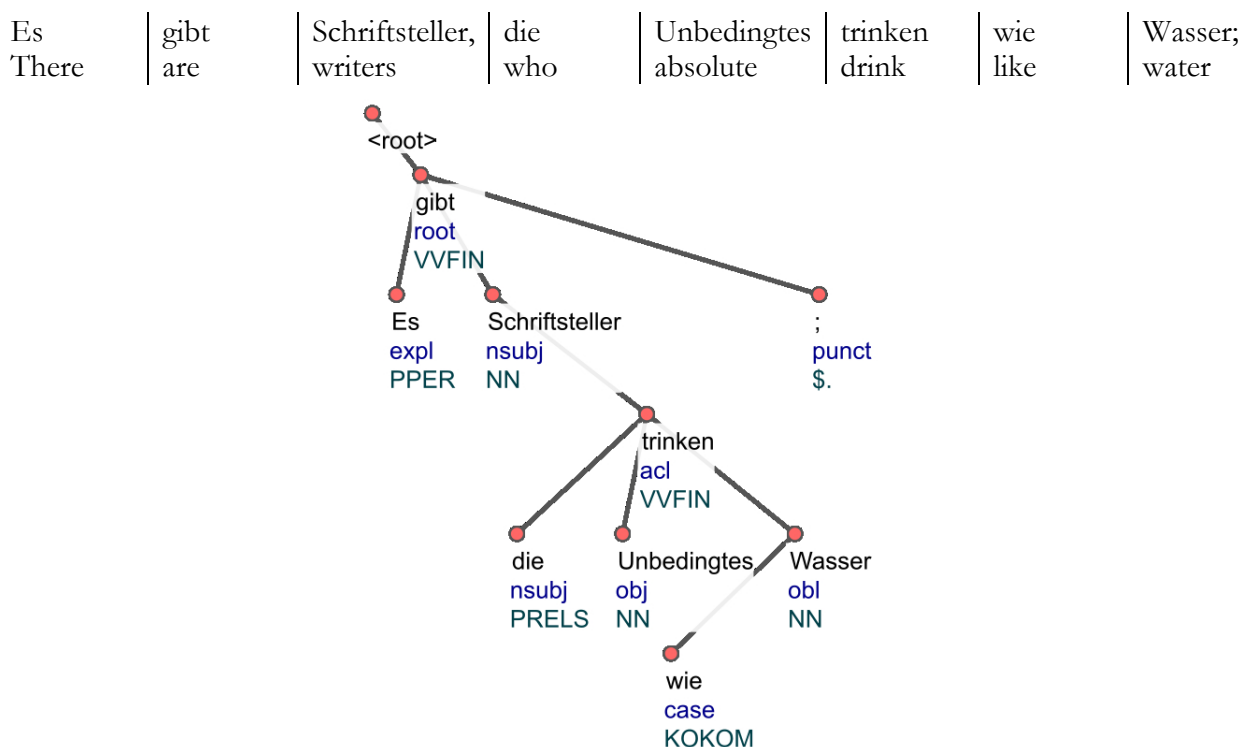


Figure 35 Dependency representation in tree-like form of the sentence "Es gibt Schriftsteller die Unbedingtes trinken wie Wasser;" from *Kritische Fragmente* by F. Schlegel, according to the UD2.0 scheme.

sent_id = lyceum-f54-s1

¹²⁰ F. Schlegel, *Kritische Fragmente*, fragment 54.

¹²¹ Friedrich Schlegel's *Lucinde and the Fragments*, University of Minnesota Press, 1971. ProQuest Ebook Central, <http://ebookcentral.proquest.com/lib/unibg-ebooks/detail.action?docID=345421>.

In Figure 36, there is a complex sentence consisting of a main clause and a relative clause. In the relative clause, a comparative structure is introduced by *wie*, which, therefore, works as a comparative conjunction here. To determine the best annotation of this element, I reasoned according to the deletion rule again. Indeed, if we delete the whole comparative phrase from the comparative conjunction on, the sentence still makes sense, both semantically and syntactically:

Es gibt Schriftsteller die Unbedingtes trinken.

Therefore, the phrase *wie Wasser* could have the function of a non-core argument of a predicate or a noun modifier. In functional terms, this phrase adds extra information to the predication, since it works as an adverb specifying the way in which some authors drink the absolute. This is a role usually played by an oblique argument that add some specification to the meaning of the verb. In addition, unlike the previous two examples of comparative structures, there is no element in the higher clause which can be clearly said to syntactically trigger this comparative phrase. Therefore, the status of non-core dependent of the subordinate predicate looks like the most appropriate one. Therefore, here, *wie* depends on *Wasser* through the *case* relation, exactly like any preposition introducing a noun phrase, while its status of comparative conjunction is signalled thanks to the XPOS (KOKOM). Then, *Wasser* depends on *trinken* through the *obl* relation.

3.3.5 Ellipsis

3.3.5.1 Gapping, (Appositional Modifier)

2. [...] Die Sprachlehre ist die Dynamik des Geisterreichs. **Ein Kommandowort bewegt Armeen; das Wort Freyheit Nazionen.**¹²²

2.[...] The theory of language of the dynamic of the spiritual realm! **One word of command moves armies – the word liberty – nations.**¹²³

Ein	Kommandowort	bewegt	Armeen;	das	Wort	Freyheit	Nazionen.
A	command	moves	armies;	the	word	liberty	Nations

¹²² Novalis, *Blütenstaub*, fragment 2.

¹²³ STOLJAR, Margaret Mahony, et al. (ed.). Novalis: Philosophical Writings. SUNY Press, 1997.

In Figure 37, there is a complex sentence consisting of two clauses. The main predicate is the verb *bewegt*. In the second clause, there is an apposition, i.e. a noun juxtaposed to another noun, which modify the first noun. Here, the noun *Freyheit* modifies the noun *Wort*, playing a role of specification. Therefore, *Freyheit* depends on *Wort* through the *appos* relation (*appos* stands for apposition).

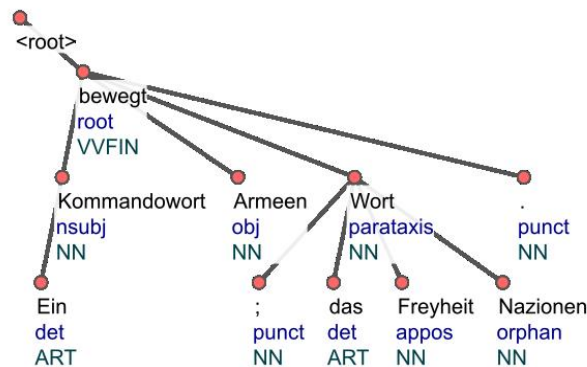


Figure 36 Dependency representation in tree-like form of the sentence "Ein Kommandowort bewegt Armeen; das Wort Freyheit Nazionen." from *Blüthenstaub* by Novalis, according to the UD 2.0 scheme.

sent_id = bluethenstaub-f2-s4.

Most importantly, there is no predicate in the second clause, because of a particular kind of verbal ellipsis, which is usually referred to as *gapping* (Sag 1976), (Schuster, Lamm, and Manning 2017). It consists in two coordinated clauses, in which the same predicate should be expected, but the predicate stands in the first one only, while it is totally omitted in the second one. In other words, the predicate of the first clause should be repeated in the second clause also, but it is actually not explicit. For instance, in the sentence reported in Figure 37, the verb *bewegt* should stand in the second clause as well, between *Freyheit* and *Nazionen*, which play the roles of subject and object of the second clause, respectively. On the contrary, the verb occurs only in the first clause. *Gapping* is particularly problematic in dependency grammar, since, in this formalism, the predicate is the core of the sentence. Therefore, when it is completely elided, a solution should be found to preserve the dependency structure. In UD 2.0, the introduction of fake nodes to replace ellipsis in the surface structure is avoided. Conversely, according to the UD 2.0 guidelines, one of its dependents has to be promoted as head of the clause (Droganova and Zeman 2017) according to the following rules:¹²⁴

- If there is an auxiliary, or copula, or infinitival marker, it should be promoted as head. In this case, an element that was originally part of the predicate phrase (as a copula) is used. In this way, all the other elements in the sentence maintain the syntactic relation that they would have with

¹²⁴ Cf. <https://universaldependencies.org/u/overview/specific-syntax.html#ellipsis>

the missing verb. Such a solution especially works for English, where auxiliaries such as *do* or *will* can replace entire clauses or even entire sentences. In this case, no extra relation is added.

- When none of these elements of the verbal phrase stands in the clause, as in the sentence in Figure 37, a promotion of another element is needed. In fact, in order to avoid the introduction of a fake node in the surface structure, an element of the clause has necessarily to be the node that should be occupied by the predicate. The following hierarchy should be observed:

nsubj > obj > iobj > obl > advmod > csubj > xcomp > ccomp > advcl

At the same time, a new relation for all the dependents that are orphan of the missing predicates has to be introduced: the *orphan* relation. This is necessary to avoid unnatural dependencies due to the replacement of the predicate, as explained below.

Let us move back to the sentence in Figure 37. In the second clause, according to the above-mentioned hierarchy, the noun *Wort* is promoted to the position of predicate, since it would have been the nominal subject of the predicate *bewegt* if this had not been omitted. Therefore, it depends back on the main predicate *bewegt* through the function that the predicate would have had if present, which is *parataxis* in this case (since the two clauses are linked through asyndetic coordination). *Freyheit* regularly depends on *Wort* through the relation *appos*, since the introduction of the *orphan* relation does not affect any of the relations involving the children of the other nodes. *Conversely*, the noun *Nationen* has necessarily to depend on the noun *Wort* through the *orphan* relation. Without the verbal ellipsis, it would have played the role of direct object of the verb, thus depending on it through the *obj* relation. If it maintained its original relation, it would generate an unnatural syntactic relation, being the direct object of a noun. But this is unacceptable, apart from the case of nominal predicates.

3.3.5.2 Modal Verb Promotion as Head in *Afinite Konstruktion*

[131] **Der Dichter kann wenig vom Philosophen, dieser aber viel von ihm lernen.** Es ist sogar zu befürchten, daß die Nachtlampe des Weisen den irre führen möchte, der gewohnt ist im Licht der Offenbarung zu wandeln.¹²⁵

¹²⁵ A.W. Schlegel, *Athenäums Fragmente*, fragment 131.

131. **The poet can learn little from the philosopher, but the philosopher much from the poet.** It's even to be feared that the night lamp of the sage may lead someone astray who is given to walking by the light of revelation.¹²⁶

Der	Dichter	kann	wenig	von	Philosophen,	dieser	aber	viel	von	ihm	Lernen.
the	poet	can	little	from	philosopher	this	but	much	from	him	learn
						one					

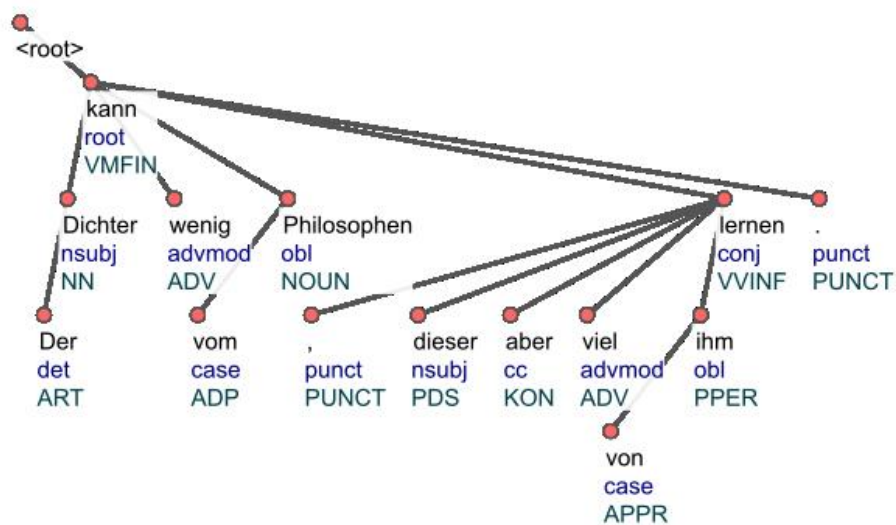


Figure 37 Dependency representation in tree-like form of the sentence " Der Dichter kann wenig vom Philosophen, dieser aber viel von ihm lernen." from Athenäums Fragmente, according to the UD 2.0 scheme.

sent_id = athenaeum-fl31-s1

In Figure 38, there is a complex sentence consisting of two coordinate clauses, which shows an unusual structure. In fact, let us imagine to divide the sentence as follows:

Der Dichter kann wenig vom Philosophen,

Dieser aber viel von ihm lernen.

We can consider it as a complex sentence, made up of two coordinate clauses, each one showing a missing element in the verbal phrase. Precisely, the non-finite verb which should be the head of the modal verb *kann* is missing in the first clause, while the modal verb *kann* itself is missing in the second clause, in which, on the contrary, the non-finite verb *lernen* is explicit, and stands at the end of the clause. The verbal phrase is expected to appear in the same form in both the sentences, as follows:

¹²⁶ Friedrich Schlegel's *Lucinde and the Fragments*, University of Minnesota Press, 1971. ProQuest Ebook Central, <http://ebookcentral.proquest.com/lib/unibg-ebooks/detail.action?docID=345421>.

Der Dichter *kann* wenig vom Philosophen *lernen*,

Dieser aber *kann* viel von ihm *lernen*.

In fact, the subjects of the two clauses, i.e. the noun *Dichter* and the pronoun *Dieser*, both require the same predicate in this context, but the verbal phrase is apparently divided between the two subjects. This construction can be regarded as a form of auxiliary ellipsis that was rather common in written German texts in prose in early-modern German, but it was still in use in some text typologies until the end the 18th century. It is commonly known as *Afinite Konstruktion* (Öhl 2009), which literally means non-complete construction. In fact, the verbal phrase is not complete, neither in the first clause, nor in the second clause. In the first case, the verb is missing, while in the second case, the auxiliary is missing. This poses significant problems for the dependency annotation, because of the (partial) omission of the predicate, and because of the disposition of the elements of the verbal phrase as well. To syntactically annotate this sentence, I opted for considering the two clauses as two coordinate items. Then, I promoted the auxiliary *kann* as predicate of the first clause, considering the first clause as the main clause. This is allowed by the rules adopted in UD to annotate the verbal ellipsis. Then, I considered the non-finite verb *lernen* as predicate of the second clause, therefore I let it depend on the node of the main predicate *kann* through the *conj* relation. The other dependency relations were assigned consequently.

4 Analysis

4.1 Goals and Methodology

The language of the Fragments is here empirically investigated for the first time through a treebank-based approach. The aim of this analysis is twofold. On the one hand, this analysis attempts to demonstrate which features a treebank-based approach can detect in the language of a literary genre, whose description would be hardly feasible through traditional methods. In this regard, I especially aim to highlight those benefits offered by a treebank based-approach with respect to a common corpus-based approach that cannot exploit any linguistic annotation going beyond the level of parts of speech. On the other hand, the linguistic features of Fragments are not investigated in absolute terms, but against two different textual genres. I chose the two genres that are currently represented in the two main UD treebanks for German, i.e. web texts from the GSD Treebank (McDonald et al. 2013b), mainly taken from Wikipedia¹²⁷, and web-news texts¹²⁸ from the HDT treebank (Völker et al. 2019), which come from a websites o news about technology world. Such a comparison aims to provide empirical evidence about the differences, if any, between the new literary variety that is here given representation in a dependency treebank for the first time, and those that are usually considered the *de facto* standard variety to work with (dependency) treebanks, not only in UD but in treebanking in general.

To perform the investigation, I selected a dataset in CoNLL-U format for each textual genre, and I selected a tool that allows to retrieve specific evidence from the datasets through formal queries, especially concerning dependency relations. I opted for SETS, a free online tool integrated the UD infrastructure, which is maintained by the Turku NLP group.¹²⁹ The query language implemented in this tool is loosely inspired by TGrep2 and TRegex, but it is specifically designed for querying general dependency graphs. For an overview on the syntax of the search expressions in this language, I redirect to the official page, where the query language is explained through different practical examples¹³⁰. However, I clarify the syntax of the expressions during the analysis, when I report the queries used to extract data. The tool consists in a search bar in which the user can write and run a search expression, and a page where those sentences of the datasets matching the query are returned. The results can be downloaded in CoNLL-U format. Moreover, the tool displays not only the dependency graph (in linear

¹²⁷ This genre is not the only one represented in this treebank, but it is however the main one (see 4.2).

¹²⁸ More in detail, they come from a specific subdomain of web news concerning technological products (see 4.2).

¹²⁹ http://bionlp-www.utu.fi/dep_search/.

¹³⁰ <https://bionlp.utu.fi/searchexpressions-new.html>.

form) of each returned sentence, but also the respective unannotated counterpart in the original context, making it especially suitable for searching literary genres, where the context can help better understand some linguistic phenomena. For a matter of convenience and clarity, I report the graph-structure output by SETS only in some specific cases, where it can help better illustrate the issue under investigation. Most of the times, I only report the unannotated sentences in linear form, in which only the specific dependency relation which is under investigation is highlighted. As for the drawbacks of the tool, it does not allow to search any specific subsets of the UD treebanks, since the queries can be run on the whole treebank only¹³¹. Consequently, if one wants to work on specific subsets of the treebank, such as the training set or the test set, post-processing on the downloaded file is needed to filter the results. When needed, I usually post-processed the output of the query in a text editor through regular expressions. Moreover, I encountered some problems in the extraction of data from the HDT treebank, probably due to the very large amount of data collected in this dataset. I signal them during the analysis.

The analysis is structured as follows. I first provide an insight into the distribution of parts of speech (POS), showing how the syntactic information can help read the distribution of POS more in detail. I then provide an overview of the distribution of syntactic relations in each dataset. The analysis focuses then on predicates, which are the core of dependency-based syntax. I therefore exploited the dependency annotation to provide a quantitative and, when possible, qualitatively, portrait of the distribution and the use of predicates in the genres. I considered nonverbal and verbal predicates, verbal forms, existential constructions, modal verbs, as well as several direct dependents of the predicates within the sentence. For each of these features, I reported the relative frequency (RF) of the specific parameters that are investigated, and I discussed the results. The RF was always calculated through a script written in R and ran in R Studio. During the analysis, I provide further indications concerning the method of the quantitative analysis when needed.

In the end, it is worth underlying that this is not a comprehensive treebank-based stylistic analysis of Fragments. First, such an analysis would require many further parameters to be considered. Second, such an analysis should compare the features of Fragments with respect to a textual genre from the same age, or, in any case, it should be motivated by research questions that especially move from literary considerations. For instance, the necessity to compare two genres that embodied the style of two distinct or clashing literary movements¹³², or the works by two authors that are considered literary opponents. Unfortunately, the lack of syntactically annotated data from the same age of Fragments is a strong bottleneck toward such a goal. In any case, such a stylistic analysis could not fall within the scope of this thesis. By contrast, this analysis represents the necessary first step toward that goal. In fact, I here aim to

¹³¹ The tool searches the treebank partitions, i.e. training file, testing file and development file, in the dev branch of the GitHub repository. However, I was assured by the that they correspond to those published in the UD file.

¹³² To this respect, see the conclusions of this thesis for a suggestion about a proper stylistic analysis based on this treebank.

preliminary test the treebank-based methodology for the literary-linguistic analysis, which is still a totally unexplored area in the field of corpus-based linguistic approaches to literature. The results of this analysis show that a series of hidden linguistic features of a literary text, especially syntactic features that peculiarly characterize its language, can actually be detected and investigated thanks to a dependency treebank. This can hopefully pave the way for the development of a proper treebank-based, specifically dependency-based, literary stylistics in future.

4.2 Datasets

The analysis is conducted on three datasets from three German treebanks hosted in UD 2.5 (Zeman et al. 2019): GSD, HDT and LIT¹³³. In particular, I considered the training files of both GSD and HDT, while I considered the test file of LIT¹³⁴. Table 1 summarizes the whole dataset, as well as the textual composition. From now on, I will use the official IDs of the three UD treebanks to refer to the respective dataset used in the analysis. As shown in Table 39, the three datasets are very different in size, therefore I considered only the RF of features in the analysis.

ID	Tokens	Sentences	Genre
LIT	40,545	1,922	Fragments (100%)
GSD	268,414	13,814	Web news (11%) Web reviews (5%) Wikipedia (84%)
HDT	2,653,628	153,035	Web news about technology (100%)

Table 39 Dataset used for the analysis.

All the three treebanks from which the data were taken passed the official UD validation test to be published in the 2.5 release. It means that all of them respect a set of fundamental requirements of the UD guidelines. However, some deviations are expected in all the datasets, especially due to the automatic

¹³³ LIT, GSD and HDT are the three official identifiers (ID) of the three treebanks in UD.

¹³⁴ For this treebank, data were not split into training, test and development set, but only the test file was available in UD 2.5.

annotation, or to the automatically conversion from other schemes. Table 40 summarizes the source of the annotation of each dataset in detail.¹³⁵

Annotation	Source LIT	Source GSD	Source HDT
Lemmas	assigned by a program, with some manual corrections, but not a full manual verification	assigned by a program, not checked manually	annotated manually in non-UD style, automatically converted to UD
UPOS	annotated manually in non-UD style, automatically converted to UD, with some manual corrections of the conversion	annotated manually in non-UD style, automatically converted to UD	annotated manually in non-UD style, automatically converted to UD
XPOS	assigned by a program, with some manual corrections, but not a full manual verification	assigned by a program, not checked manually	assigned by a program, with some manual corrections, but not a full manual verification
Features	not available	assigned by a program, not checked manually	annotated manually in non-UD style, automatically converted to UD
Relations	annotated manually, natively in UD style	annotated manually in non-UD style, automatically converted to UD	annotated manually in non-UD style, automatically converted to UD, with some manual corrections of the conversion

Table 40 Sources of the annotation of each dataset.

¹³⁵ For LIT, see https://universaldependencies.org/treebanks/de_lit/index.html; For GSD, see https://universaldependencies.org/treebanks/de_gsd/index.html; For HDT, see https://universaldependencies.org/treebanks/de_hdt/index.html.

Moreover, a certain degree of errors and inconsistencies in the annotation is expected in all the treebanks. Some of them are reported and discussed when detected during the analysis.

4.3 Overall Distribution of Parts of Speech (UPOS)

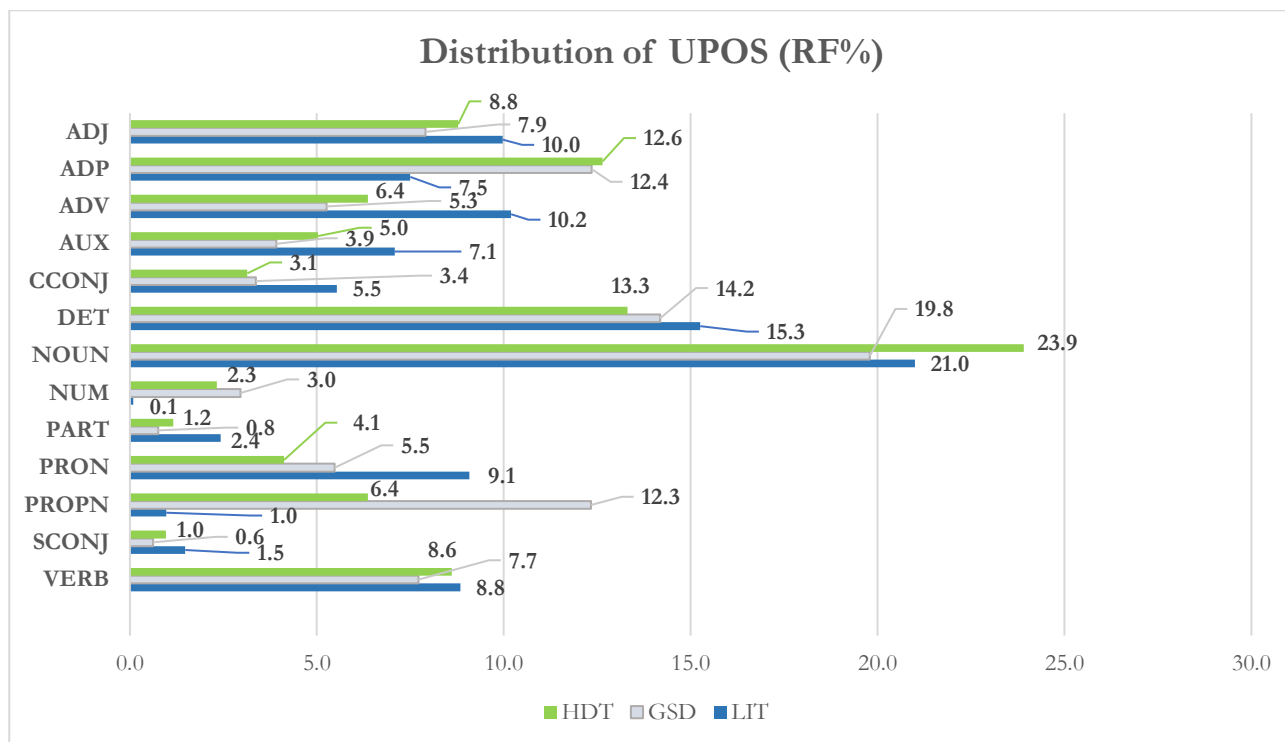


Chart 7 Overall distribution of UPOS.

Chart 7 displays the overall distribution (RF %) of parts of speech (UPOS) in each dataset¹³⁶. The role of this section of the analysis is twofold. On the one hand, it provides a first insight into the lexical features of the genres, which also anticipate some of their syntactic features. On the other hand, it offers the opportunity to show some of the benefits offered by a treebank-based approach with respect to a common corpus-based approach that cannot exploit dependency relations. In fact, dependency relations let us better interpret some of the data that we observe in the distribution of the parts of speech. I will show how through some examples.

Nouns are the most common part of speech across all the datasets, especially in HDT. One may ask whether all genres show a high frequency of nouns for the same reason. In fact, with respect to other parts of speech, nouns can fill a very wide range of syntactic functions: among the others, nouns can be

¹³⁶ Punctuation, i.e. the UPOS PUNCT, was excluded from the count.

nominal subjects, direct objects, predicates (in nonverbal predicates), oblique cases, and many others. We may therefore ask whether nouns are equally distributed in these functions in each data set. In terms of dependency relations, it means investigating the distribution of the most frequent dependency relations governing nouns.¹³⁷ To perform this investigation, I extracted the list of all the rows of tokens tagged as NOUN in the UPOS field from the CoNLL-U file (for the CoNLL-U format, see Chapter 2) of each dataset, and I calculated the distribution of their dependency relations (which fill the DEPREL field) as relative frequency (RF). For an overview on the dependency relations used in UD and on how they encode syntax, see Chapter 3. Table 41 summarizes the results.

	LIT		GSD		HDT	
RANK	DEPREL	RF	DEPREL	RF	DEPREL	RF
1	nmod	24.1	nmod	25.4	obl	25.2
2	nsubj	19.0	obl	23.5	nmod	23.4
3	obl	17.7	nsubj	14.4	obj	17.2
4	conj	12.8	obj	10.8	nsubj	16.4
5	obj	12.5	conj	9.2	conj	6.4
6	root	6.2	nsubj:pass	4.0	appos	3.6
7	parataxis	1.6	root	3.6	nsubj:pass	3.1
8	xcomp	1.3	appos	2.3	root	2.3
9	iobj	1.1	compound	1.9	nummod	1.5
10	appos	0.9	flat	1.7	iobj	0.5
11	nsubj:pass	0.7	iobj	1.2	parataxis	0.2
12	acl	0.5	xcomp	0.9	xcomp	0.1

Table 41 Distribution of dependency relations governing nouns.

As shown in Table 41, the distribution of the functions of nouns varies across the data sets. The most frequent function in both LIT and GSD is nominal modifier (nmod), while the most frequent one in HDT is oblique modifier (obl). In both GSD and HDT, the functions of oblique modifiers and nominal modifier are overall the two most frequent ones, both showing a rather similar RF, and, moreover, there is a considerable gap in RF between them and the third most frequent functions. Conversely, RF of oblique arguments is sensibly lower in Fragments with respect to the other two genres, even if they are,

¹³⁷ The dependency relations encode the syntactic function of the dependent (child node) with respect to the head (parent node).

however, the third most frequent function for nouns. This leads to two first conclusions about the distribution of nouns. The first one concerns the relation between nouns and functional categories (for the functional categories, see Chapter 3). In fact, the distribution observed in Table 41 suggests that nouns tend to fill the functional categories of non-core dependents or nominal modifiers much more frequently than those of core-arguments, in all the three genres. In this respect, Table 42 summarizes the distribution of nouns grouped per functional category. The category *core-arguments* groups RF of those nominals filling core arguments, i.e. nominal subjects, direct objects (obj) and indirect objects (iobj), while the category *dependents* groups those nominals filling both nominal non-core dependents, i.e. oblique arguments, and nominal dependents, i.e. nominal modifiers¹³⁸. In addition, I added the categories of *root* for those nouns playing the function of main predicates (in nonverbal predicates), and I grouped RF of all the syntactic relations belonging to other functional categories under *other*, in order to have a more comprehensive portrait of the distribution of nouns across the UD functional categories.

FUNCTIONAL CATEGORY	RF LIT	RF GSD	RF HDT
Core-Argument	33.3	30.4	37.1
Dependent	42.6	51.3	48.6
Root	6.2	3.6	2.3
Other	17.8	14.7	14.7

Table 42 Distribution of nouns in functional categories.

On the other hand, a higher distribution of *obl* relation in GSD and HDT with respect to Fragments suggests a more frequent use of specification in these two genres, such as locative specification and time specification. This is surely correlated to the distribution of prepositions (ADP) observed in Chart 7, which in fact shows a huge gap between Fragments and the other two genres: RF of prepositions is very high in both GSD and HDT, while it remarkably decreases in Fragments. Prepositions notoriously attach to nouns building prepositional phrases, which functionally work as nominal modifiers or oblique arguments. When encoding locative and time information, they usually work as oblique modifiers. If we consider the text typologies collected in both GSD and HDT, the high distribution of prepositional phrases encoding locative and time specification in these datasets should not be very surprising. As for GSD, most texts come from Wikipedia. We can therefore define them as bibliographical and encyclopaedic texts, which notoriously encode a lot of information about space and time. Moreover,

¹³⁸ Both the categories of core arguments and dependents in Table 4 include nominals only. For the distribution of nouns as clausal core arguments and clausal dependents, such as clausal subjects and adverbial clausal modifiers, see (4.5.5).

other texts of GSD come from the genre of web news, in which the need to encode that type of specification is notoriously high as well. As for HDT, the texts entirety belong to the macro genre of web news. The sentence in (1) exemplifies the use of prepositional phrases for this purpose in GSD. The prepositional phrases which generates oblique modifiers are highlighted in bold (both the preposition and the nominal head), and each sentence extracted from the dataset is followed by a free translation in English. The prepositional phrases *auf Dauer* ('in the long run') encodes and *am Mittwoch* ('on Wednesday') encodes time information, while *in Bonn* ('in Bonn') encodes locative information. It is worth noting how the spatial information could be encoded through toponyms, which are proper nouns, as in the case of *in Bonn*. Therefore, the high frequency of proper nouns observed in both GSD and HDT must also be correlated to the frequency of prepositions as well. The distribution of proper nouns is addressed later.

(1) Die Grundbedürfnis könne **auf Dauer** nicht unterdrückt werden, sagte die CDU -- Politikerin **am Mittwoch in Bonn**.¹³⁹

The basic need could not be suppressed **in the long run**, said the CDU politician **on Wednesday in Bonn**.

Both examples were retrieved though the following query, which returns any noun¹⁴⁰ (NOUN) which is governed (<) through *obl* relation by any token (␣), and which, at the same time, governs (>) a preposition (ADP) through case relation:

NOUN <obl_␣>case ADP

Furthermore, the high frequency of numerals observed in both GSD and HDT (much lower in Fragments) must also be correlated to the need to encode time information in these two genres, and the consequently high frequency of propositions and oblique arguments. The example in (2) illustrates a prepositional phrase encoding time information through a numeral in GSD. The time information is encoded through the prepositional phrase *ab 1994* ('from 1994').

(2) Die Wehrpflicht wird **ab 1994** abgeschafft.¹⁴¹

Conscription is abolished **from 1994**.

The example in (2) was retrieved through the following query, which returns any numeral (NUM) that is governed (<) through *obl* relation, and that, in turn, governs (>) a preposition (ADP) through *case* relation.

¹³⁹ sent_id = train-s1940.

¹⁴⁰ Note that the left-most token in the expression is always the target of the search.

¹⁴¹ sent_id = train-s1604.

As for Fragments, they are mostly speculative texts, in which the need to encode locative and time information is supposed to be much lower with respect to the other two genres. In fact, they mainly comment and judge aesthetic principles behind art, and different aspects of intellectual and cultural issues in general. Therefore, they mostly use declarative sentences, whose message aims to have a broad scope, and is not limited to a particular action, which is contextualized in a specific time or space. Moreover, both references to historical events and toponyms tend to occur very rarely in this genre. This must be the reason behind the frequency of oblique modifiers, as well as behind the frequency of both prepositions and numerals, which is for much lower with respect to the other genres. If Table 42 showed a trend that was similar across the three datasets, i.e. the distribution of nouns in functional categories, we have now detected a first difference between Fragments and the other two genres, which is not only syntactic, but also semantic: Fragments tend to avoid a very frequent use oblique modifiers, presumably to avoid encoding spatial and locative information through nouns and proper nouns. The sentence in (3) exemplifies the use of a declarative sentence in Fragments, which conveys a judgment about an intellectual issue concerning their age, in which no time or locative information is provided:

(3) Man hat schon so viele Theorien der Dichtarten¹⁴².

We already have so many theories about the typologies of poetry.

When referring to issues of previous ages, both time and space information are frequently omitted as well. An example is shown in (4). In this Fragment, the author refers to an aspect of the thought of Sophocles¹⁴³, which is not contextualized using any oblique argument encoding space, or time, or any other information. The statement remains therefore at a sort of universal level.

(4) Schon Sophokles glaubte treuherzig, seine dargestellten Menschen seien besser als die wirklichen.¹⁴⁴

Already Sophocles naively believed that the men that he represented were better than the real ones.

Both examples (3) and (4) were retrieved through the following query, which returns any verb (VERB) that does not govern (!>) any token through *obl* relation¹⁴⁵:

¹⁴² sent_id = lyceum-f62-s1.

¹⁴³ An author of the ancient Greece.

¹⁴⁴ sent_id = lyceum-f125-s1.

¹⁴⁵ This query aims at returning some relevant examples, i.e. at least a verbal predicate that does not govern any token through *obl* relation. It did not aim to retrieve all the sentences that do not contain any *obl* relation.

Let us return to the distribution of the functions of nouns observed in Table 41. Fragments show a higher distribution of nouns in the role of root node (root), i.e. as main predicates. This could correspond to a wider use of nouns as nonverbal predicates in this genre with respect to the others. For a detailed analysis of the distribution of verbal and nonverbal predicates across the datasets, see (4.5.1). Moreover, there is another clear difference between Fragments and the other genres. In fact, the use of nouns as subjects of predicate in passive forms (*nsubj:pass*) is remarkably higher in both HDT and especially in GSD, with respect to Fragments. The use of passive voice seems therefore much more frequent in these two genres. This is strictly correlated to the textual genres. Both news and historical texts frequently use the passive form to put the focus on the action rather than on the agent, since they usually aim to report and describe historical events, or in general processes (also referring to the present) in an objective way. Examples from (5) to (7) show the use of the passive voice for this purpose in GSD. In (5), the predicate in passive form is the verb *ruhiggestellt* ('immobilized'), while the passive auxiliary is *wurden*, i.e. the past of *werden* ('were'). In (6), the verb in passive form is *fortgesetzt* ('carried on'), while the auxiliary is *wurden*. In (7), the passive verb is *untersucht* ('examined'), while the auxiliary is always *wurden*. In German, the auxiliary verb of the passive voice usually occupies the second position of the clause, while the past-participle form occupies the last position of the clause.

- (5) In China **wurden** zwei Arten von Gegnern politisch **ruhiggestellt**: Umweltschützer und das Militär..¹⁴⁶
In China, two types of opponents **were** politically **immobilized**: environmentalists and army.
- (6) Nach dem Tod von Abt Guéranger **wurden** die Arbeiten vor allem durch den Mönch André Mocquereau **fortgesetzt**.¹⁴⁷
After the death of the abbot Guéranger, the works **were carried on** by the Monk André Mocquereau.
- (7) Hierbei **wurden** 370 periprothetische Membranen, die bei einem Prothesenwechsel **entfernt wurden**, von Pathologen histologisch mit dem Mikroskop **untersucht**.¹⁴⁸
Here, 370 periprosthetic membranes, which were removed when the prosthesis **was changed**, **were** histologically **examined** with a microscope by pathologists.

¹⁴⁶ sent_id = train-s5368.

¹⁴⁷ sent_id = train-s4070.

¹⁴⁸ sent_id = train-s4032.

Examples from (5) to (7) were retrieved through the following query, which returns any verb governing a noun through *nsubj:pass* relation.

VERB >nsubj:pass _

On the contrary, active forms are clearly more frequent in Fragments with respect to the passive forms: RF of *nsubj:pass* relation is only 0.7, while RF of *nsubj* is 19.1, and RF of *nsubj* relation is remarkably higher in Fragments with respect to both GSD and HDT. The passive form is however used in some cases. (8) exemplifies the use of the passive form in a typical judgment of Fragments. The subject of the passive form is the noun *Schrift* ('writing'), while the predicate in passive form is the past-participle verb *verstanden* ('understood'), while *werden* ('be') is the non-finite auxiliary verb to build the passive form. Moreover, the sentence has negative polarity, thanks to the adverb *nie* (never), which also stresses the universal scope of the message. In this case, the passive form is clearly exploited to bring the focus on the action of understanding, which, according to the author, should never be accomplished when reading classical authors. The passive becomes therefore a strategy to stress the complexity of the classical literary works with respect to those who do not belong to this category by focusing on the difficulty of understanding.

(8) Eine klassische Schrift muß nie ganz **verstanden werden** können.¹⁴⁹

A classical writing must never be understood.

As shown in Chart 7, proper nouns (PROPN) occur very rarely in Fragments, while they are very frequent in the other two genres, but especially in GSD. Fragments' authors use proper nouns only in some cases to refer to authors or to their works, as already shown in (8). Overall, Fragments aim to be ambiguous and universal at the same time in their messages, therefore the direct denotation embodied by proper nouns is mostly avoided. Moreover, most fragments reflect about abstract concepts and values, therefore denotation of real-world entities is mostly useless. On the contrary, such denotation is fundamental in Wikipedia's texts, but also in web news, since both these genres continuously refer to real entities, especially public figures, both current and historical ones, institutions, and geographical entities. An Example of the use of a proper noun to denote these classes of entities in GSD is shown in (9). It was retrieved by searching for any occurrence of the UPOS PROPN in the dataset. In particular, (9) shows three toponyms. The first one, i.e. *Chancy*, syntactically works as nominal subject of the main nonverbal

¹⁴⁹ sent_id = lyceum-f20-s1.

predicate, which is the noun *Gemeinde* (‘community’), while the other two syntactically work as appositions (*appos* relation), i.e. *Somme* and *Picardie*.

- (9) **Canchy** ist eine französische Gemeinde mit Einwohnern (Stand) im Département **Somme** in der Region **Picardie**;¹⁵⁰

Speaking of the *appos* relation, we observed that it is rather frequent among the syntactic functions embodied by nouns both in GSD and GSD (Cf. Table 41). (10) exemplifies the use of a noun in the role of apposition in GSD. It was retrieved through the following query, which returns any noun governed through *appos* relation:

NOUN <appos _

The noun is *Offizier* (‘officer’), which works as apposition of the proper noun *Oleg Wladimirowitsch Penkowski*, to specify the title of the person embodied by the proper noun. The role of apposition specifying titles, i.e. through common nouns and not through proper nouns, is expected very frequent in both GSD and HDT, especially when referring to public figures. Obviously, this is directly correlated to the high frequency of proper nouns in both these datasets.

- (10) Eine dieser Quellen war Oleg Wladimirowitsch Penkowski, ein **Offizier** der GRU [...].¹⁵¹
One of these sources was Oleg Wladimirowitsch Penkowski, an **officer** if the GRU [...].

As for Fragments, the role of nouns as appositions is much less frequent. An example from Fragments is reported in (11). Interestingly, unlike in the previous cases observed in GSD and HDT, apposition is here used to specify an abstract category, not a real entity, which is embodied by the noun *Wort* (‘word’). The noun playing the function of apposition is the word *Freyheit* (‘freedom’).

- (11) Ein Kommandowort bewegt Armeen; das Wort **Freyheit** Nazionen.¹⁵²
An order moves armies; the word freedom nations.

Let us now focus on the distribution of those UPOS that seem to particularly differentiate¹⁵³ the lexicon of Fragments with respect to the that of the other genres. The use of adverbs is particularly frequent in Fragments (see Chart 7). Notoriously, adverbs can encode locative and time information, such as in the

¹⁵⁰ sent_id = train-s2504.

¹⁵¹ sent_id = train-s6185.

¹⁵² sent_id = bluethenstaub-f2-s4.

¹⁵³ In terms of frequency.

case of *hier* ('here'), or *heute* ('today'), respectively. Since the use of oblique cases in Fragments is much lower with respect to the other genres, one may ask whether they exploit adverbs rather than nouns to encode this information in sentences. I therefore extracted the list of all the forms with the UPOS ADV from the CoNLL-U file of the Fragments, to verify whether there was any correlation between the frequent use of adverbs and the encoding of spatial and/or time information. Table 43 summarizes the results.

RANK	FORM	RF LIT
1	auch	8.4
2	nur	3.7
3	noch	3.3
4	sehr	2.6
5	so	2.1
6	wieder	1.8
7	jedoch	1.7
8	mehr	1.6
9	aber	1.6
10	etwa	1.6
11	dort	1.5
12	heute	1.5

Table 43 Distribution of forms of adverbs in LIT.

As shown in Table 43, there is no correlation between the need to encode temporal and locative information and the higher distribution of adverbs in Fragments. The first temporal adverbs in the ranking is *dort* ('there'), which only ranks 11th, and the first locative adverb is *heute* ('today'), which ranks 12th. Conversely, the most frequent adverb is *auch* (also), the second most frequent one is *nur* ('only'), the third most frequent one is *noch* ('again'), and the fourth most frequent one is *sehr* ('very'). An instance of *auch* in Fragments is reported in (12). In this case, it is governed by the verb *liegt* ('is'). It was retrieved through the following query, which returns any token whose form is *auch*, and which is governed through *advmod* relation.

auch <advmod _¹⁵⁴

(12) Die instinktartige Universalpolitik und Tendenz der Römer liegt **auch** in dem Deutschen Volk.¹⁵⁵

The Universal politics moved by good instinct and the trend of Romans **also** is in the German people.

Interestingly, Fragments also shows a much higher distribution of both adverbs and pronouns with respect to the other two genres (Cf. Chart 7). As done for adverbs, I therefore investigated the distribution of the most common lemmas of pronouns in this dataset¹⁵⁶. I extracted the list of all those word forms tagged as PRON in the UPOS field from the CoNLL-U file of LIT, and I calculated their distribution as RF. Table 44 summarizes the results. I then repeated the same extraction considering the lemmas of pronouns rather than the forms. Table 45 summarizes the results.

RANK	FORM	RF LIT
1	sich	10.3
2	es	8.5
3	sie	8.3
4	man	7.3
5	die	5.5
6	er	5.2
7	was	4.2
8	Es	3.6
9	der	3.3
10	alles	2.1
11	das	2.0
12	nichts	2.0

Table 44 Distribution of pronominal forms in LIT.

¹⁵⁴ If not explicitly mentioned, no search of word forms is case-sensitive.

¹⁵⁵ sent_id = bluethenstaub-f64-s7.

¹⁵⁶ In this way, we can also verify whether there is a significant frequency of false positives causing the RF of ADV to be so high, i.e. forms which are not actually adverbs, but which are POS-tagged as adverbs due to errors in the annotation.

As shown in Table 44, the most common form of pronouns in Fragments is the reflexive pronoun *sich*. (13) and (14) exemplify two different uses of *sich* in this genre. They were both retrieved through the following query, which returns any occurrence of the token *sich* that is tagged as PRON (UPOS):

sich&PRON

In (13), *sich* is used as core argument of the verbal predicate *wundern* ('surprise'), within a non-finite clause that has the function of clausal subject of the main clause. The predicate of the main clause is the adjective *indelikat* ('indelicate'). In this case, the pronoun *sich* depends on the verb through *obj* relation.



- (13) Es ist indelikat, **sich** drüber zu **wundern**, wenn etwas schön ist, oder groß.¹⁵⁷
It is indelicate to surprise yourself when something is beautiful or big.

In (14), the pronoun *sich* is used in the role of oblique argument of the verb *lebt* ('lives') (in the second clause), through the prepositional phrase *in sich* ('in herself'), therefore *sich* depends on *lebt* through *obl* relation, while the noun *Leben* ('life') is the direct object of the verb *leben*, therefore it depends on *lebt* through *obj* relation.



- (14) Der Wahn lebt von der Wahrheit; die Wahrheit **lebt** ihr Leben in **sich**.¹⁵⁸
The delusion lives of truth; the truth lives her life in herself.

RANK	LEMMA	RF LIT
1	der	14.0
2	es	12.5
3	sich	10.7
4	sie	10.2
5	man	8.8
6	er	6.3
7	was	5.0


¹⁵⁷ sent_id = lyceum-f127-s1.

¹⁵⁸ sent_id = bluethenstaub-f8-s2.

8	welcher	2.5
9	alle	2.5
10	nichts	2.3
11	dieser	2.0
12	wir	1.8

Table 45 Distribution of pronominal lemmas in LIT.

Conversely, if we consider lemmas (see Table 45), the most frequent pronominal lemma is *der*, which is responsible for all the inflected forms of personal pronouns, such as *der*, *dem*, *die*, and others. All these forms of *der* can frequently occur as relative pronouns, i.e. introducing a relative clause, which modifies a higher predicate as adjectival clausal modifier (*acl* relation). The example in (15) shows the use of *der* as relative pronoun in a relative clause. The pronoun *der* is the nominal subject of the verb *wiederkaut* (‘ruminates’), and refers back to the nonverbal predicate of the higher clause, i.e. the noun *Leser* (‘reader’). Therefore, the predicate of the relative clause depends back on *Leser* through *acl* relation.

- (15) Ein Kritiker ist ein Leser, **der** wiederkaut.¹⁵⁹
 A reader is a critic who ruminates.
- 

One may ask to what extent the relative clauses have a role in the widespread use of *der*, and in massive use of pronominalization in Fragments. I therefore extracted all those occurrences of the lemma *der* working as relative pronoun through the following query, which returns all the occurrences of the lemma *der* which is governed by a token that, in turn, is governed by another token through *acl* relation:

L=der < (_ <acl _)

I divided the absolute frequency of the returned hits (AF = 301) by the total number of occurrences of the lemma *der* (AF = 501). The result is that 67.5% of the occurrences of the lemma *der* in Fragments are actually relative pronouns. I repeated the same investigation with the following query, i.e. considering all those pronouns that are governed by a token that, in turn, is governed by another token through *acl* relation:

¹⁵⁹ sent_id = lyceum-f27-s1.

PRON < (_ <acl _)

I divided the absolute frequency of the returned hits (579) by the total number of occurrences of the UPOS PRON (3,195). The result is that 18% of pronouns are relative pronouns in Fragments. It therefore seems that adjectival clausal modifiers occur rather frequently in Fragments, and they contribute to the high frequency of pronouns in this genre. For the detailed distribution of these syntactic relations across datasets, and for the role of those relations encoding subordination in general, see (4.4) and (4.5.5).

Let us consider other forms contributing to the high frequency of pronouns in Fragments. The second most frequent pronominal lemma in Fragments is the neuter pronoun *es*. *Es* is frequently used as expletive element in existential clauses (see 4.5.3). A high frequency of *es* could be due to a widespread use of these clauses in this genre. For an investigation into the use of existential clauses in Fragments, see (4.5.3). The impersonal pronoun *man* occurs frequently as well. I retrieved an example of the use of *man* with the following query, which return all the occurrences of *man* tagged POS-tagged with PRON (UPOS):

L=man&PRON

The sentence in (16) exemplifies the use of *man* in Fragments. In this case, *Man* is the nominal subject of the verb *sagen* ('say') in the main clause. In fact, the impersonal form moves the focus on the action rather than on the agent. Therefore, it goes without saying that the use of the impersonal form perfectly matches the communicative purpose of Fragments, which is performing judgements aiming to sound universal in their scope. The frequent use of declarative sentences without any personal nominal subject¹⁶⁰ serves this purpose.

(16) **Man** kann nicht sagen, daß etwas ist, ohne zu sagen, was es ist.¹⁶¹

One cannot say that something exists without saying what it is.

Table 46 shows the most frequent relations governing pronouns in Fragments. In contrast to what observed for nouns (Cf. Table 42), pronouns tend to mostly fill the syntactic functions of nominal subject and objects, while they are used more much more rarely as oblique arguments and nominal modifiers.

¹⁶⁰ 99% of occurrences of the lemma *man* in Fragments are nominal subjects (nsubj).

¹⁶¹ sent_id = athenaeum-f226-s2.

RANK	DEPREL	RF LIT
1	nsubj	53.0
2	obj	20.4
3	obl	6.8
4	expl	5.0
5	iobj	3.5
6	nmod	3.3
7	nsubj:pass	1.9
8	det	1.5
9	conj	1.3
10	root	1.3
11	parataxis	0.5
12	ccomp	0.3

Table 46 Distribution of dependency relations governing pronouns.

As previously done for nouns, I grouped RF of the syntactic functions played by pronouns according to the UD functional category. Table 47 summarizes the results. In this case too, the category of dependents includes the nominals only. For an inquire into the distribution of pronouns in the role of nonverbal clausal dependents, see (4.5.1).

FUNCTIONAL CATEGORY	RF LIT
Core Argument	78.8
Dependent	15.1
Root	1.3
Other	4.7

Table 47 Distribution of pronouns per functional category in LIT.

In the end, as for the subordinating conjunctions (SCONJ) and coordinating conjunctions (CCONJ), both of them are more frequent in Fragments with respect to both GSD and HDT (RF SCONJ_{LIT} =

1.5, RF *SCONJ*_{GSD} = 0.6, RF *SCONJ*_{HDT} = 1.0; RF *CCONJ*_{LIT} = 5.5, RF *CCONJ*_{GSD} 3.4, RF *CCONJ*_{HDT} = 3.1). This suggests a larger use of both coordination and subordination in Fragments with respect to the other genres. For the distribution of the dependency relations encoding subordination, see (4.4), while for an investigation of their syntactic position within sentences, see. As for *AUX*, this UPOS includes both auxiliaries and modal verbs. For a disambiguation, and an analysis of the modal verbs, see (4.5.4). As for verbs, their distribution is investigated in (4.5.2), within the section that focuses on some features of predicates.

4.4 Overall Distribution of Dependency Relations

Chart 8 displays the overall distribution of dependency relations in each dataset. Overall, the most frequent relation is *det* relation, showing a rather similar RF in all the datasets. This is not actually surprising, since we already observed that determiners are the second most frequent part of speech in all the datasets, after nouns (see Chart 7). As previously observed in Table 41, nouns can play a wide range of syntactic functions. On the contrary, determiners can only fulfil the function of determiners, therefore they are mostly governed through *det* relation. Even if it is intuitive, I however extracted all the rows containing the UPOS *DET* from the CoNLL-U file of both *LIT* and *GSD*, and I calculated how many of them bear the relation *det* in the *DEPREL* field: 95.4 % of all the UPOS *DET* in *LIT* are governed through *det* relation, and 98.5% in *GSD*. The *det* relation mostly spans from a noun to a determiner, therefore most determiners (*DET*) are supposed to have a noun as head. To test this assumption, and to check potential inconsistencies in the annotation as well, which can generate false positives causing the RF of *det* relations increases, I extracted all the nouns (*NOUN*) governing a determiner (*DET*) through *det* relation from both *LIT* and *GSD*¹⁶². I ran the following query:

NOUN >det DET

I divided them by the total occurrences of *det* relation in each data set. The result is that 95 % of *det* relations is governed by nouns (*NOUN*) in *LIT*, while 85% of *det* relations is governed by nouns in *GSD*. If every determiner must be headed by nouns, the opposite does not hold necessarily true, i.e. it is

¹⁶² I encountered problems in extracting data through the same query from *HDT*, since the file downloaded in CoNLL-U format only contained a small part of the results of the query, which did not even belong to the portion of the treebank used as dataset in this analysis, i.e. the training file.

Distribution of Dependency Relations (RF %)

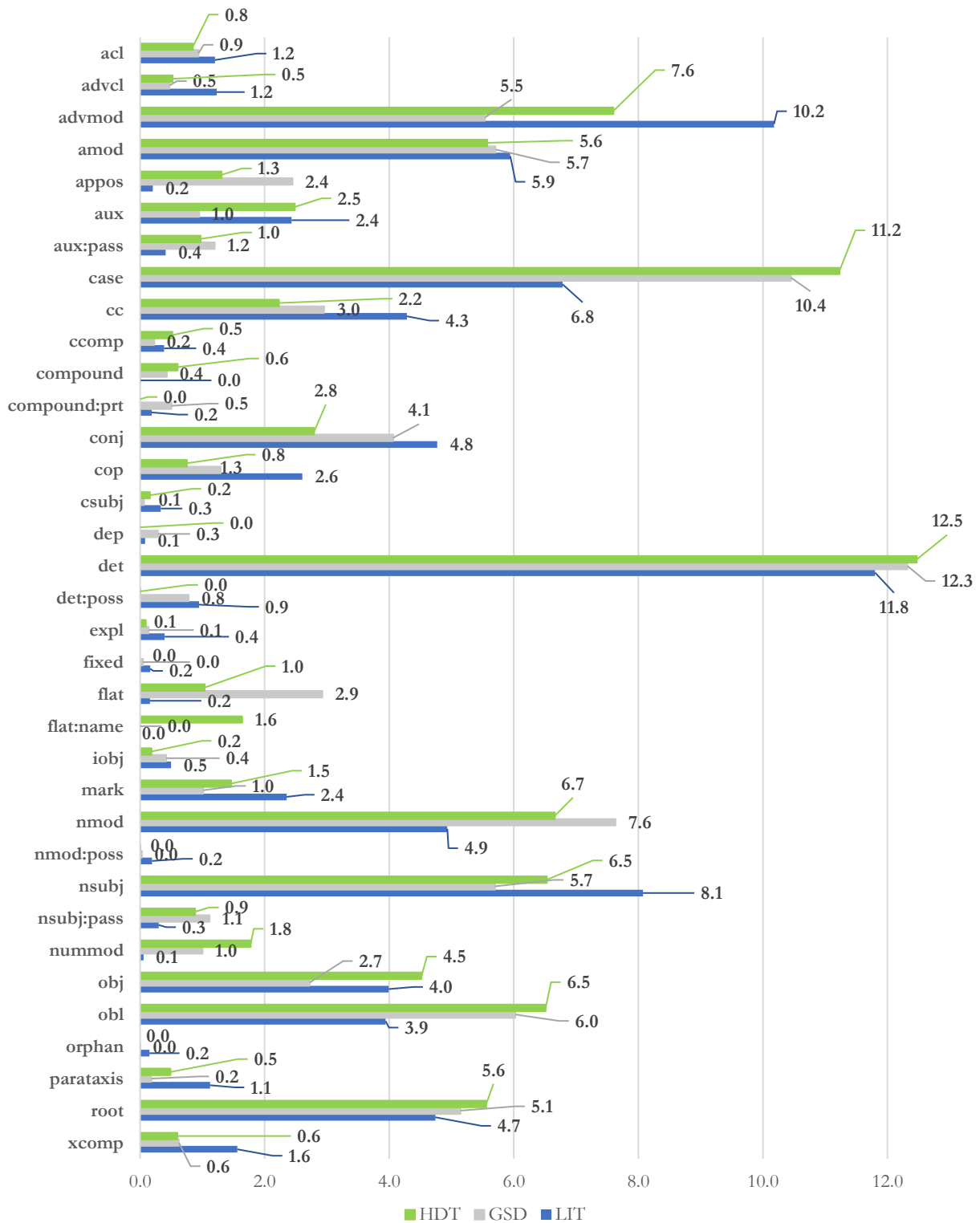


Chart 8 Overall distribution of dependency relations.

not automatic that any noun governs a determiner. For instance, a noun can be part of a prepositional phrases introduced by a simple preposition, or the determiner can be simply omitted from the noun phrase. This would also explain the gap between the distribution of the UPOS DET and that of the UPOS NOUN that we previously observed in Chart 7. The sentence in (17) exemplifies the use in Fragments of different nouns that are not modified by any determiner. The noun *Witz* ('wit') has no modifiers at all, while the nouns *Geist* ('spirit') and *Genialität* ('geniality') are both modified by adjectives and not by determiners.

(17) **Witz** ist unbedingt geselliger **Geist**, oder fragmentarische **Genialität**.¹⁶³

Wit is incredibly social Spirit, or fragmented geniality.

The example was retrieved through the following query, which returns any noun that does not govern any token, neither through *det* relation, nor through the subclass *det:poss* relation¹⁶⁴:

NOUN !(>det|>det:poss) _

Overall, many observations previously deduced from the distribution of UPOS and their syntactic functions are here confirmed by the distribution of dependency relations. The *case* relation is the second most frequent relation in both GSD and HDT. It mostly spans from nouns to prepositions (ADP), therefore it is obviously correlated to both the high RF of the UPOS ADP and of high RF of *obl* relation governing nouns observed in both these data sets. In this regard, the *obl* relation is overall much more frequent in both GSD and HDT. Similarly, *nmod* relation is more frequent in both GSD and HDT with respect to the Fragments. This tell us something more about the distribution of this relations in the datasets with respect to what observed about the role of this function among nouns. In fact, RF of this *deprel* among nouns was rather similar across all the three datasets. Conversely, Chart 8, shows a gap between the distribution of *nmod* between Fragments and the other two genres, especially with respect to GSD. Consequently, it means that even if a similar percentage of nouns is governed through this relation with respect to the whole number of nouns in each data, however, the number of nominal modifiers has overall a different quantitative impact on the global syntax of each dataset, if measured with respect to the whole number of syntactic relations. Therefore, we can say that the specification embodied through nominal modifiers appears to be much more frequent in GSD and HDT with respect to Fragments. *Nummod* relation is also much more frequent in GSD and especially HDT, and it is correlated to the higher use of temporal specification in this dataset addressed in 4.1. The frequency of *nsubj:pass* relation

¹⁶³ sent_id = lyceum-f9-s1.

¹⁶⁴ Which is used for possessive determiners.

is higher in HDT and especially in HDT. Passive voice is therefore confirmed being much more frequent in both GSD and LIT rather than in Fragments. It goes without saying that the distribution of the *aux:pass* relation, which governs the auxiliaries of the passive form, is correlated to the distribution of *nsubj.pass* relation, and it is due to the same reasons. Both *appos* relation and *flat* relation are more frequent in GSD, in accordance to what observed about the use of these relations for nouns. In fact, *appos* relation governs appositions, while *flat* relation especially governs members of exocentric (headless) semi-fixed MWEs to their heads, as in the case of names (proper nouns), which were demonstrated being very frequent in GSD.

As for those relation that are more frequent in Fragments with respect to both GSD and HDT, *aux* relation is more frequent in this genre, exactly as was the UPOS AUX. All the auxiliaries¹⁶⁵ are always governed by the *aux* relation. *expl* relation is more frequent in Fragments, which could be correlated to the rather high frequency of the neuter pronoun *es* observed in this genre, which can work as expletive element in existential clauses depending on the verb through *expl* relation (see Chapter 3). Fragments show a high RF of *advmod* relation, which is the second most frequent relation in this genre. This is certainly due to the high frequency of the UPOS ADV in this genre. In fact, *advmod* relation always spans from the predicate to the adverb. RF of *cc* relation is rather high, which is correlated to the distribution of coordinating conjunctions (CCONJ), which was observed much higher in Fragments with respect to the other genres. As a consequence, the distribution of *cconj* relations, which attaches to those items related to another item through syndetic coordination, is also higher in Fragments with respect to the other genres. One may ask whether this is due to a larger use of coordination between items within the same clauses, for instance between two nouns with the function of nominal subjects, or between two predicates. I therefore extracted a list of those *cconj* relations occurring between two predicates, both verbal and nonverbal. I ran the following query, which returns the occurrences of all verbs (verbal predicates) or tokens governing a copula (this was the formalization for extracting the nonverbal predicates) which govern a verb or another token governing a copula through *cconj* relation:

(VERB|(_ >cop _)) >conj (VERB|(_ >cop _))

The result is that 23.7% of *cconj* relation occur between predicates. An example of a *cconj* relation occurring between two predicates is reported in (18). In this case, the first item of the coordination is the main verb *gibt*, which is used in the existential construction *es gibt* ('there is'), while the second item is a nonverbal predicate, i.e. the adjective *seltner* ('rarer'). The two coordinate predicates and the coordinating conjunction are highlighted in bold.

¹⁶⁵ Both auxiliaries and modal verbs.

- (18) Es **gibt** so viel Poesie, **und** doch ist nichts **seltner** als ein Poem!¹⁶⁶
There is so much Poetry, **and** certainly nothing is **rarer** than a Poem!

Besides a larger use of syndetic coordination, Fragments appears to be marked by wider use of asyndetic coordination as well. In fact, the RF of *parataxis* relations, which attach two items which are coordinated though any explicit coordinating conjunction, is higher in this genre with respect to the others. (19) exemplifies the use of parataxis relation in Fragments. In this case, the parataxis relations spans from the main nonverbal predicate, which is the noun *Meister* ('masters') to the coordinated verbal predicate, which is the verb *haben* ('to have'). It was retrieved through the following query, which returns any token governed through *parataxis* relation:

_ <parataxis _

- (19) Die Alten sind **Meister** der poetischen Abstraktion: die Modernen **haben** mehr poetische Spekulation.¹⁶⁷

The ancients are masters of the poetic abstraction: the moderns have more poetic speculation.

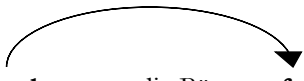
The distribution in Chart 8 also provides further information about the use of subordination in the three genres. In Chart 7, we observed a higher distribution of subordinating conjunctions (SCONJ) in Fragments with respect to the other genres, which suggested a more frequent use of subordination in this genre with respect to the other genres. This is confirmed by the distribution of *mark* relation, which attaches to markers of subordination, and whose frequency is much higher in Fragments with respect to the other genres. Overall, the distribution of those dependency relations encoding subordination not only confirms the assumption about a frequent use of subordination in Fragments, but also tells us more about how subordination is used in this genre. Overall, there is certainly a more frequent use of both clausal core-arguments and clausal dependents with respect to the other genres. The first category includes clausal subjects (csubj), clausal objects (ccomp), and open clausal complements (xcomp). In particular, *xcomp* relation is much more frequent in Fragments with respect to other genres. This relation is used for both predicates of non-finite clauses and for secondary predicates. For a distribution of the *xcomp* relation between these two classes of predicates, see Chapter 3. The second category includes adverbial clausal modifiers (advcl), i.e. the adverbial clauses, and adjectival clausal modifiers (acl), i.e. the relative clauses.

¹⁶⁶ sent_id = lyceum-f4-s1.

¹⁶⁷ sent_id = lyceum-f107-s1.

The distribution of adjectival clauses is slightly higher in Fragments, while the gap in the distribution of the adverbial clauses is remarkably higher. (20) exemplifies the use of an adverbial clause in Fragments. In this case, the adverbial clause spans from the verb *anfangen* to the past participle form *gebraucht* ('used'), which is the verbal part of the main predicate with passive voice (the auxiliary of the passive form is *werden*). It was retrieved through the following query, which return all the tokens which are governed by any token through *advcl* relation:

_ <advcl _



(20) Besonders in der Schlußcadence **werden** sie stark **gebraucht**, wenn die Bässe **anfangen** recht voll zu arbeiten.¹⁶⁸

They are strongly used in the closing cadence, when basses begin to fully work.

In the end, the higher frequency of nominal subjects in Fragments (*nsubj*) should be correlated to the higher frequency of both subordinate and coordinate clauses. Intuitively, the higher the number of clauses, i.e. of complex sentences, the higher the number of nominal subjects, since each clause can have only one subject. An overall analysis of the distribution of the dependency relations was provided. Some of the dependency relations reported in Chart 2 will be analyse more in detail over the next section.

4.5 An Investigation of Predicates

4.5.1 Distribution of Verbal and Nonverbal Predicates

Predicates can be either verbal or nonverbal. As for verbal predicates, the predicate only consists in a verb, either simple or complex (in this last case, it consists in an auxiliary occurring with the verbal form). As for nonverbal predicates, broadly speaking, the predicate consists in a nonverbal element usually occurring with a copula. The nonverbal element is mostly filled by a noun or an adjective, but, more rarely, it can also be filled a pronoun or other parts of speeches. For an overview on both verbal and nonverbal predication, see e.g. (Payne 1997), (Dixon 2012), (Roy 2013). (21) exemplifies the use of a verbal predicate, while (22) the use of a nonverbal predicate, both coming from the development set of the GSD treebank. In (21), the predicate is filled by the verb *empfehlen* ('to advice'), which is a non-finite form in this case, because the verbal predicate is complex: it consists in an auxiliary verb, i.e. *würde* ('to

¹⁶⁸ sent_id = lyceum-f49-s2. The topic of this Fragment is the musical cadence.

would’), which is combined with the verbal element. In German, in case of complex predicates, the auxiliary usually occupies the second position of the clause after the *Vorfeld*, i.e. after the first position of the clause, while the verbal or nonverbal element occupies the last position of the clause. In (22), the nonverbal predicate is filled by the adjective *Freundlich* (‘welcoming’), which indeed occupies the last position of the clause, while the copula is *war*, i.e. the *Präteritum* form of the verb *sein* (‘to be’), which occupies the second position.

(21) Ich **würde** nicht **empfehlen**.¹⁶⁹

I **would** not **advise**.

(22) Der Empfang **war** sehr **freundlich**.¹⁷⁰

The reception **was** very **friendly**.

I measured the distribution of both verbal and nonverbal predicates across the datasets. In terms of dependency relations, this can show how UPOS are distributed in the role of heads in predicates. Overall, this can tell us more about how the predication is embodied in each textual genre, also allowing for hypothesis about the functions of predication that are mostly used in one genre with respect to another one. In this regard, I extracted all those dependency relations spanning from predicates in main clauses, subordinate clauses, and coordinate clauses, and I measured the distribution of those UPOS playing the role of heads of these dependencies. In doing so, I also compared the distribution of both verbal and nonverbal predicates across different syntactic functions, for instance, in adverbial clauses with respect to main clauses or relative clauses. In detail, I considered the following relations: *root* relation, for the main predicates; *subj* relation, for clausal subjects, i.e. those predicates playing the role of subject of a higher clause; *comp* relation, for clausal complements, i.e. those predicates playing the role of object of a higher clause; *xcomp* relation, for open clausal complements, those predicates, which correspond to two classes of predicates: those in non-finite clauses introduced by the infinitival marker *zu*, and those working as secondary predicates.; *advcl* relation, for predicates in adverbial clauses, i.e. those predicates modifying the predicate of a higher clause as if adverbs; *acl* relation, for those predicates working as adjectival clausal modifiers, i.e. predicates occurring in relative clauses, which therefore modify nouns or pronouns of a higher clause; *parataxis relation*, for those predicates depending on other predicates through asyndetic coordination; *conj* relation, for those predicates depending on higher predicates through syndetic coordination. For the annotation of these syntactic relations in Fragments, see Chapter 3.

¹⁶⁹ sent_id = dev-s116

¹⁷⁰ sent_id = dev-s30

For each of these relations but *parataxis* and *conj*, I directly extracted all the rows from the CoNLL-U file of each dataset that bear a target dependency in the field DEPREL, and I calculated the RF of each of the following UPOS: VERB, NOUN, ADJ, and PRON. In fact, all the following relations necessarily govern predicates: *root*, *csubj*, *comp*, *xcomp*, *advcl*, *acl*. For instance, I extracted all the lines of the LIT file that had the relation *root* in the DEPREL field, and I measured the RF of each of the target UPOS, i.e. NOUN, PRON, ADJ, and VERB, with respect to the total number of lines. By contrast, in the case of the two relations encoding coordination, i.e. *parataxis* and *conj*, I cannot directly extract all the tokens bearing the deprel *parataxis* or *conj* in the DEPREL field, since coordination can also occur between elements that are not predicates (see Chapter 3). I therefore used SETS to extract only those occurrences of both *conj* and *parataxis* that actually governs predicates, both verbal and nonverbal. I run a different query for each of these dependency relations. As for verbal predicates, I searched for all the verbs governed by another token through both *parataxis* and *conj* relation. I run the following queries:

VERB <conj _

VERB <parataxis _

As for nonverbal predicates, I searched for all the nouns, adjectives, and pronouns that depend on another token through *parataxis* and *conj* relation, but this token, in turn, must depend on the lemma *sein* through *cop* relation.¹⁷¹ I run the following queries:

NOUN | ADJ | PRON <conj _ >cop L=sein

NOUN | ADJ | PRON <parataxis _ >cop L=sein

According to what observed in Chart 8 about the distribution of the *cop* relation across the datasets, we should expect RF of nonverbal predicates to be higher in LIT with respect to both GSD and HDT. Chart 3 summarizes the results concerning the overall distribution of predicates in each dataset. Those predicates POS-tagged with VERB are grouped under the *verbal*, while all those tagged with NOUN, ADJ and PRON are grouped as under *nonverbal*. The category *other* includes all those occurrences of tokens bearing one of the before-mentioned dependency relations in the DEPREL field, but which are POS-tagged with none of the target UPOS (i.e. neither NOUN, nor ADJ, nor PRON). The distribution of the single UPOS in the role of predicate is analysed better later.

¹⁷¹ I opted for declaring the UPOS and the lemma of *sein* in order to reduce the number of false positives, if any.

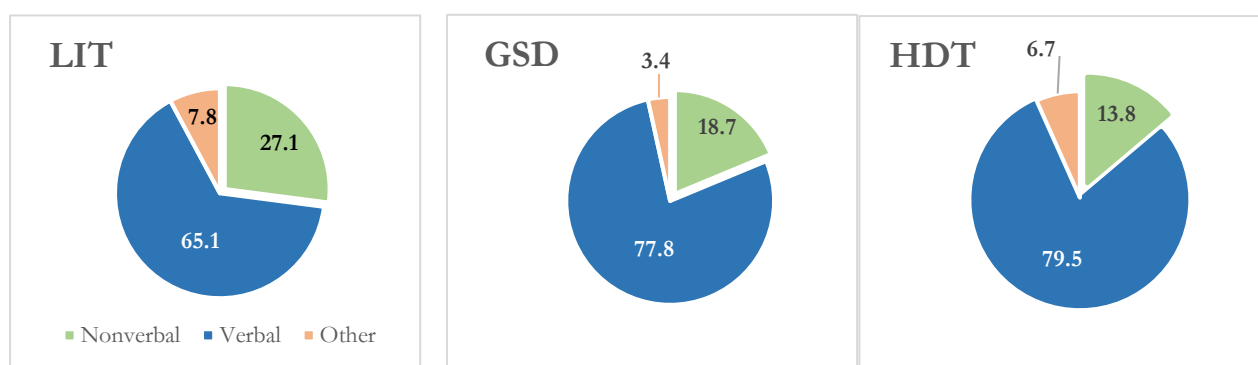




Chart 9 Distribution of verbal and nonverbal predicates (RF%).


As shown in Chart 9, as expected, the overall RF of nonverbal predicates is much higher in LIT with respect to both GSD and, especially, to HDT. In terms of textual genres, therefore, we can say that Fragments tend to use the nonverbal predication more frequently with respect to both the texts of Wikipedia and the web news. By contrast, web-news texts tend to use nonverbal predication rather rarely, while texts from Wikipedia tend to use it more frequently with respect to web news, but however at a much lower extent with respect to Fragments.

This result allows a possible semantic interpretation of the use of predication across the three genres. Let us consider three of the semantic macro-functions that are usually associated with nonverbal predication, i.e. attribution, inclusion, and equation, on which see, among the others, (Payne 1997), (Dryer 2007). Attribution is embodied through adjectives in the role of nonverbal predicates. According to Payne (1997): “Predicate adjectives are clauses in which the main semantic content is expressed by an adjective”. In other words, attribution generates clauses in which the adjective used as nonverbal predicate convey a quality of the subject. Inclusion is embodied through nonverbal predication (Payne 1997) “when a specific entity is asserted to be among the class of items specified in the nominal predicate”. Equation generates equative clauses, which are (Payne 1997) “those which assert that a particular entity (the subject of the clause) is identical to the entity specified in the predicate nominal”. In terms of communicative purposes, as said above, Fragments are very judgement-oriented: even if often in a cryptic and ironic way, they ultimately aim to judge many aspects of their age, especially concerning artists, work of arts, and, broadly speaking, intellectual issues in general. Consequently, the function of attribution must be very widely used in these texts. Some cases of nonverbal predicates encoding attribution in Fragments are reported in (23), (24), and (25). Both the copula and the predicate adjective are highlighted in bold. I retrieved them through the following query, which returns all the adjectives which governs any token through *nsubj* relation and which, at the same time, govern a copula.

ADJ >nsubj _ >cop _


(23) Nichts **ist verächtlicher** als trauriger Witz.¹⁷²
Nothing **is more despicable** than the sad wit.


(24) Die Römer **sind** uns **näher und begreiflicher** als die Griechen.¹⁷³
The Romans **are closer and more understandable** for us than the Greeks.


(25) Alle Gattungen **sind gut**, sagt Voltaire, ausgenommen die langweilige Gattung.¹⁷⁴
All genres **are good**, Voltaire said, apart from the boring genres.

In (23), the indefinite pronoun *Nichts* ('nothing') is the nominal subject of the nonverbal predicate, whose nonverbal element is the comparative adjective *verächtlicher* ('more despicable'). In this case, the attribution is used to judge a certain type of "way to be witty" (*Witz*), especially by some authors in their literary works. In (24), the attribution is expressed through a comparative adjective as well, i.e. *näher* ('closer'), which is coordinated to another comparative adjective, i.e. *begreiflicher* ('more understandable'). In this case, the author performs a judgement about cultural values by comparing some values of his age with those of the two classical civilizations of Romans and Greeks. In (25), the author reports a quotation of Voltaire¹⁷⁵, which somehow anticipates the viewpoint of the early Romanticism about the use of irony, which is one of the key values of this literary movement, and therefore he considers this value as always desirable in a work of art. In this case, the function of attribution works on the nominal subject *Gattung* ('genre'), while the nonverbal predicate is the adjective *gut* ('gut').

However, the function of attribution must not be the sole one causing a rather frequent use of nonverbal predication in Fragments. In fact, the function of equation is supposed to be largely used as well. I report some examples in (26), (27), and (28). I ran the following query, which returns all the nouns governing any token through *nsubj* relation and, at the same time, a copula.

NOUN >nsubj _ >cop _

¹⁷² sent_id = lyceum-f17-s1.

¹⁷³ sent_id = lyceum-f46-s1.

¹⁷⁴ sent_id = athenaeum-f324-s1.

¹⁷⁵ The famous French author of the XVIII century.

(26) Die Romane **sind** die sokratischen **Dialoge** unserer Zeit.¹⁷⁶
 Romances **are** the **Socratic Dialogues** of our age.

(27) Humor **ist** gleichsam der **Witz** der Empfindung.¹⁷⁷
 Humor is, so to speak, the wit of the emotion.

(28) Anmut **ist** korrektes **Leben**.¹⁷⁸
 Gracefulness **is** correct **life**.

In (26), the nominal subject *Romane* ('Romance') is equated with the *sokratische Dialoge* ('Socratic Dialogues') through a nonverbal predicate, whose nonverbal element is the noun *Dialogue*. In this case, the author uses the equation to compare two distinct literary genres belonging to two different historical ages, in order to underline a strong parallelism between them. In (27), an equation is drawn between two abstract entities: the nominal subject *Humor* ('Humor') is equated with the predicative noun *Witz* ('wit'), which is followed by the nominal modifier *Empfindung*, to specify the scope of the entity embodied in the nonverbal predicate. In (28), again, the author draws an equation between two abstract entities: the nominal subject *Anmut* ('grace') is equated to the nominal predicate *Leben* ('life'), with the adjective *korrektes* ('correct') specifying the entity embodied in the nonverbal predication. In all these cases, the equation clearly serves the purpose of giving a sort of universal and peremptory scope to the message that conveys a property of the entity that undergoes equation. This strategy perfectly matches the general speech act performed by many Fragments, which, as said before, essentially performs judgements. These examples may therefore suggest a widespread use of the equation in this genre. Moreover, it could be largely used to trace identities between abstract entities rather than between concrete entities. The combination of the abstract use of this function with that of attribution could be the reason behind the higher frequency of nonverbal predication in Fragments. Obviously, we are not arguing that other functions, such as inclusion, are totally absent in this genre. In this respect, an example of a nonverbal predicate conveying inclusion in Fragments is reported in (29), where the noun *Kritiker* ('critic') is included

¹⁷⁶ sent_id = lyceum-f26-s1.

¹⁷⁷ sent_id = athenaeum-f237-s1.

¹⁷⁸ sent_id = lyceum-f29-s1.

in the class embodied by the noun *Leser* ('reader'), which, in turn, is then modified by a relative clause, which then specifies a property of a member of this class.

(29) Ein Kritiker **ist** ein **Leser** , der wiederkäut.¹⁷⁹
A critic **is** a **reader** who ruminates.

However, from what observed by these examples, and from what we know about the nature of the genre and its communicative purposes, this function must contribute to the high frequency of nonverbal predicates in this genre to a lower extent. At the same time, we are not claiming that these two functions are not used in the other genres. But we do not expect neither attribution nor equation, especially equation between abstract entities, occurring very frequently neither in GSD nor in HDT. In fact, most of the texts from the GSD come from Wikipedia, and, in this genre, I would expect a larger use of inclusion. If we think of the biographical and encyclopaedic nature of most of Wikipedia's texts, inclusion should be especially used to define the category to which the entity described in a Wikipedia' page belongs to. I report some examples of this use of the inclusion in GSD' in (30), (31), and (32).

(30) Trancecore **ist** ein **Subgenre** des Hardcore Techno, der aus der Vermischung von Trance und Hardcore Techno entstand.¹⁸⁰
Trancecore **is** a **subgenre** of Hardcore Techno, which consists in a mix of Trance and Hardcore Techno.

(31) Duell an dem Missouri **ist** ein **Spätwestern** aus dem Jahr 1976.¹⁸¹
Duel on the Missouri is a late western of the year 1976.

(32) Er **war** der erste **Senator** der Class - 3 - Kategorie aus Maryland.¹⁸²
He was the first Senator from the third-class category from Maryland.

¹⁷⁹ sent_id = lyceum-f27-s1.

¹⁸⁰ sent_id = train-s2376


¹⁸¹ sent_id = train-s2342.

¹⁸² sent_id = train-s2282.


Both (30) and (31) exemplifies a common definition of Wikipedia. In (30), the nominal subject *Trancecore*, a music genre, is included in the broader category embodied by the noun *Subgenre*, which, in turn, is modified by the nominal modifier *Hardcore Techno*. In (31), the proper noun *Duell an dem Missouri* ('Duel on the Missouri') is included in the category of *Spätwestern* ('late western'), a film genre. In (31), the personal pronoun *Er*¹⁸³ is included in the category embodied by the noun *Senator*, which is modified by the compound noun *Class - 3 – Kategorie* to specify a sort of subtype of the category.

In HDT, I expect a similar use of the nonverbal predication, which is therefore very influenced by the function of inclusion. (33) and (34) exemplify the use of this function in web-news texts. In (33), the noun *US-amerikanischen Online-Medien* ('big American online media') embodies the entity which is included in the category embodied by the noun *Mitglieder* ('member'), which is therefore the predicative part of the nonverbal predicate, as well as root of the sentence. In (34), the compound noun *Zehn-Punkte-Plan* ('ten-point plan') is included in the category of *Kernstück* ('core').

(33) Fast alle großen US-amerikanischen Online-Medien **sind** ebenfalls **Mitglieder** des IAB.¹⁸⁴
 Almost all the big American online media companies **are** also **members** of the IAB.



(34) **Kernstück** der Initiative **ist** ein **Zehn-Punkte-Plan**.¹⁸⁵
 The core of the initiative is a ten-point plan.



In conclusion, I assume that the higher frequency of nonverbal predicates in Fragments is caused by a large use of all the three macro-functions associated to nonverbal predication, but especially of two functions, i.e. both attribution and equation. In fact, both these functions match the speculative nature of this literary genre, and its dominant communicative purpose, which is mainly judgment oriented. By contrast, I suppose that the lower frequency of nonverbal predicates in both GSD and HDT is caused by a limitation in the use of functions embodied by nonverbal predication, which should be more oriented toward inclusion rather than on attribution or equation. In fact, this should be the main function in texts with a high degree of reference to real-world entities, and with a communicative purpose that is mainly informative. Let us now move on to consider how the verbal and nonverbal predication is distributed across different syntactic functions. In this respect, I report the distribution of UPOS in the role of heads of each of the dependency relations encoding predication, for each dataset. Table 48 shows the

¹⁸³ More precisely, the entity to which the personal pronoun refers, i.e. a person.

¹⁸⁴ sent_id = hdt-s112081

¹⁸⁵ sent_id = hdt-s112324


distribution of UPOS as heads of the *root* relation, i.e. in the role of main predicates. To retrieve them, I ran the following query, which returns all the nouns, adjectives and pronouns that are not governed by any relation (!<)¹⁸⁶, and that govern the lemma *sein* through cop relation.

NOUN|ADJ|PRON !< _>cop L=sein

DEPREL	UPOS	RF LIT	RF GSD	RF HDT
root	NOUN	23.7	12.4	9.0
	PRON	2.2	0.4	0.7
	ADJ	11.6	8.1	4.9
	VERB	54.6	76.4	78.9
	OTHERS	7.8	2.7	6.4

Table 48 Distribution of UPOS in the role of main predicates (root nodes).

As shown in Table 48, the distribution of UPOS as main predicates is rather similar between GSD and HDT. In both these datasets, the vast majority of main predicates are verbal predicates. Nonverbal predicates are mostly filled by nouns, while the RF of adjectives in this role is overall rather low, especially in HDT. On the contrary, the RF of nonverbal predicates in the role of main predicates is very high in Fragments, and it is clearly above the average reported in Chart 9. In fact, if we sum the distribution of adjectives, nouns and pronouns playing the role of head of the *root relation* in this genre, RF of nonverbal predicates add up to 36.5%: more than one main predicate out of three is filled by a nonverbal element, and the majority of these nonverbal elements are nouns. An example of a nonverbal predicate in a main clause in Fragments is reported in (35), where the noun *Instikt* is the nonverbal predicate, while the noun *Ironie* is the nominal subject of the predicate, which, in turn, is modified by the possessive nominal modifier *Lessings* ('Lessing's'¹⁸⁷):

- (35)  Lessings Ironie **ist Instinkt.**¹⁸⁸
Lessing's Irony **is Instict.**

¹⁸⁶ The *root* relation cannot be specified in SETS. However, the *root* relation is the sole one fulfilling the condition of not depending by any other token, since the head of the root node is always 0. I therefore exploited this property to extract root nodes.

¹⁸⁷ Lessing is the surname of the German writer Gotthold Ephraim Lessing, who lived in the 18th century.

¹⁸⁸ sent_id = lyceum-f108-s10.

DEPREL	UPOS	RF LIT	RF GSD	RF HDT
csubj	NOUN	4.4	3.6	2.0
	PRON	1.5	0.6	0.1
	ADJ	7.4	3.6	3.6
	VERB	76.5	92.3	89.5
	OTHERS	10.3	0.0	4.8
ccomp	NOUN	12.2	6.0	3.5
	PRON	7.1	0.2	0.4
	ADJ	12.8	7.7	6.7
	VERB	57.7	80.1	84.5
	OTHERS	10.3	6.0	4.9
xcomp	NOUN	15.5	24.2	3.4
	PRON	1.7	1.1	1.9
	ADJ	23.8	24.5	16.8
	VERB	54.6	34.4	72.4
	OTHERS	4.4	15.9	5.6

Table 49 Distribution of UPOS in clausal core-arguments.

Table 49 shows the distribution of verbal and nonverbal predicates in those dependencies that are classified as clausal core-arguments in UD, i.e. those dependent clauses playing the role of clausal subjects, clausal objects, or open clausal complements (Cf. Chapter 3). I retrieved them through the following query:

NOUN|ADJ|PRON <csubj _>cop L=sein

NOUN|ADJ|PRON <ccomp _>cop L=sein

NOUN|ADJ|PRON <xcomp _>cop L=sein

As for *csubj* relation, i.e. clausal subjects, the distribution is rather similar between GSD and HDT, with verbs filling the vast majority of the predicates with this syntactic function, in both these datasets. As for Fragments, unlike the *root* relation, the distribution of verbal predicates is closer to the average distribution observed in Chart 9 for this dataset. By contrast, the distribution of nonverbal predicates clearly deviates from the average, and clearly remains beneath it. Verbal predicates therefore appear to occur much more frequently in this function with respect to nonverbal predicates. The frequency of the category *other* turned out surprisingly high. I therefore investigated this deviation. The result is that the

low RF of nonverbal predicates is correct, while the category *other* includes some false positives. In detail, almost 87% of the tokens that were assigned to this category are POS-tagged as AUX. This is due to a mistake in POS-tagging: 66% of these tokens tagged as AUX are actually occurrences of the verb *haben* ('to have') used as a verb in possessive clauses, and not as an auxiliary, while 34% of them are occurrences of either *sein* or *werden* ('to become'), which are actually used as verbs and not as auxiliaries. An occurrence of the verb *haben* as clausal subject that was incorrectly tagged as AUX in Fragments is reported in (36).

- (36) Wer Fantasie, oder Pathos, oder mimisches Talent **hat**, müßte die Poesie lernen können, wie jedes andre Mechanische.¹⁸⁹
 Whoever **has** Fantasy, Phatos, or Talent for mimics should be able to learn poetry, like every other Mechanik.

Consequently, RF of verbal predicates in the role of clausal subjects in LIT is even higher with respect to that reported in Table 49. Nonverbal predicates therefore tend to fill the role of clausal subject rather rarely in all the genres. (37) exemplifies the use of a nonverbal predicate as clausal subject in Fragments. In this case, the clausal subject follows the main clause, and the nonverbal predicate in the role of clausal subject is the adjective *unbeweglich*, followed by the copula *ist* ('is'), i.e. the third-person singular of *sein*. In those subordinate clauses with nonverbal predicates, the copula usually occupies the last position of the clause, immediately after the nonverbal element. Here, the main predicate is the verb *erscheint*, i.e. the third singular person of *erscheinen* ('to seem'), which is followed by a secondary predicate, i.e. the adjective *ruhig* ('quite').

- (37) Schlechthin ruhig **erscheint**, was in Rücksicht der Außenwelt schlechthin **unbeweglich ist**.¹⁹⁰
 It seems very calm, what is immovable to the eyes of the outside world.

As for *comp* relation, the distribution of nonverbal predicates is much higher in Fragments with respect to the *subj* relation. In this genre, nouns and adjectives are almost equally distributed in the role of clausal complements, and the use of pronouns in this role increases with respect to the *subj* relation. (38) exemplifies the use of a nonverbal clausal complement in Fragments. In this case, the comparative adjective *besser* ('better') is the nonverbal predicate in the role of clausal object, while the main predicate

¹⁸⁹ sent_id = athenaeum-f250-s1.

¹⁹⁰ sent_id = bluethenstaub-f111-s1.

is the verb *glaubte*, which is the third-person singular of *glauben* ('to believe') at *Praeteritum*. Interestingly, the structure of the subordinate clause does not show the canonical syntactic order, since the verbal predicate occupies the second position of the clause as if it were in a main clause, which is a verb-second clause, and not the last position, as it should be for a subordinate structure, which is usually a verb-final clause. For this reason, the copula, i.e. the *Konjunktiv I* form *seien* ('would be'), precedes the nonverbal element of the subordinate predicate, unlike what overserved in (37). Moreover, the complementizer *dass* at the beginning of the clausal object is missing. This clause therefore shows a word order that frequently occurs in subordinate clauses in the contemporary spoken German. In literature, the verb-second order in the subordinate clauses of the Spoken German has been mainly studied in causative clauses, see e.g. (Gaumann 1983), (Antomo and Steinbach 2010).

- (38) Schon Sophokles **glaubte** treuherzig, seine dargestellten Menschen **seien besser** als die wirklichen.¹⁹¹
 Already Sophocles naïvely believed that the men that he represented **were better** than the real ones.

As for GSD and HDT, again, the distribution of nonverbal predicates as clausal complement is rather similar between them, and it is very low in both cases. Moving to the distribution of nonverbal predicates in the role of open clausal complements (*xcomp* relation), it varies a lot across the datasets. First of all, the relation *xcomp* can encode two main different types of predicates in UD for German. On the one hand, it is used for non-finite predicates in infinitival clauses, i.e. those subordinate clauses without their own subject, and whose predicate, both verbal and nonverbal, is in non-finite form, introduced by the infinitival marker *zu*. When a non-finite clause is a final clause (or purpose clause), it is also introduced by the infinitival marker *um*. In any case, in this type of clauses, all the infinitival markers depend on the verbal or nonverbal element of the non-finite predicate, as well as the copula. In turn, the verbal or nonverbal element of the predicate depends back to the higher predicate through *xcomp* relation (Cf. Chapter 3). An example of this use of the *xcomp* relation in Fragments is reported in (39). In this case, the non-finite clause is a final clause introduced by *um*, while the non-finite predicate is a nonverbal predicate filled by the adjective *naiv* ('naive'). The non-finite form of the copula, i.e. *sein*, occupies the last position of the non-finite clause.

- (39) Es gibt sentimentale Kunsturteile, denen nichts fehlt als eine Vignette und ein Motto, **um** auch


vollkommen **naiv** zu **sein**.¹⁹²
 There are impulsive evaluations on art, which only lack a cartoon or a motto to **be** totally **naiv** as well.

¹⁹¹ sent_id = lyceum-f125-s1

¹⁹² There are impulsive evaluations on art, which only lack a cartoon or a motto to **be** totally **naiv** as well.

On the other hand, the *xcomp* relation is also used for those so-called secondary predicates. They occur, for instance, in this English construction: “She declared the cake *beautiful*”. In this case, *beautiful* is an adjective playing the role of nonverbal secondary predicate of the verb *declared*¹⁹³. In secondary predicates, *xcomp* relation spans from the first predicate to the secondary (i.e. the secondary predicate is the dependent), which is usually a verb. (40) exemplifies the use of *xcomp* in the double predication in Fragments. The two predicates are highlighted in bold. In this case, the second predicate is the adjective *idealish* (‘ideal’), which depends on the verb *wird*, i.e. the third-person singular of *werden* (‘to become’), through *xcomp* relation. The form *wird* is the main predicate.

(40) Dadurch **wird** es **idealisch**.¹⁹⁴
 Through this it **becomes** **ideal**.



Let us focus on Fragments. In this genre, RF of nonverbal predicates as open clausal complements is overall very high with respect to the average value observed in Chart 9: overall, 41% of predicates in this syntactic role are nonverbal, and a significant part of them are adjectives. One may now ask how many of these predicates are predicates of non-finite clauses, and how many are secondary predicates instead. To measure such distribution, I run three queries. To retrieve verbal predicates in non-finite clauses, I extracted all those verbs (VERB) governed through *xcomp* relation by a token that, in turn, governs the infinitival marker *zu* through *mark* relation. I ran the following query:

VERB <xcomp _>mark zu

Second, to retrieve all those nonverbal predicates in non-finite clauses, I extracted all those nonverbal predicative elements (NOUN, ADJ, and PRON) governed through *xcomp* relation by a token that, in turn, governs both an infinitival marker *zu* through *mark* relation, and, at the same time, the lemma *sein* through *cop* relation. I ran the following query:

NOUN|ADJ|PRON <xcomp _>mark zu >cop L=sein

¹⁹³ The *xcomp* relation is used for core arguments of clausal predicates only, therefore it is never used for other instances of secondary predication (Cf <https://universaldependencies.org/u/dep/xcomp.html>).

¹⁹⁴ sent_id = lyceum-f23-s2

Third, to retrieve all the occurrences of secondary predicates, I extracted all the verbal and nonverbal elements that are governed through *xcomp* relation, but that, in turn, do not govern neither a copula nor the marker *zu*. I therefore ran the following query:

VERB|NOUN|ADJ|PRON <xcomp _ !>mark _ !>cop _ L=sein

In the end, I measured RF of each of these groups of open clausal complements with respect to the total occurrences of *xcomp* relations. Table 50 summarizes the results.

Open Clausal Complement	RF LIT
non-finite verbal predicate	31.5
non-finite nonverbal predicate	1.9
secondary predicate	61.4
other	5.2

Table 50 Distribution of predicates in the role of open clausal complements in LIT, out of the total occurrences of *xcomp*.


Interestingly, as shown in Table 50, most predicates occurring as open clausal complements in LIT are actually secondary predicates. Table 51 reports the distribution of the POS in the role of secondary predicates in Fragments.

UPOS (Secondary Predicate)	RF LIT
ADJ	37.0
NOUN	23.4
VERB	37.0
PRON	2.6

Table 51 Distribution of UPOS in the role of secondary predicates.


As shown in Table 51, the two most frequent UPOS in the role of secondary predicates in Fragments are adjectives and verbs, while nouns occur less frequently in this function. An example of a verb used as secondary predicate is shown in (41). In this case, the first predicate is the verb *lassen* ('leave'), whose

nominal subject is the pronoun *sie* ('they'), while the secondary predicate is the non-finite verb *greifen* ('to grasp'), which depends on *lassen* through *xcomp* relation, and, in this case, is also coordinated to the verb *vorhalten*. Both the first and the second predicate are highlighted in bold.




(41) sie **lassen** sich nicht mit Händen **greifen**, und dem andern vorhalten.¹⁹⁵
 they don't **leave** themselves be **grasped** with hands, and be reproached.

An example of a noun in the role of secondary predicate in Fragments is reported in (42). In this case, the nonverbal secondary predicate is the noun *Imperativ* ('imperative'), which depends on the first predicate, i.e. the verb *nennen* ('to name'), through *xcomp* relation. The noun *Kantianer* ('Kantian') is the nominal subject of *nennen* ('call'), while the pronoun *dies* is the direct object.



(42) Ein Kantianer würde dies den kategorischen **Imperativ** der Genialität **nennen**.¹⁹⁶
 A Kantian would call this the categoric **imperative** of Geniality.

An example of an adjective used as secondary predicate is reported in (43). In this case, the adjective *ruhig* ('quiet') is the second predicate, while the verb *erscheint*, i.e. the third-person singular of *erscheinen* ('to appear') is the main predicate.



(43) Daher **erscheint** das Schöne so **ruhig**.¹⁹⁷
 From there appears the Beautiful so quiet.

As far as GSD is concerned, RF of nonverbal predicates as open clausal complements increases with respect to LIT: almost 50% of open clausal complements are nonverbal. (44) exemplifies a nonverbal predicate used in this role. In this case, the nonverbal predicate is the noun *Eigenkonstruktion* ('unique

¹⁹⁵ sent_id = lyceum-f44-s3.

¹⁹⁶ sent_id = lyceum-f16-s3.

¹⁹⁷ sent_id = bluethenstaub-f111-s4.

construction’), while the copula is the non-finite form *sein* occupying the last position of the non-finite clause. The main predicate is the verb *schien*.



- (44) Das Fenster schien eine **Eigenkonstruktion** aus zwei einfachverglasten Fenstern zu **sein**.¹⁹⁸
 The window seems to **be** a **unique construction** of two simple glassed windwos.

However, RF of the category *other* is unexpectedly high in this dataset as well (RF = 15.9). I therefore investigated this deviation. It turned out that the 15.9% is made up as follows: 8.2% are actually proper nouns (PROPN), while 6.6% are actually prepositions (ADP)¹⁹⁹. Proper nouns can actually play the role of open clausal complements in some constructions. Furthermore, given the high RF of proper nouns in GSD, one may expect that a certain amount of them can occur as open clausal complements. (45) exemplifies the use of a proper noun as nonverbal secondary predicate. In this case, the secondary predicate is the acronym CSU (which is tagged as PROPN in UD) while *wird*, i.e. the third-person singular of *werden* (‘become’), is the main predicate. For the matter of clarity, Figure 39 shows the same sentence of (45) in the format of the SETS’ output, in which the dependencies relations are displayed. The acronym CSU is the first item of a compound noun, whose second token is the hyphen, and whose third token is the noun *Generalsekretär* (general secretary’). According to UD 2.5, these two tokens should both depend on the first item of the compound, i.e. *CSU*, through *compound* relation (the compound is endocentric)²⁰⁰. Here there is therefore a parsing error, since Figure 38 shows how *Generalsekretär* depends back on the main verb *wird*, actually.²⁰¹

- (45) Sein Nachfolger **wird** der frühere **CSU - Generalsekretär** Erwin Huber.²⁰²
 Her successor becomes the previous general secretary of CSU Erwin Huber.

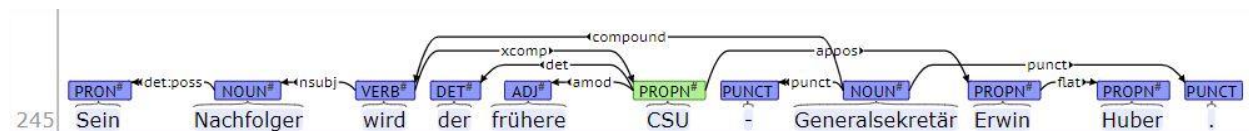


Figure 38 An example of a proper noun with the function of nonverbal secondary predicate in GSD.

¹⁹⁸ sent_id = train-s806.

¹⁹⁹ RF of both PROPN and ADP is intended with respect to the total occurrences of dependents of the *xcomp* relation.

²⁰⁰ Cf. <https://universaldependencies.org/u/dep/flat.html>.

²⁰¹ Furthermore, the choice of the root node in this sentence is arguable. I would have chosen *Nachfolger* as root node, and *CSU* as nominal subject.

²⁰² sent_id = test-s916 (this example was mistakenly taken from the development set rather than from the training set).

Conversely, the presence of some ADP in the role of nonverbal predicates, therefore as dependents of *xcomp* relation, is clearly an error of POS-tagging. In fact, function words are never allowed to be heads of dependencies in the UD scheme, apart from some peculiar cases of ellipsis.²⁰³ (46) illustrates a case of a preposition that was mistakenly parsed as dependent of the relation *xcomp* in GSD. In this case, the *xcomp* relation spans from the main predicate *gilt* (the head of the relation), i.e. the third-person singular of *gelten* ('to be valid'), to the preposition *als* ('as'), which precedes the actual second predicate, i.e. the past-participle *erfüllt* ('completed'). Therefore, the *xcomp* relation was expected to span from *gilt* to *erfüllt*, with *als* depending on *erfüllt* through *case* relation. For the matter of clarity, Figure 40 illustrates the sentence reported in (46) in the form of the SETS' output.

- (46) Die Mission gilt **als erfüllt**, sobald die Cruise Missile zerstört wurde.²⁰⁴
 The mission was considered **as** completed, as soon as the Cruise Missile was destroyed.

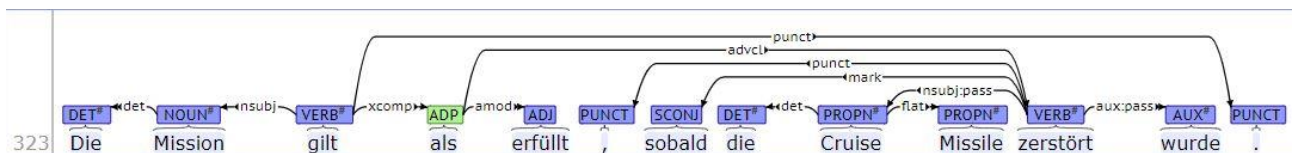


Figure 39 An example of a preposition that was incorrectly parsed as dependent of an *xcomp* relation in GSD.

As for the distribution of *xcomp* relation between non-finite clauses and secondary predicates, I ran the same queries that I had previously run on Fragments (see above). Given the significant RF of proper nouns as open clausal complements in GSD, I also included the UPOS PROPN in the queries:

NOUN|PRON|ADJ|PROPN <xcomp _>mark zu >cop L=sein

VERB|NOUN|PRON|ADJ|PROPN <xcomp _!>mark _!>cop _

Table 52 reports the results.

²⁰³ Cf. <https://universaldependencies.org/u/overview/specific-syntax.html#ellipsis>.

²⁰⁴ sent_id = train-s5018.

Open Clausal Complements	RF GSD
verbal non-finite predicates	18.0
nonverbal non-finite predicates	1.0
secondary predicates	64.8
others	16.3

Table 52 Distribution of predicates in the role of open clausal complements in GSD, out of the total occurrences of *xcomp*.

As shown in Table 52, most of the open clausal complements are secondary predicates in GSD too. Interestingly, RF of the category *others* is rather high (16.3%). Given the problem detected above about prepositions, I tested how many prepositions actually occur as secondary predicates. I ran the following query:

ADP <xcomp _ !>mark _ !>cop _

As a result, it turned out that 8% of the open clausal are actually filled by prepositions (ADP) in the role of secondary predicate. An example is reported in (47). In this case, the secondary predicate should be the noun *Sondierungsbrief* ('exploratory letter'), which should therefore depend on the main predicate, i.e. the verb *sieht*, ('saw'), with the preposition *als* depending on *Sondierungsbrief* through *case* relation. Conversely, *als* depends on *sieht* through *xcomp* relation, while *Sondierungsbrief* depends on *als* through *nmod*. For the matter of clarity, I also report the SETS's output in Figure 41.

- (47) Grayson **sieht** dieses Schreiben **als** Sondierungsbrief.²⁰⁵
 Grayson saw this writing as exploratory letters.

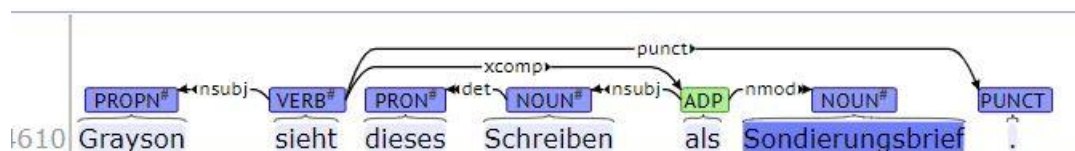


Figure 40 An example of a preposition that was incorrectly parsed as secondary predicate in GSD.

Table 53 shows the distribution of UPOS in the role of secondary predicates in GSD.

²⁰⁵ sent_id = train-s7945.

UPOS (secondary predicate)	RF GSD
ADJ	30.3
NOUN	29.7
VERB	23.5
ADP	10.2
PROPN	8.5

Table 53 Distribution of UPOS in the role of secondary predicates in GSD.

As shown in Table 53, RF of verbal secondary predicates is remarkably lower with respect to GSD. As for HDT, RF of nonverbal predicates as open clausal complements is much lower with respect to both LIT and GSD, with more than 76% of *xcomp* relation occurring with verbs. Since I already outlined a detailed comparison between LIT and GSD concerning the distribution of open clausal complements, I opted for not investigating further into the distribution of the specific functions of *xcomp* relation in this dataset.

Table 54 shows the distribution of UPOS in the role of adverbial clauses. The RF of nonverbal predicates in LIT and HDT overall follows the average trend observed in Chart 9. In both cases, adjectives occur more frequently in this role with respect to nouns, especially in Fragments.

DEPREL	UPOS	RF LIT	RF GSD	RF HDT
advcl	NOUN	5.6	1.4	1.4
	PRON	1.2	0.1	0.1
	ADJ	10.4	6.1	17.9
	VERB	74.6	91.7	70.2
	OTHERS	8.2	0.7	10.3

Table 54 Distribution of UPOS in the role of adverbial clause.

Conversely, the situation is rather different in GSD, where more than 90% of predicates in this role are verbal predicates. An example of an adverbial clause with an adjective as nonverbal predicate is reported in (48) from Fragments. In this case, the predicate of the adverbial clause introduced by *wenn* is the

adjective *ruchlos* ('heinous'), and the copula is *ist*, while the noun *Dichterwerk* ('work of poetry') is the nominal subject of the adverbial clause. The main predicate is the verb *haben*.



- (48) Eigentlich haben sie 's recht gern , wenn ein Dichterwerk ein wenig **ruchlos ist**, besonders in der Mitte;²⁰⁶
Actually they appreciate it, when a work of poetry **is** a bit **shocking**, especially in the middle.

The use of a verb as verbal predicate in an adverbial clause in GSD is exemplified in (49). In this case, the adverbial clause is introduced by the subordinating conjunction *Nachdem* ('After that'), and the adverbial predicate is the past participle *gerissen* ('broken'), followed by the auxiliary *ist* ('is'), while the subject of the adverbial clause is the noun *Kette* ('chain'). The main predicate is the past participle *gebracht* ('brought'), which occupies the last position of the main clause before the coordinating conjunction *und*, since the auxiliary *habe* occupies the first position of the main clause, immediately followed by the subject in postverbal position, i.e. the pronoun *ich* (there is therefore an inversion between subject and auxiliary in the main clause).



- (49) Nachdem an meinem Fahrrad die Kette **gerissen ist**, habe ich es dahin gebracht [...].²⁰⁷
After that the chain of my bicycle **had broken**, I brought it there [...].

Table 55 shows the distribution of UPOS in the role of adjectival modifiers, i.e. as predicates in relative clauses. As for the LIT, the distribution of nonverbal predicates in this role resembles the distribution of nonverbal predicates observed for adverbial clauses, and it is therefore in accordance with the average value observed in Chart 9 for this dataset. Moreover, adjectives and nouns are almost equally distributed in this role. By contrast, for both GSD and HDT, the use of nonverbal predicates in this role dramatically decreases with respect to the average values, especially for GSD.

²⁰⁶ sent_id = lyceum-f72-s1.

²⁰⁷ sent_id = train-s306.

DEPREL	UPOS	RF LIT	RF GSD	RF HDT
acl	NOUN	7.8	2.3	1.4
	PRON	1.4	0.7	0.3
	ADJ	8.0	5.1	3.6
	VERB	77.0	91.1	91.4
	OTHERS	5.7	0.8	3.3

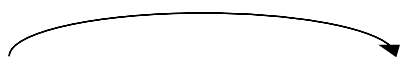
Table 55 Distribution of UPOS in the role of adjectival modifiers.

Example (50) illustrates the use of a nonverbal predicate as adjectival modifier in Fragments. In this case, the relative clause is introduced by the relative pronoun *was* ('which'), which is the nominal subject of the relative clause, and refers back to the pronoun *alles* in the main clause. The nonverbal predicate of the relative clause is the adjective *gut* ('good'), which, in turn, is coordinated to a second adjective, i.e. *groß* ('big'). The copula of the relative clause is *ist* ('is'), which occupies the last position of the relative clause. The main predicate is a nonverbal predicate, whose nonverbal element is the pronoun *alles*, modified by the relative clause.



(50) Paradox ist **alles**, **was** zugleich **gut** und **groß** ist.²⁰⁸
 Paradox is **everything**, which **is good** and **big** at the same time.

By contrast, example (51) exemplifies the use of a verbal predicate in a relative clause in GSD. In this case, the verbal predicate of the relative clause is the verb *anspricht* ('consider'), while the nominal subject of the relative clause is the proper noun *Perot*. The relative pronoun *die* plays the role of direct object of *anspricht*, and refers back to the plural noun *Themen* ('topics') in the main clause. The two nouns *Rezession* ('Recession') and *Bürokratie* ('Burocracy') occurring after the relative clause are both appositions of the noun *Themen*, and the main predicate is the verb *machen* ('to make').



(51) Zwei **Themen**, die Perot immer wieder **anspricht**, Rezession und Bürokratie, machen ihnen besonders zu schaffen.²⁰⁹
 Two topics, which Perot always considers, Recession and Burocracy, seems to them particularly worth menaging.

²⁰⁸ sent_id = lyceum-f48-s2.

²⁰⁹ sent_id = test-s621.

Table 56 shows the distribution of UPOS in the role of coordinate predicates. The distribution of nonverbal predicates involved in asyndetic coordination (*parataxis*) almost corresponds to the overall trend observed in Chart 9, with the nonverbal predicates being more frequent in LIT with respect to both GSD and HDT. An example of a nonverbal predicate coordinated to a main predicate through *parataxis* relation in the Fragments is reported in (52). The role of nonverbal predicate is filled by the noun *Selbstmord* ('suicide'), while the copula is *ist* in the second position of the coordinate clause. The predicate of the main clause is the verb *verschwindet* ('disappear').

(52) Das innere Leben verschwindet unter dieser Behandlung; sie **ist** der jämmerlichste **Selbstmord**.²¹⁰
 The inward nature disappears under the treatment; this is the pathetic suicide.


DEPREL	UPOS	RF LIT	RF GSD	RF HDT
parataxis	NOUN	12.5	5.0	4.6
	PRON	1.9	0.2	4.6
	ADJ	8.0	11.7	6.3
	VERB	77.6	83.0	84.4
	OTHERS	0.0	0.0	0.0
conj	NOUN	6.8	3.2	13.8
	PRON	1.2	0.1	1.0
	ADJ	5.8	6.6	25.9
	VERB	86.1	90.0	59.3
	OTHERS	0	0.0	0.0

Table 56 Distribution of UPOS in the role of coordinate predicates.


As for the syndetic coordination (*conj* relation), the frequency of coordinated nonverbal predicates decreases with respect to the asyndetic coordination, both in LIT and in GSD. Conversely, it dramatically increases in HDT, with about 40% of predicates in nonverbal form, and a significant part of them filled by adjectives. (53) exemplifies the use of an adjective as nonverbal predicate coordinate through *conj* relation in HDT. The role of the coordinated predicate in the clause introduced by the conjunction *und*

²¹⁰ sent_id = athenaeum-f336-s10.

(‘and) is filled by the adjective *ersichtlich* (‘visible’), which occupies the last position, while the copula occupies the position right after the conjunction. The main predicate is the verb *tritt* (‘move’).

- (53) Die URL **tritt** dabei in den Hintergrund **und ist** für den Benutzer nicht **ersichtlich**.²¹¹
The URL moves back in the background, and it **is** not **visible** for the user.
- 

The sentence in (53) exemplifies the use of a verbal predicate coordinated through syndetic coordination in Fragments. The coordinated predicate is the verb *finden* (‘to find’), which depends back on the main predicate *suchen* (to search). In this case, the two coordinated clauses share the same nominal subject, which is the pronoun *Wir* (‘We’) in the main clause.

- (54) Wir **suchen** überall das Unbedingte, und **finden** immer nur Dinge.²¹²
We **search** everywhere the absolut, un we **find** only things.
- 

4.5.2 Verbal Forms

Let us now focus on the word class of verbs, which embodies verbal predicates. In each dataset, there are five different subclasses of verbs²¹³ which are grouped under the coarse-grained UPOS VERB. They are encoded through the STTS in the field XPOS (see Chapter 2). They are summarized in Table 57. Chart 10 shows the distribution of these classes across the three datasets.

²¹¹ sent_id = hdt-s15310.

²¹² sent_id = bluethenstaub-f1-s1.

²¹³ Auxiliaries, copulas and modal verbs are all excluded from this class, since they are all tagged as AUX (UPOS) in the UD.

UPOS (UTS)	XPOS (STTS)	Meaning	Example
VERB	VVFIN	Finite form	gehet
	VVINFIN	Non-finite form	gehen
	VVIZU	Non-finite form with zu	einzugehen
	VVIMP	Imperative form	gehe!
	VVPP	Past-participle form	gegangen

Table 57 Correspondence between the UPOS VERB and its subclasses (XPOS).

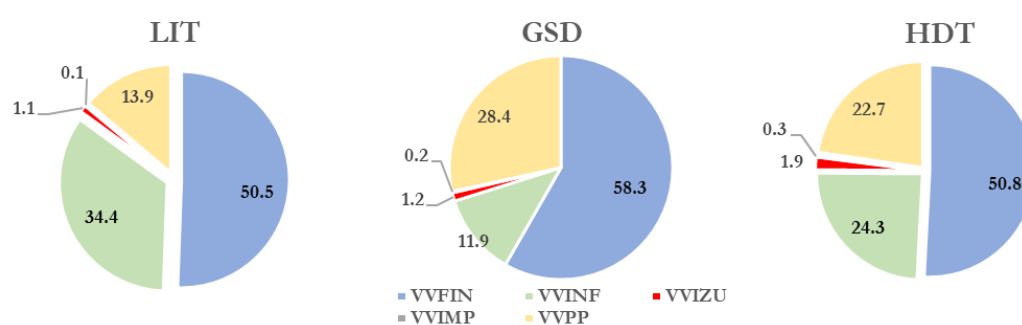


Chart 10 RF % Distribution of subclasses of verbs out of the total count of verbs (RF%).

As shown in Chart 10, more than 50% of verbs occur in finite form (VVFIN) in all the datasets. This is a rather expected result, since the finite forms are intuitively more frequent than non-finite forms and past-participle forms in texts.¹⁴To obtain more empirical evidence about this phenomenon, I conducted a quick corpus-based investigation on the German corpus *deWac* (Baroni et al. 2009), which is hosted on the online platform *Sketch Engine* (Kilgarriff et al. 2014) and collects texts from different sources on the web. I extracted the occurrences (AF) of the finite verbal forms, i.e. those tagged as VVFIN (parts of speech are encoded in *deWac* with STTS too) by using CQL, and I calculated RF of this class with respect to the total count of verbs in the corpus (excluding auxiliaries and modals). I here report the results: $TOTAL\ COUNT_{VERBS} = 122,636,598$, $AF_{VVFIN} = 64,954,433$, $RF_{VVFIN} = 53\%$. This result also confirms that the finite forms should be overall more frequent than the other forms. However, as shown in Chart 4, the RF of the finite forms (VVFIN) is clearly higher (almost + 8%) in GSD with respect to the other datasets. This must be correlated to the text typology. In fact, GSD mostly consists of texts from Wikipedia, which, notoriously, frequently report about past events²¹⁴. The verbal form which is usually

²¹⁴ I do not argue that they only report about past events, since they are frequently explanatory texts as well (therefore present oriented). However, being encyclopaedic texts, references to past actions must be more frequent.

used in German to narrate events taking place in the past, especially in those formal texts dealing with historical topics or biographical events, is the simple perfective form of the past tense, i.e. the *Präteritum*, which is usually preferred to the perfective compound form of the past, i.e. the *Perfekt*, for this purpose²¹⁵. An example of the use of the *Präteritum* as historical tense in GSD is shown in (55) and (56):

- (55) Für das Schwedische Fernsehen **arbeitete** er als Drehbuchautor, bevor er mit dem Schreiben von Horrorliteratur **begann**.²¹⁶
He **worked** as a screenwriter for Swedish television before he **began** writing horror literature.
- (56) Nach einer kurzen Stagnationsphase **begann** 1990 durch den verstärkten Neubau von Wohnungen eine zweite bis heute anhaltende Wachstumsphase.
After a brief period of stagnation, a new phase of growth **began** in 1990 due to the increased construction of new apartments.
- (57) In 1840 **verlegten** er und seine Frau Caroline Mathilde Bohlen (1800-1882) ihren Wohnort in das badische Mannheim.²¹⁷
1840 he and his wife Caroline Mathilde Bohlen (1800-1882) **moved** to Mannheim, in Baden.

On the contrary, the use of the *Präteritum* is expected to be less frequent in both LIT and GSD. As for Fragments, even if they are literary texts, they are quintessentially speculative texts, therefore the narration of events, both real and fictional, does not play a prominent role in the communicative intentions of these texts. In terms of verbal aspect, they will tend to prefer the present tense rather than the past, since most of the them are very judgment oriented, as already stated before. The present must be therefore very often used to convey the sense of judgment, as well as the universal scope of the message. An example of the typical use of the *Präsens* in Fragments is shown in (58), (59) and (60):

- (58) Man **nennt** viele Künstler, die eigentlich Kunstwerke der Natur sind.²¹⁸
We **call** many people artist, who are actually works of art.
- (59) Durch Humor **wird** das eigenthümlich Bedingte allgemein interessant, und **erhält** objektiven Werth.²¹⁹
Through humor **becomes** the conditioned in overall interesting, and receives objective value.
- (60) Schlegels Ironie **scheint** mir ächter Humor zu seyn.²²⁰
It **seems** to me that Schlegel's irony is true humor.

²¹⁵ Even if it can also be used with an imperfective aspect, *Präteritum* is mainly used as perfective historical tense especially in formal writing and literary writing. Cf. (Wermke, Kunkel-Razum, and Scholze-Stubenrecht 2005).

²¹⁶ sent_id = train-s4688.

²¹⁷ sent_id = train-s7154.

²¹⁸ sent_id = lyceum-f1-s1.

²¹⁹ sent_id = bluethenstaub-f29-s3.

²²⁰ sent_id = bluethenstaub-f29-s6.

Also when referring to authors from past historical ages, and to their thought, Fragments' authors tend to use the *Präsens* as well. An example is shown in (61):

(61) Im Plato **finden** sich alle reinen Arten der griechischen Prosa in klassischer Individualität unvermischt, und oft schneidend nebeneinander: die logische, die physische, die mimische, die panegyrische, und die mythische.²²¹

In Plato, we **find** all the authentic types of Greek prosa, mixed with classic individuality, and they clash with each other: the logical one, the physical one, the mimetic one, the panegyric one, and the mythical one.

The *Präteritum* is used very rarely in Fragments. (68) exemplifies the use of *Präteritum* to refer to some core values and processes of the classic world of the ancient Greece.

(62) Aus Dichtung und Gesetzgebung **bildete** sich die griechische Weisheit.²²²

From Poetry and legislation the Greek Wisdom build itself.

One may say that, even if the *Präteritum* is used very rarely in Fragments, a massive use of the present tense could however cause the RF of the finite forms to be very high in this dataset. In other words, the high RF of present-tense forms alone would not explain the difference in the distribution of finite forms with respect to GSD, since the forms of the *Präsens* also are finite forms. In fact, the different distribution of finite forms cannot be understood without considering the frequency of the non-finite forms (VVINF). As shown in Chart 10, LIT shows a remarkably higher RF of non-finite forms (in green) with respect to both the other datasets, especially to GSD (+ 22.5% with respect to GSD, + 10% with respect to HDT). Consequently, all those syntactic phenomena causing the distribution of non-finite forms to increase are expected to be frequent in Fragments, i.e. mainly modality and infinitival clauses. As previously shown in Chart 8, high RF of both *aux* relation (which governs both auxiliaries and modal verbs) and *xcomp* relation (which governs infinitival clauses) in this dataset seems to confirm this assumption. A rather frequent use of these structures at present tense can cause the frequency of finite forms to decrease out of the total number of verbs. Moreover, a high RF of nonverbal predicates could also contribute to mitigate the weight of the final forms in the verbal system of Fragments. In other words, when Fragments tend to use a finite form, they tend to use the *Präsens* rather than the *Präteritum*, while it should be exactly the opposite in GSD. But, at the same time, GSD's texts tend to speak most

²²¹ sent_id = athenaeum-f165-s1

²²² sent_id = athenaeum-f165-s1.

about the past, therefore the RF of *Präteritum*, and therefore of finite forms, is high in this dataset for this reason. By contrast, Fragments tend to speak most in the *Präsens*, but they frequently do this through modal verbs, infinitival clauses and nonverbal predicates. these structures should occur more rarely in GSD's texts. The important role of nonverbal predication in Fragments was already widely considered. An example of the other structures used in the *Präsens*, which can cause the frequency of non-finite verbs to increase in this genre is reported in (63) (modal verb), and (64) (non-finite clause).

(63) Man **muß** das Brett **bohren**, wo es am dicksten ist.²²³

One must drill a board, where it is thickest.

(64) Es bedürfte eines neuen »Laokoon«, **um** die Grenzen der Musik und der Philosophie **zu bestimmen**.²²⁴

We need a new "Laocoon" to fix the limits of music and Philosophie.

As for HDT, the use of the *Präteritum* should be expected at a certain degree, since news texts can also recall events from a remote past. However, in general, the distribution of this past form is expected to be lower with respect to the texts from Wikipedia. In fact, news texts mainly tend to refer to actions which are still in progress or that have recently happened, and they sometimes try to anticipate some future events on the base of the current information available, therefore they will presumably show an higher RF of both present and future forms rather than of *Präteritum* forms. When referring to past events, they usually refer to recently happened past events, therefore the *Perfekt* must be used much more frequently than the *Präteritum*. Example (65) shows the use of the *Präteritum* in HDT, while example (66) shows the use of the *Perfekt*.

(65) Umgekehrt **fühlten** sich 5,2 Prozent der Befragten im Erhebungszeitraum sogar **beflügelt** und kauften mehr Musik-CDs als früher.²²⁵

Conversely, 5.2 percent of those surveyed even **felt** inspired during the survey period and bought more music CDs than before.

(66) Ein Experiment **hat gezeigt**, dass der Pilz erst ab 30 Grad Celsius und einer relativen Luftfeuchtigkeit von 90 Prozent Appetit auf CDs bekommt.²²⁶

²²³ sent_id = lyceum-f10-s1.

²²⁴ sent_id = lyceum-f64-s1.

²²⁵ sent_id = hdt-s195190.

²²⁶ sent_id = hdt-s206769.

The past-participle forms are rather frequent in HSD and especially in GSD. As for HDT, the frequent use of the *Perfekt* tense when speaking about recent past events would also contribute to the rather high RF of the finite past participle forms (VVPP) in HDT (RF = 22.7), even if it should be mainly caused by a frequent use of the passive form (Cf. Chart 8). As for GSD, the high frequency of past participle forms should be mostly due to a large use of the passive voice, as previously demonstrated both in Table 41 and Chart 8. In general, we can say that this dataset shows a more balanced distribution of verbs between the two subclasses VVPP and VVFIN. On the contrary, the distribution of these classes is clearly unbalanced in the other two datasets. As for the phenomena causing the distribution of non-finite forms, they are investigated later.

The list of the 15 most frequent word forms tagged as VVFIN can help understand to what extent the use of the *Präteritum* plays a role in the RF of finite forms in the GSD dataset and in the other datasets as well. Table 58 reports the list of the most common verbal forms for each dataset. The forms of the *Präteritum* are highlighted in bold.

Rank	LIT		GSD		HDT	
	Form	RF	Form	RF	Form	RF
1	gibt	4.0	gibt	1.5	sagte	1.9
2	läßt	2.0	liegt	1.4	gibt	1.2
3	scheint	2.0	kam	1.4	geben	1.0
4	weiß	1.5	gab	1.0	stehen	0.9
5	wird	1.5	erhielt	1.0	sagen	0.9
6	ist	1.4	befindet	1.0	kommen	0.9
7	macht	1.3	begann	0.7	steht	0.9
8	geht	1.3	gehört	0.7	bietet	0.8
9	hat	1.3	führte	0.7	gehen	0.8
10	sagt	1.2	nahm	0.7	geht	0.8
11	bleibt	1.2	besteht	0.7	lassen	0.8
12	haben	1.0	spielte	0.7	kommt	0.7
13	sieht	1.0	studierte	0.6	sieht	0.7
14	entsteht	0.9	steht	0.6	stellt	0.7
15	gehört	0.9	arbeitete	0.6	gab	0.7

Table 58 Most frequent verbal finite forms.

As shown in Table 58, the hypothesis about a widespread use of the *Präteritum* in the GSD is correct: 9 forms out of 15 most frequent verbal forms in this dataset are *Präteritum* forms. On the contrary, the RF of forms of the *Präteritum* is very low in HDT (only 2 forms out of 15), even if the most frequent finite

verb form of this dataset is at the *Praeteritum*, which is *sagte* ('said'), i.e. the *Praeteritum* third-person singular of the verb *sagen* ('to tell'). This must be mostly caused by the typical use of this form to introduce both indirect speech and direct speech in news texts. An example of this use of the verb *sagen* in the HDT is shown in (67) (indirect speech) and in (68) (direct speech):

(67) Am Mittwoch **sagte** ein BSI-Sprecher, er rechne noch im Juni mit einer Einigung über die Prüfungsmodalitäten.²²⁷

On Wednesday, a BSI spokesman **said** he would expect an agreement on the exam modalities in June.

(68) "Windows und Outlook wurden nur deshalb als Angriffsziele gewählt, weil sie die populärsten Programme auf dem Markt sind", **sagte** Bernhard Grander von der deutschen Microsoft GmbH.²²⁸

As far as Fragments are concerned, none of the 15 most frequent forms in in this genre are at *Präteritum*. Overall, we can now state the GSD's texts belongs to a very past-oriented textual genre. On the contrary, Fragments are a very present-oriented textual genre. Moreover, many finite forms in HDT are third-person plurals: *geben* ('to give'), *stehen* ('to stay'), *sagen* ('to tell'), *kommen* ('to come'), *gehen* (to go), *lassen* ('to leave'). Conversely, all the fifteen most frequent forms of finite verbs in both LIT and GSD are third-person singulars.²²⁹ It seems therefore that news texts tent to prefer third-person plurals in the role of subjects of finite forms, while both Fragments and the texts from Wikipedia tent to privilege third-person plurals in this role.

2.1.1 Existential Clauses

I now focus on the verb *geben* (to give'), whose third-person singular (*gibt*) is the most frequent finite form in both LIT and GSD, and the second most frequent form in HDT. In German, *geben* is frequently used in the impersonal construction *es gibt*, which is used as predicate in the existential clauses, i.e. those clauses encoding the existence of someone or something, either an abstract or a real entity. In other words, the phrase *es gibt* has the same function of the constructions *there is* or *there are* in English, which states the existence of the entity that follows, such as in the sentence "there are many apples on the table". Before inquiring into the distribution of the existential constructions, I first report about an issue concerning

²²⁷ sent_id = hdt-s206114.

²²⁸ sent_id = hdt-s206040.

²²⁹ With the sole exception of *haben* ('to have') in LIT (rank 12), which is a third-person plural.

the annotation of the existential constructions in UD for German. In the English sentence reported above, according to the UD scheme, *apples* is the subject of the verb *are* (which is not used as a copula), while *there* has the role of expletive element. In the UD scheme, the expletive element occupies an argument position of the verb, i.e. it is grammatically necessary to build the verbal phrase, but it does not play any of the semantic roles that fulfils the predicate.²³⁰ In the case of there-BE constructions, *there* is necessary to build the construction with existential vale, but it does not have any semantic role with respect to the verb to be. Therefore, it depends on the verb *are/is* through *expl* relation, while the entity whose existence is stated by the there-BE construction depends on the verb *are/is* (i.e. *apples* in the example above) as nominal subject (*nsubj*). As far as the German existential construction *es gibt* is concerned, *es* should be the expletive element, since it satisfies the conditions to be considered as expletive element, while the entity should depend on the verb *gibt* as nominal subject. An example of the existential use of *geben* in the phrase *es gibt* in GSD is shown in (69), in which the personal pronoun *es* and the entity of which the existence is stated are highlighted in bold:

(69) In Brasilien **gibt es** große **Lagerstätten**, in denen besonders große Kristalle gefunden wurden.²³¹

There are large **deposits** in Brazil, in which they have been mostly found big crystals.

However, the syntactic status of both the pronoun *es* and the entity in the German existence clauses is problematic. In fact, the entity is always in accusative case, therefore, grammatically speaking, it should be treated as a direct object. At the same time, the pronoun *es* could be considered as a nominal subject. This issue has generated inconsistencies in the annotation of the UD German treebanks²³². As for LIT, I considered all the entities of the existential clauses as nominal subjects in the annotation phase, and the pronouns *es* as expletive elements (see Chapter 3). Therefore, the entity depends on the verb through *nsubj* relation, while the pronoun through *expl* relation. I tested the annotation in the GSD treebank²³³. I extracted all the tokens in accusative case²³⁴ that are governed by the form *gibt*, which, in turn, governs the pronoun *es*. I ran the following query:

Case=Acc < (gibt > es)

²³⁰ Cf. <https://universaldependencies.org/u/dep/expl.html>.

²³¹ sent_id = train-s14092.

²³² There is only a recent guideline mentioning the entities in the existential clauses, from which one can assume that they should be treated as objects. Cf. <https://universaldependencies.org/de/index.html>. However, this issue is not dealt with explicitly, and it should be fixed in future releases. In any case, from now on, I will refer to the entity of the *es-gibt* construction as object.

²³³ In this section, only data concerning LIT and HDT are provided. I was not able to retrieve any data from HDT due to technical problems in the extraction of data.

²³⁴ In this case, I could exploit the morphological features, since they are encoded in GSD.

I then counted how many of these tokens are governed through *nsubj* relation and how many through *obj* relation instead. The result is that 33% is governed through *obj* relation, while 55% through *nsubj* relation, and the remaining 12% by other relations. The annotation of the relation spanning from *gibt* to the entity of the existential clauses in GSD appears therefore affected by a rather high degree of arbitrariness. I performed the same test on the annotation of the pronoun *es* in the *es-gibt* constructions. I extracted all the occurrences of *es* that are governed by the verbal form *gibt*, which, in turn, governs another token in accusative case.

es < (gibt > Case=acc)

The result is that 64.2% of the occurrences if *es* depends on *gibt* through *expl* relation, while 35.7% depends on *gibt* through *nsubj* relation. Therefore, the annotation of the pronoun *es* in *es-gibt* in GSD turned out rather inconsistent, even if almost two thirds of the occurrences of *es* are annotated as expletive elements. As demonstrated, the annotation of the existential clauses in UD for German suffers from a rather high degree of arbitrariness.

Let us get back to the form *gibt*. It is worth remembering that *geben* can be also used in its “canonical” transitive form, i.e. in the construction ‘to give someone something’. In this case, besides the subject, the verb usually requires a direct object, and optionally an indirect object in dative case. An example of the transitive use *geben* from the Fragments is shown in (70), in which the nominal subject is the pronoun *das*, while the direct object is the noun *Vorgefühl* (‘feeling’).

- (70) Man kann etwas innig lieben, eben weil mans nicht hat: **das gibt** wenigstens ein Vorgefühl ohne Nachsatz.²³⁵
 One can deeply love something, just because one has nothing: this gives at least a presentiment without afterthought.

One may ask how many occurrences of *gibt* in each dataset are due to the existential use, and how many to the transitive form instead. I therefore extracted the occurrences of *gibt* governing the personal pronoun *es* from both LIT and GSD, and I divided them per the total occurrences of the form *gibt* in each dataset. As for LIT, I ran the following query, which returns all the occurrences of *gibt* that govern the pronoun *es* through *expl* relation:

gibt >expl es

²³⁵ sent_id = lyceum-f69-s2.

As for GSD, I ran the following query:

gibt > es >Case=Acc

Table 59 reports the results.

	LIT	GSD	HDT
RF es gibt	84.1.0	77.6	-

Table 59 Distribution of the existential construction *es gibt* out of the total occurrences of the form *gibt*.

As shown in Table 59, the use of *gibt* in the existential construction *es gibt* is very frequent both in LIT and in GSD, even if it occurs at a lower extent in the GSD with respect to LIT. German is a language with a relatively free word order. One may now ask how many times the pronoun *es* occurs in preverbal position, and how many in postverbal position instead. In terms of dependency relations, it means counting how many dependency relations spanning from “gibt” to “es” are left oriented²³⁶ (preverbal *es*), and how many are right oriented²³⁷ (postverbal *es*) instead. I therefore run the queries previously used to extract the occurrences of the *es-gibt* constructions, but I added the orientation of the dependency relation. As for LIT, I run the following queries, which extract all the occurrences of *gibt* respectively governing *es* from right to left (@L), i.e. *es* is in preverbal position, and from left to right (@R), i.e. *es* is in postverbal position.

gibt >expl@L es

gibt >expl@R es

As for GSD, I run the following queries:

gibt >@L es >Case=Acc

gibt >@R es >Case=Acc

Table 60 summarizes the results.


²³⁶ The head of the relation is on the right.


²³⁷ The head of the relation is on the left.

	LIT	GSD
Preverbal	81.1	38.1
Postverbal	18.9	61.9

Table 60 Distribution of preverbal and postverbal pronouns in the form *es gibt* in each dataset.

As shown in Table 60, the distribution of preverbal and postverbal pronouns in the *es-gibt* constructions varies a lot across the datasets. In Fragments, the pronoun *es* occurs in preverbal position in the majority of cases. On the contrary, the preverbal use of *es* significantly decreases in GSD, where the majority of pronouns occurs in postverbal position instead. An example of a preverbal use of *es* from the Fragments is shown in (71), while a postverbal occurrence of *es* in the GSD dataset is shown in (72)


 (71) **Es gibt** so viel Poesie, und doch ist nichts seltner als ein Poem!²³⁸
There is so much Poetry, **and** certainly nothing is **rarer** than a Poem!


 (72) In Berlin **gibt es** mittlerweile Strandbars wie Sand an dem Meer.²³⁹
 In Berlin **there are** now beach bars like sand on the sea.

As shown in (73), *es* occurs in postverbal position when the *Vorfeld*, i.e. the field before the finite verb (or the auxiliary), is occupied by a verbal modifier. For instance, it can be a prepositional phrase with locative function, which encodes spatial information, such as *in Berlin* in (72), or encoding time information, as shown in (73) from GSD, where the time information is encoded through the proposition *seit* (“from”) followed by the cardinal 2001:

²³⁸ sent_id = lyceum-f4-s1.

²³⁹ sent_id = dev-s352.

(73) **Seit 2001 gibt es** nun vier Staffeln à sechs Teams, wobei die beiden erstplatzierten Clubs das Viertelfinale erreichen.²⁴⁰

Since 2001 there have been four seasons with six teams each, with the two first-placed clubs reaching the quarter-finals.

In addition, the first position could be occupied by an adverb, such as *da* ('there'), or *hier* ('here'). An example of an adverb occupying the *Vorfeld* immediately followed by *es-gibt* construction is reported in (74) from GSD.

(74) **Da gibt es** sogar echte Tropische Pflanzen aus dem Süden²⁴¹.

There, there are even authentic tropical plants from the south.

Presumably, all these constructions could be very frequent in GSD, especially because of the biographical and historical nature of the texts from Wikipedia, which notoriously tend to encode both spatial information and time information. Nevertheless, the *Vorfeld* could also be directly occupied by the object of the *es-gibt* construction, i.e. by the token governed by *gibt* through *nsubj* or *obj* relation. An example of this inversion of the object with respect to the pronouns *es* from the GSD is shown in (30), in which the noun **Firma** ('company') has the role of object of *gibt*:

(75) Diese **Firma gibt es** seit 1989 nicht mehr.²⁴²

This **company there is** not anymore since 1989.

One may now ask whether the *Vorfeld* of the existential clauses in the GSD dataset²⁴³ is mostly occupied by objects or by modifiers, and also what type of modifier tend to occur most in that position, i.e. whether adverbs or phrases. I therefore run three queries. First, I extracted all those tokens governed by *gibt* from

²⁴⁰ sent_id = train-s8298.

²⁴¹ sent_id = dev-s499.

²⁴² sent_id = test-s113.

²⁴³ I focused on this dataset only, since the phenomenon of the postverbal pronoun mostly occurs here.

right to left through *nsubj/obj* relation²⁴⁴, with *gibt*, in turn, governing *es* from left to right. I ran the following query:

_ <nsubj@R|<obj@R (gibt >@R es)

Second, to extract adverbs in preverbal position, if any, I varied the first part of the query only, i.e. I changed the conditions concerning the right-to-left dependent of *gibt*, and I extracted all those tokens governed by *gibt* from the right through *advmod* relation. I ran the following query:

_ <advmod@R (gibt >@R es)

Third, to extract the oblique arguments, I varied the first part of the query again, i.e. I extracted all those tokens governed by *gibt* on the right through *obl* relation. I ran the following query:

_ <obl@R ("gibt" >@R "es"|"Es")

Table 24 summarizes the results²⁴⁵.

DEPREL	RF GSD
nsubj/obj	21.3
obl	48.0
advmod	29.3
others	1.3

Table 61 Distribution of syntactic functions of preverbal tokens in *es-gibt* constructions when *es* is in postverbal position.

As reported in Table 24, most of the tokens preceding *gibt* in GSD are oblique arguments of *gibt*, i.e. they depend on *gibt* through *obl* relation. Therefore, the necessity to encode specification through prepositional, in particular time and space information, appears to play a decisive role in shaping the subcategorization frame of *geben* in this dataset, when used in *es-gibt* construction. The second most frequent element occupying the *Vorfeld* when *es gibt* is used are adverbs, while the object of the *es-gibt* construction tends to occur less frequently in preverbal position with respect to the verbal modifiers.

Another syntactic property of the form *es gibt* that one may investigate thanks to the dependency relations is the capacity of this finite form to generate complex sentences. In other words, one may ask whether the existential clauses are more used as simple sentences in one dataset with respect to another, or, by

²⁴⁴ I considered both these relations due to the annotation problem highlighted above.

²⁴⁵ RF refers to the absolute frequency of each *deprel* with respect to the total number of the preverbal dependents of *gibt* occurring in the form *es gibt* in the dataset, when *es* is postponed.

contrast, if they are more likely to generate complex sentences, i.e. coordinate or subordinate structures. An example of an existential clause consisting in a simple clause is shown in (76) from the GSD treebank.

- (76) Diese Firma **gibt es** seit 1989 nicht mehr.²⁴⁶
This **company there is** not anymore since 1989

As far as the coordination is concerned, the secondary clause can depend on the main clause either through the relation *conj* (syndetic coordination) or through the *parataxis* relation (asyndetic coordination). An example of an existential clause generating coordination through the *conj* relation is reported in (77) from LIT, while an example of an existential clause generating asyndetic coordination is reported in (78) from GSD:

- (77) **Es gibt** so viel Poesie, und doch ist nichts seltner als ein Poem!²⁴⁷
There is so much Poetry, **and** certainly nothing is **rarer** than a Poem!

- (78) Am Ende **gibt es** eine Dreifachhochzeit: Franz und Marion - Madeleine, Eva und Peter sowie Gustav und Mariele heiraten.²⁴⁸
In the end **there is** a triple wedding: Franz and Marion - Madeleine, Eva and Peter as well as Gustav and Mariele **get married**.

As far as subordination is concerned, the subordinate predicate can depend on the main existential clause through three distinct types of dependency relations: *advcl* (adverbial clause); *xcomp*, (non-finite clause), and *acl* (relative clause). In this last case, subordination does not directly work on the main predicate, but on a nominal of the main clause. I here deliberately omitted both the *csubj* and the *ccomp* relation, which respectively encode clausal subjects and clausal objects.²⁴⁹ An example for each of the dependency relations modifying an existential clause is respectively in (79) (*advcl*), (80) (*xcomp*) and (81) (*acl*), all from GSD. In each example, I highlighted both the main predicate and the subordinate predicate in bold. As for (37), both the noun i.e. *Schafe*, which is the element of the main clause that is modified by the


²⁴⁶ sent_id = test-s113.


²⁴⁷ sent_id = lyceum-f4-s1.


²⁴⁸ sent_id = train-s4779.

²⁴⁹ I preliminary searched for occurrences of both these types of clause modifiers depending on the form *es gibt* in both LIT and GSD, and no occurrence of these phenomenon was returned.

subordinate predicate of the relative clause (*asunutzen*), and the relative pronoun *die*, which is the subject of the relative clause, are highlighted too.

- (79) Doch **es gibt** Momente, da **scheint** der Wurm drin zu sein.²⁵⁰
 But **there are** moments when the worm **seems** to be inside.
- 

- (80) Es **gibt** von meiner Seite nur zu **bemängeln**, dass ich nicht meiner Mobilität gefragt wurde und auf die Strassenbahn angewiesen war.²⁵¹
There is nothing **to complain** about on my part, that I was not asked about my mobility and that I was dependent on the tram.
- 

- (81) Es **gibt** in dieser Branche zu viele schwarze **Schafe die** das Handykap der Kunden voll **ausnutzen**.²⁵²
There are too many black **sheep** in this industry **that take** full advantage of the invalidity of customers.
- 

To test this syntactic property of the existential clauses, I run different queries on the datasets. First, I searched for all those adverbial and non-finite clausal modifiers depending on *gibt* in *es gibt* constructions, therefore I extracted all those tokens depending on *gibt* through *advcl* or *xcomp* relation, with *gibt*, in turn, governing the pronoun *es* (I maintained this second part of the query in all the other following queries as well).

`_ <advcl | <xcomp (gibt >es)`

Second, I varied the first part of the query to search for those sentences depending on the main clause as relative clauses, therefore I extracted all those tokens depending on an element through *acl* relation, which, in turn, depends on *gibt* through *nsubj*, *obj* or *obl* relation. I ran the following query:

`_ <acl (_ <nsubj | <obj | <obl (gibt > es))`

²⁵⁰ sent_id = test-s118.

²⁵¹ sent_id = test-s142.

²⁵² sent_id = dev-s327.

Third, I searched for all those cases of existential clauses generating coordination, therefore I extracted all those tokens depending on *es-gibt* construction through *conj* or *parataxis* relation. I ran the following query:

_ <conj|<parataxis (gibt > es)

In the end, I searched for those existential clauses occurring as simple sentences. In this case, the form *gibt* must be root node, therefore it is the sole non-headed node of the sentence. Moreover, the form *gibt* must not govern any token through any of the before-searched dependency relations, and there must be no relative clause modifying the main clause neither. I therefore ran the following query:

gibt !< _ > es !(>advcl|>xcomp|>conj|>parataxis) _ !> (_ >acl _)

Chart 11 summarizes the results by grouping results concerning *advcl*, *xcomp* and *acl* in the category of *subordination*, and those of both *parataxis* and *conj* relations in *coordination*. All the results are expressed in terms of RF with respect to the total occurrences of *es gibt* in each dataset.

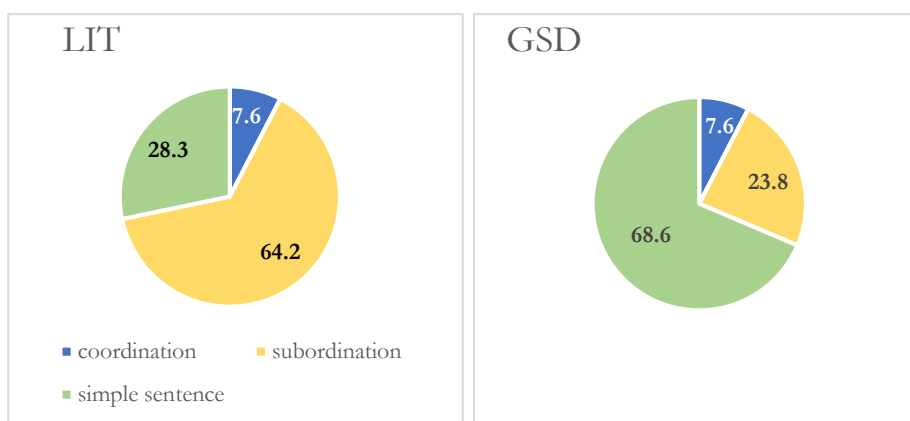


Chart 11 Distribution of coordination, subordination and simple sentences generated by the existential clauses, out of the total occurrences of the *es gibt* construction (RF%).


As shown in Chart 11, the distribution of syntactic constructions generated by existential clauses varies a lot between LIT and GSD. In Fragments, the majority of them generate subordinate structures. On the contrary, in GSD, the majority of them is used as simple sentence, and generates subordination much more rarely. In both cases, existential clauses tend to generate coordinate clauses very rarely. Table 62 reports the distribution of the dependency relations in each of the category shown in Chart 5, in terms of RF with respect to the total occurrences of *es gibt* in each dataset.

DEPREL	LIT	GSD
advcl	5.7	8.5
xcomp	1.9	0
acl	56.6	15.3
conj	5.7	5.1
parataxis	1.9	2.5
root	28.3	68.6

Table 62 Distribution of syntactic functions of those subordinate predicates (apart from root) generated by *es-gibt* constructions (RF%).

As shown in Table 62, the high degree of subordination generated by *es gibt* in Fragments is mainly caused by relative clauses, which are therefore the most frequent subordinate clause associated with the use of existential clauses in this dataset. In addition, I calculated that only 3.8% of the *acl* relations that depends of an item in a higher existential clause depends on an oblique argument, therefore the vast majority of them actually modify the token that plays the role of object of the *es-gibt* construction in the main clause. An example of a relative clause modifying the object of *es gibt* from the Fragments is reported in (82), in which the noun *Schriftsteller* is the object of *es gibt* in the main clause, while the verb *trinken* is the subordinate predicate modifying *Schrtsteller*, and *die* is the relative pronoun depending on *trinken* as nominal subject (and referring back to *Schrtsteller*).

(82) **Es gibt Schriftsteller die** Unbedingtes **trinken** wie Wasser;²⁵³
There are writers who drink the Absolute as it was water.



An example of a relative clause modifying an oblique argument of *es gibt* is reported in (83). In this case, there is an existential clause modified by two relative clauses at the same time, one preceding and one

²⁵³ sent_id = lyceum-f54-s1.

following the main clause. The first one, which consists in the pronoun *die* referring to *Menschen* and the verb *fortgehen*, modifies the noun *Menschen*, which, in turn, depends on *gibt* as oblique argument indeed. In addition, there is a relative clause introduced by the pronoun *welche*, whose predicate is the verb *stehen*, which, in turn, modifies the pronoun *manche* in the main clause.

(83) Unter den **Menschen, die** mit der Zeit **fortgehn, gibt es manche, welche**, wie die fortlaufenden Kommentare, bei den schwierigen Stellen nicht still **stehn** wollen.²⁵⁴

Among the men that go away, there are those, like the continuous opinions, who do not want to stand at difficult positions.

As for the other subordinate clauses, both adverbial clauses and non-finite clauses tend to occur very rarely when *es gibt* is used, especially the second ones. As for coordination, even if rare, the syndetic coordination occurs slightly more frequently with respect to the asyndetic coordination. As for GSD, the distribution of the *conj*, *parataxis* and *xcomp* is rather similar with respect to LIT. Among the subordinate structures, relative clauses (*acl* relation) are the most frequent form of subordination as well. However, almost 70% of the form of *es gibt* in this dataset occur in simple sentences. In conclusion, we can say that the syntax of existential clauses in Fragments is clearly marked by the necessity to encode specification about those entities whose existence is state. The favourite syntactic form to encode such specification are relative clauses. Conversely, the texts of GSD tend to encode the specification of the entities in existential clauses much more rarely.

4.5.3 Modal Verbs

In UD, the UPOS AUX is used to tag both auxiliaries and modal verbs. For the matter of clarity, Table 63 reports the distribution of this POS-tag in each data displayed earlier in Chart 7. It is expressed as RF with respect to the total number of UPOS in each dataset.

UPOS	RF LIT	RF GSD	RF HDT
AUX	7.1	3.9	5.0

Table 63 Distribution of the UPOS AUX.

²⁵⁴ sent_id = athenaeum-f332-s1.

Table 63 shows that there is a considerable gap between the distribution of AUX across the datasets, especially between LIT, which shows the highest RF of this UPOS, and GSD, which, on the contrary, shows the lowest RF of this UPOS. However, it is not clear to what extent these values are caused by a high distribution of “pure” auxiliaries, or of modal verbs, or by an equal distribution of both. The analysis of the XPOS is therefore necessary to disambiguate. Chart 12 reports the composition of the UPOS AUX in each dataset, distinguishing between the two macro subclasses of auxiliaries and modal verbs.

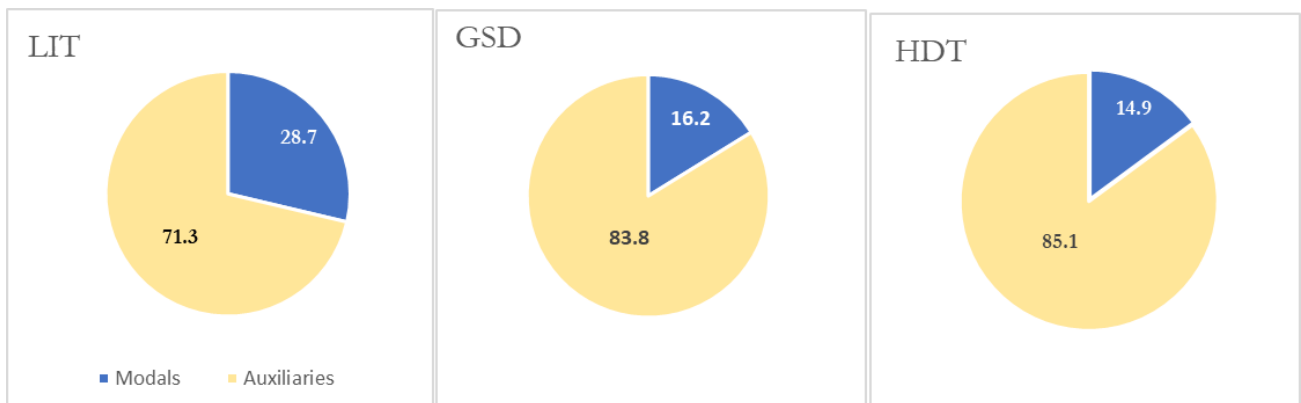


Chart 12 Composition of the UPOS AUX (RF %).

As shown in Chart 12, most of the tokens tagged as AUX are auxiliaries rather than modal verbs, in all the datasets. Nevertheless, the distribution of the modal verbs varies significantly across them: LIT shows a much higher RF of modal verbs with respect to both the other datasets (+ 12.5% with respect to GSD, and + 13.8% with respect to GSD). Chart 13 illustrates the composition of the UPOS AUX in detail. It displays the distribution subclasses of both auxiliaries and modal verbs in each dataset, which are encoded through XPOS (STTS). VA stands for auxiliary verb, while VM stands for modal verb; INF stands for non-finite form, FIN for finite form, while PP for past-participle form (for auxiliaries only).

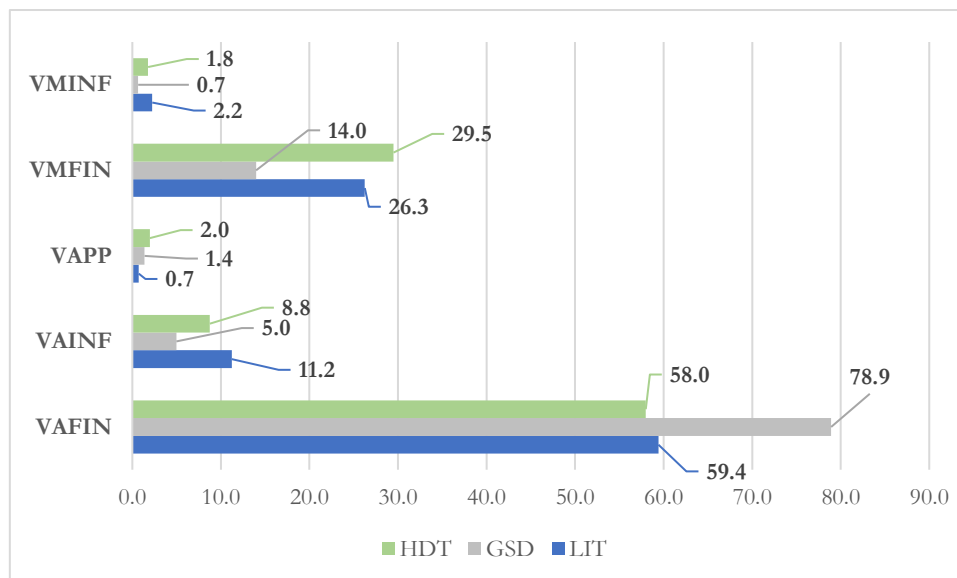


Chart 13 Distribution of subclasses of both auxiliaries and modal verbs out of the total number of tokens tagged as AUX. (RF%).

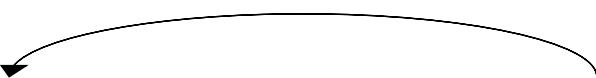
As shown in Chart 13, the finite forms are much more frequent than the non-finite forms for both auxiliaries and modal verbs in each dataset. Let us now focus on the modal verbs. Overall, we can say that the use of modality marks the syntax of the genre of Fragments more than the other textual genres collected both in GSD and HDT. I will now try to analyse this issue more in depth.

Semantically and pragmatically, modal verbs are those verbs conveying the attitude of the speaker toward the *Aussagengehalt* ('the content of the utterance'), i.e. toward the meaning of the action embodied by the predicate. For instance, in English, the sentence "I open the door" is rather different in meaning from the sentence "I can open the door", which, in turn, is rather different from "I must open the door". In each of these sentences, the agent is always "I", while the patient is always "the door", and the action is always the same one, i.e. the action of opening encoded by the verb *to open*. However, the use of a different modal verb cast a different meaning on the action of opening in each utterance, even if the participants to the act of predication are always the same ones. In the first case, the utterance "I open the door" describes the action of opening the door in an objective way. In the second case, through the use of the modal verb *to can*, the agent conveys a totally different message with respect to the previous sentence: he/she does not actually describe the action of opening that he/she performs, but he/she tells us that is able to do perform that action. In the last case, using the modal verb *must*, the speaker conveys a sort of necessity to perform the action, but, in this case too, he/she does not tell nothing about the factual performance of that action. Consequently, in terms of analysis of the predication, a high distribution of modal verbs in a text means that there is high intervention of the author on the way actions are evoked

through the predicate. I will show some real cases of the use of modal verbs in a while, but I want to focus on some syntactic aspects of the modal verbs first.

In German, the use of modal verbs also has a deep impact on the syntax of the verbal phrase and of the whole sentence in general. In fact, in the main clauses, modal verbs usually fill the second position (or *Linke Klammer*), which is usually occupied by the finite verb (or auxiliary), i.e. the second position of the clause after the *Vorfeld*, while the non-finite verb to which they refer occupies the final position of the clause (*Rechte Klammer*). This generates a discontinuous verbal phrase, in which other lexical items can occur between the first part of the verbal phrase (finite verb/auxiliary) and the last part of the verbal phrase. This part of the clause between the two elements of the verbal phrase is usually referred to as *Mittelfeld*. For an overview on topological structure of the German sentence, see e. g. (Pittner and Berman 2015). In terms of dependency relations, a high distribution of modal verbs in a dataset can cause widespread long-standing dependency relations. In fact, in UD, the modal verb is a direct dependent of the verb to which it refers, therefore one may expect a certain number of tokens occurring between the governor and the dependent. An example of the use of a modal verb in a German main clause from the GSD dataset is shown in (84). In this case, the modal verb *kann* occupies the second position after the nominal subject *Frau*, and both depend on the main predicate, which is the non-finite verb *werden* (‘become’), in the final position. The modal verb and its head are highlighted in bold:

(84) Die Frau **kann** auch durch Adoption eines Kindes zur Mutter **werden**.²⁵⁵
The woman **can become** a mother also through the adoption of a child.



In the subordinate clauses, modal verbs usually occupy the final position of the clause, immediately following the non-finite form of the verb to which they refer. An example of the use of a modal verb in a subordinate clause is reported in (85) from the GSD treebank. In this case, the non-finite verb *schlafen* (‘sleep’) is the predicate of an adverbial clause introduced by the subordinate conjunction *damit* (‘in this way’) and whose subject is the noun *Kinder* (‘guys’); *können* is the modal verb depending on *schlafen*, therefore it occupies the final position of the adverbial clause, immediately after the verb.

²⁵⁵ sent_id = train-s5982.

(85) Vielmehr fordern wir : Bleibt dem Rauschgold fern , damit die Kinder in Ruhe **schlafen können**.²⁵⁶

We increasingly demand: stay away from Rauschgold [the name of a pub], so that the children **can sleep** in peace.

When modal verbs are used in infinitival clauses, they always occupy the final position of the clause, and they are always preceded by the infinitival marker *zu*. An example from the Fragments is reported in (86). In this case, there is an infinitival clause preceding the main clause, introduced by the subordinating conjunction *um*, in which the non-finite verb is *schreiben* ('write'), while *können* is the modal verb depending on *schreiben*.

(86) Um über einen Gegenstand gut **schreiben zu können**, muß man sich nicht mehr für ihn interessieren;²⁵⁷

To **be able to write** good about an object, one must no longer interest in it.

In German, there are six modal verbs, i.e. *wollen* ('to will'), *sollen* ('to should'), *müssen* ('to must'), *mögen* ('like' / 'would like', especially in the form *möchten*), *können* ('to be able' / 'to can'), *dürfen* ('may')²⁵⁸. Table 64 reports the Average Dependency Length (ADL) of those dependency relations (*aux* relations) involving modal verbs, expressed as average number of tokens occurring between the head and the dependent. It was automatically calculated for each dataset by exploiting the ID of the head (ID_{head}) and the dependent (ID_{modal}) of each target relation. First, I searched for all the lines in the CoNLL-U file with a lemma of a modal verb. Second, I applied the following formula, where N is the number of modal verbs (AF) retrieved in each dataset.

$$\sum_N^1 | ID_{\text{head}} - ID_{\text{modal}} | / N$$

As expected, the ADL confirms that there are on average several tokens occurring between a modal verb and its head. However, the average number of tokens varies across the datasets. Fragments have the lowest ADL for modal verbs, while the news-texts of HDT are the genre showing the highest ADL.

²⁵⁶ sent_id = dev-s226.

²⁵⁷ sent_id = lyceum-f37-s1.

²⁵⁸ The translations provided for each modal verb are approximate.

	LIT	GSD	HDT
ADL Modals	3.8	5.2	6

Table 64 Average Dependency Length (ADL) of the dependency relations governing modal verbs.

The gap in ADL could be caused by a different distribution of the modal verbs between main clauses and subordinate clauses across the datasets. Intuitively, the higher the usage of modal verbs in the subordinate clauses, the lower the ADL of modal verbs in the dataset, since only the particle *zu* should be expected occurring between the head and the dependent. On the contrary, the higher their usage in the main clauses, potentially the higher their ADL, since more tokens should be expected occurring between the modal and the verb in a main clause, as previously shown in (84). We already observed a higher general distribution of subordination in Fragments with respect to the other datasets (see 4.4), therefore there could be a correlation between these two distributions. To check the distribution of the modal verbs in the subordinate clauses, I extracted those occurrences of lemmas of modal verbs that depend on any token, which, in turn, depend on another token through one of the relations encoding subordination. I considered all the following relations: *advcl* (adverbial clauses), *csubj* (clausal subjects), *ccomp* (clausal complement), *xcomp* (in this case, encoding infinitival clauses), and *acl* (relative clauses). I ran the following query:²⁵⁹

```
(L=können|L=müssen|L=sollen|L=dürfen|L=mögen|L=wollen) < (
<advcl|<acl|<csubj|<ccomp|<xcomp _).
```

Those occurrences of modal verbs not retrieved by the query were considered as modal verbs in main clauses. I then divided both the retrieved and non-retrieved occurrences by the total occurrences of modal verbs in each dataset. Table 27 summarizes the results expressed as RF.

	RF LIT	RF GSD	RF HDT
Subordinate clauses	37.3	29.1	12.2
Main clauses	62.7	70.9	87.8

Table 65 Distribution of modal verbs in main clauses and subordinate clauses.

As shown in Table 65, the majority of the modal verbs occur in main clauses in all the datasets. However, the distribution of modal verbs occurring in subordinate clauses remarkably varies across them. As

²⁵⁹ I had to specify the single lemma of each modal verb, because SETS does not allow for querying the XPOS (VMFIN/VMINF).

expected, Fragments shows the highest RF of modal verbs governed by subordinate predicates, while the use of modal verbs in subordinate clauses progressively decreases in both GSD and HDT. This distribution follows the trend observed in the increase of ADL. In fact, the hypothesis about a correlation between the distribution of modal verbs in subordinate clauses and their ADL was correct: HDT, i.e. the dataset showing the highest ADL for modal verbs, shows the lower distribution of modal verbs in subordinate clauses; while LIT, i.e. the dataset showing the lower ADL of modal verbs, shows the highest distribution of modal verbs in subordinate clauses.

Let us now consider the lemmas of modal verbs. Chart 14 displays the distribution of the lemmas of these modal verbs in each dataset out of the total number of tokens tagged as VMFIN or VMINF (RF%)²⁶⁰.

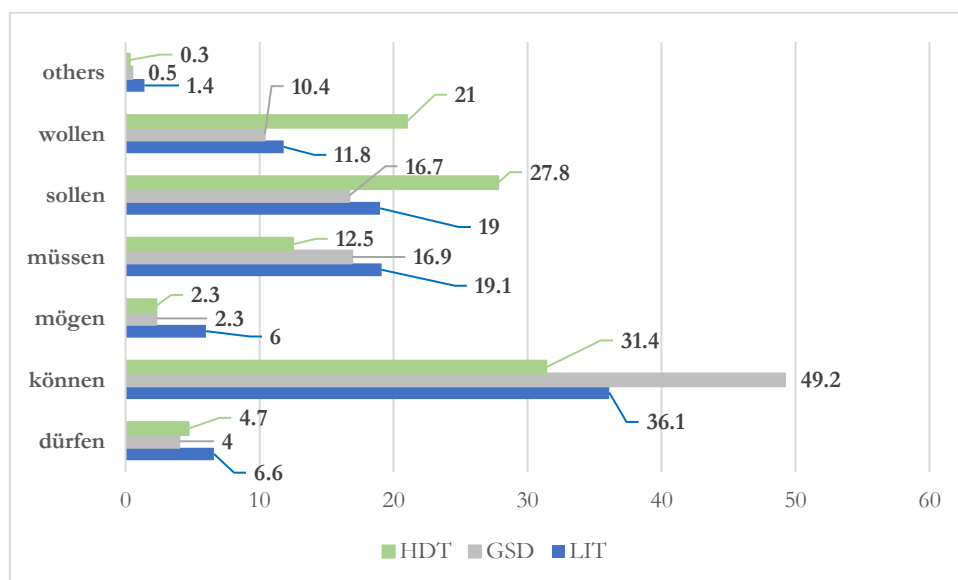


Chart 14 Distribution of lemmas of modal verbs (RF%).

As shown in Chart 14, the most common modal verb in each dataset is *können*, which, however, shows a remarkably higher RF in GSD with respect to the other two datasets. Fragments shows a higher RF of *müssen* with respect to the other datasets, especially with respect to HDT, since it is the second most frequent modal verb genre (even if it shows almost the same RF of *sollen*). *sollen* is the second most frequent modal verb in HDT, and its RF shows a remarkable gap with respect to the other two datasets, as in the case of *wollen*. Table 66 reports the most common finite forms of modal verbs in each dataset.

²⁶⁰ The label *others* in the chart includes those tokens mistakenly tagged as either VMFIN or VMINF.

RANK	FORM LIT	RF	FORM GSD	RF	FORM HDT	RF
1	kann	20.0	kann	18.7	soll	15.3
2	muß	10.8	konnte	12.0	will	12.4
3	können	8.4	können	11.8	können	10.3
4	soll	7.3	sollte	5.7	kann	8.5
5	wollen	5.0	muss	4.8	sollen	8.1
6	sollte	4.8	soll	4.7	wollen	4.2
7	will	4.8	musste	3.8	müssen	3.5
8	könnte	4.6	müssen	3.3	könnte	3.3
9	müssen	3.2	sollen	3.3	muss	3.3
10	müßte	3.2	wollte	3.2	könne	2.8
11	mag	2.9	will	3.0	konnte	2.7
12	darf	2.8	konnten	2.9	könnten	2.6
13	dürfte	2.6	mussten	2.3	wolle	2.0
14	Sollte	1.8	sollten	2.2	wollte	1.8
15	möchte	1.5	wollen	2.2	sollte	1.7

Table 66 Most frequent forms of modal verbs (RF %).

In both GSD and LIT, the most frequent form is *kann*, i.e. the third-person singular of *können*, while *soll*, i.e. the third-person singular of *sollen*, is the most frequent form in HDT. Actually, both *kann* and *soll* could be either the first-person singular or the third-person singular, because, in German, the first-singular person and the third-singular person of modal verbs have the same form, for the *Präsens* (present tense) as well as for the *Präteritum*. A disambiguation was therefore necessary. As for both GSD, I disambiguated the two forms thanks to the morphological features, i.e. I searched how many of the forms *kann* bear the tag “Person=3” in the field *features* of the CONLL-U file. The result is that 87% of the occurrences of the form *kann* in GSD are third-singular persons, and 99% of the occurrences of *soll* are third-person singular. As for LIT, the morphological features are not encoded in the treebank yet. I therefore searched for the subjects of all those verbs that govern the form *kann*, and I then searched occurrences of the personal pronoun *Ich* (‘I’) among them. I ran the following query:

```
_ <nsubj (VERB|NOUN|PRON|ADJ) >aux kann)
```

The result was that 98% of the occurrences of *kann* are third-singular persons. The second most frequent form in LIT is *muß*, i.e. the third-person singular of *müssen*. I ran the following query:

_ <nsubj (VERB | NOUN | PRON | ADJ)aux muß)

The result is that 100% of the occurrences of *muß* are third-person singulars, while the second most frequent form in GSD is *konnte*, i.e. third-person singular of *können* at the *Praeteritum* (97% of the occurrences of *konnte* are third-person singulars). As for the HDT, the second most frequent one is *will*, i.e. the third-singular person of *wollen* (99% of the occurrences of *will* are third-person singulars). Even if neither *sollen* nor *wollen* is the most frequent modal verb in this dataset, however, they are both much more frequent in HDT with respect to the other two datasets. Apart from the peak of *koennen* in GSD, the overall distribution of modal verbs seems therefore more similar between the Fragments and GSD's texts with respect to the news texts collected in HDT. However, the distribution of the finite forms of modal verbs confirms two observations made when analysing the distribution of the finite forms of verbs: first, the third-person singular is the most frequent person used for verbs in all the datasets; second, the *Präteritum* plays an important role as favourite verbal tense in the texts collected in GSD (7 forms of modals out of 15 are *Präteritum* forms), while the texts collected in both LIT and GSD are much more present-oriented (all the 15 most frequent forms of modal verbs are present forms).

Given the very frequent use of the third-person singulars in all the datasets, one may ask how many modal verbs are used in the impersonal form, i.e. with the impersonal pronoun *man* as nominal subject. An example of a modal verb used in impersonal form from the Fragments is reported in (87)

(87) **Man muß** das Brett **bohren**, wo es am dicksten ist.²⁶¹
One has to drill the board where it is thickest

In UD, the nominal subject of a modal modal verb depends on the non-finite verb to which the modal verb refers (see Chapter 3). Therefore, I extracted all those occurrences of modal verbs depending on verbs that govern the form *man* as nominal subject. I ran the following query:

```
L=können | L=müssen | L=sollen | L=dürfen | L=mögen | L=wollen <aux (VERB | NOUN | PRON | ADJ )nsubj man)
```

Table 67 summarizes the results.

²⁶¹ sent_id = lyceum-f10-s1.

	RF LIT	RF GSD	RF HDT
Impersonal	12.8	5.6	8.5
Personal	87.2	94.4	91.5

Table 67 Distribution of modals used in the impersonal form, out of the total number of modal verbs.

As shown in Table 67, most of the modal verbs are used in personal form in all the genres. However, the distribution of those modals used in impersonal form varies across the datasets. Fragments are the genre showing the highest use of the impersonal form. After all, the use of the impersonal form matches the communicative purpose of this genre. In fact, the use of the impersonal form surely helps give the utterance giving the utterances a strong sense of general scope. Besides the example previously showed in (87), I report other cases of the use of the impersonal form of modal verbs from this genre in (88), and (89). The pronoun *man*, as well as the modal verb and its head are highlighted in bold:

(88) Drittens: **man muß** die Selbstbeschränkung nicht **übertreiben**.²⁶²

Thirdly, **one must** not **exaggerate** with the self-limitation.

(89) Genie **kann man** eigentlich nie haben, nur sein.²⁶³

One **can** never **have** a genius, only be [a genius].

As for the texts in GSD, the very low frequency of impersonal forms of modal verbs was rather expected. In fact, this form is mostly used in the web reviews, to express the personal judgment by speakers about the object of the review. An example of this use from a web review from GSD is shown in (90) and (91).

(90) Im italienischen Restaurant Pinnocchio **kann man** nicht nur wunderbar Pizza **essen**, auch alle anderen Gerichte schmecken hervorragend.²⁶⁴

In the Italian restaurant Pinnocchio **one can** not only **eat** wonderful pizza, all other dishes also taste excellent.

²⁶²sent_id = lyceum-f37-s15

²⁶³ sent_id = athenaeum-f119-s4

²⁶⁴ sent_id = train-s95.

- (91) **Finden kann man** wirklich alles was man sucht.²⁶⁵
One can find indeed everything that is looking for.

However, this text typology represents only about 10% of the texts collected in GSD. Most texts from this dataset come from Wikipedia, which is notoriously a genre in which actions usually refer to specific and well-denoted subjects. Consequently, the frequency of impersonal forms should be overall very low in these texts.

As for HDT, web news about the ICT world can make use of the impersonal form with modal verbs for different kinds of communicative purposes. Some of them are shown in (92), and in (93).

- (92) An Hardware-Voraussetzungen **sollte man** dabei mindestens einen Celeron [...] **mitbringen**.²⁶⁶

In terms of hardware requirements, **one should have at disposal** at least one Celeron [...].

- (93) Durch die neuartige Technik **will man** empfindliche Bauteile ("Hot Spots") [...] **schützen können**.²⁶⁷

Through the new technology **they want to protect** sensitive components ("hot spots") [...].

In (92), the impersonal form is used to report about the minimum requirements that a user should have to use a particular software. In this case, the impersonal form is used to add a sort of objective utility to the scope of the message, since it is a sort of advice, based on a technical evaluation. In (93), the text report about the advantages brought by a new technology with respect to the pre-existing one. Through the use of the verb *wollen*, the author aims to report the goal behind the development of that technology, therefore, in this case, the use of the impersonal form with the modal verb aims to stress the benefits caused by the action of protection by somehow referring to those who made this action possible, and who aimed to let this action possible by designing that product.

In the end, I here also report some cases of modal verbs used in personal form in Fragments, which was the genre with the highest overall use of modality, in order to show other concrete uses of modality in

²⁶⁵ sent_id = train-s46.

²⁶⁶ sent_id = hdt-s67953.

²⁶⁷ sent_id = hdt-s67650.

this genre. I focused on the most frequent modal verb, i.e. *können*. Some examples are reported in (94), (95), and (96).

To retrieve them, I ran the following query, in which I negated the relation between the head of the modal verb and the impersonal pronoun *man*:

L=können <aux (VERB|NOUN|PRON|ADJ !>nsubj man)

(94) Die Poesie allein **kann** sich auch von dieser Seite bis zur Höhe der Philosophie **erheben**, und ist nicht auf ironische Stellen begründet, wie die Rhetorik.²⁶⁸

With this respect, only Poetry **can arise** until the height of Philosophie, and it is not based on ironic positions, unlike Rhetoric.

(95) Nur dann zeige ich, daß ich einen Schriftsteller verstanden habe, wenn ich in seinem Geiste **handeln kann**;²⁶⁹

Only in that moment, when I **can act** in his spirit, I show that I have understood a writer.

(96) Der Dichter **kann** wenig vom Philosophen, dieser aber viel von ihm **lernen**.²⁷⁰

The poet **can learn** little from a philosopher, but the Philosopher can learn a lot from the poet.

In these examples, the modal verb is used to convey limits and scope of an artistic mean or discipline (94), to talk about the comprehension of the work of art (95), or about the qualities and abilities of the artist himself (96). As I already mentioned, Fragments mainly deal with aesthetic issues, i.e. they reason about the scope of art in general, about its goals in the society, about the role and the attitude of the artist, and related issues. The modal verb *können* is frequently used as linguistic mean to embody the capacity and the scope of art and the artist.

Another syntactic aspect of modal verbs that is worth investigating is their distribution between nonverbal and verbal predicates. In terms of dependency syntax, we investigate the distribution of the heads of modal verbs, i.e. we test whether modal verbs are headed more frequently by verbs, or by nonverbal elements. Moreover, we not only investigate how the distribution of the heads of modal verbs may vary across the text genres, but also among the single modal verbs within the same genre. More generally, this investigation can shed light on the role of nonverbal predicates in the modality. In other words: to what extent the nonverbal predication is used to encode modality? An example of the use of a


²⁶⁸ sent_id = lyceum-f42-s4.

²⁶⁹ sent_id = athenaeum-f287-s1.

²⁷⁰ sent_id = athenaeum-f131-s1.

modal verb with a nonverbal predicate is reported in (97) from GSD. In this case, the head of the modal verb *können* is the adjective *notwendig* (“necessary”):

(97) Solche Hilfsmittel **können notwendig sein**, [...].²⁷¹
 Some aids **can necessary sein**, [...].



In this respect, I extracted all those occurrences of modal verbs that govern a noun, an adjective or a verb.²⁷² I run these three queries first, in separate sessions:

L=können < NOUN;

L=können < ADJ;

L=können < VERB.

I then repeated these three queries for each lemma of the other five modal verbs. This extraction was performed on LIT and GSD only²⁷³. Table 68 summarizes the results. In this case, I grouped the RF concerning nouns and adjective as heads under the category *nonverbal predicates*, while I considered the results concerning verbs as head in the category *verbal predicates*. Table 69 shows the results concerning each single modal verb in detail, as well as each single POS with the role of governor.

HEAD	RF LIT	RF GSD
Verbal predicate	85.6	95.0
Nonverbal Predicate	22.2	9.7

Table 68 Distribution of modal verbs between verbal and nonverbal predicates, out of the total number of modal verbs.

As shown in Table 68, the vast majority of modal verbs work on verbal predicates rather than on nominal predicates in both the datasets. However, there is a gap between the distribution of modal verbs in nominal predicates between the two datasets: modal verbs are governed by a predicative element, i.e. a nonverbal element, much more in Fragments than in the GSD’s texts. In other words, the use of nominal predicates to encode modality is much more frequent in Fragments with respect to the web texts from GSD. However, if we look at the results in Table 69, the distribution of heads of those modal verbs

²⁷¹ sent_id = train-s3888.

²⁷² I reported the condition L= können only. However, the lemma of each of the six modal verbs was included in the first part of each query separated through the operator “|”.

²⁷³ I encountered some problems in extracting data from HDT. See above.

governed by nonverbal predicates significantly varies, not only across the two datasets, but also in each single dataset.

LEMMA	HEAD	RF LIT	RF GSD
können	VERB	93.1	96.1
	NOUN	3.6	1.3
	ADJ	3.2	2.5
müssen	VERB	80.6	93.8
	NOUN	8.1	0.9
	ADJ	11.3	5.3
sollen	VERB	70.3	94.3
	NOUN	15.3	2.2
	ADJ	14.4	3.5
dürfen	VERB	86.4	90.9
	NOUN	0.0	3.6
	ADJ	13.6	5.5
wollen	VERB	93.8	96.3
	NOUN	6.3	3.7
	ADJ	0.0	0.0
mögen	VERB	82.5	84.4
	NOUN	2.5	12.5
	ADJ	15	3.1

Table 69 Distribution of the POS of the heads of each modal verb in each dataset, out of the total number of occurrences of each modal verb.

Let us focus on the Fragments, in which the amount of data about the use of the modal verbs in nonverbal predicates is more conspicuous. As for the most frequent modal verb, i.e. *können*, it is used very rarely in nonverbal predicates, and, when used in these predicates, the heads are almost equally distributed between nouns and adjectives. The situation is rather different for *müssen*: in fact, with respect to *können*, this modal verb occurs much more frequently in nominal predicates, while adjectives tend to be slightly more frequent as heads with respect to nouns. An example of *müssen* used in a nominal predicate in the Fragments is reported in (98). In this case, the head of the modal verb is the adjective *frei* (“free”):

- (98) In der Wahl dieses Mittelglieds **muß** der Mensch durchaus **frey seyn**.²⁷⁴
 In the choice of these members the man must be absolutely free.

The percentage of heads played by predicative elements increases a lot in the case of *sollen*: 30% of heads are filled by nouns and adjectives, and they are almost equally divided between the two lexical classes. An example of *sollen* used with a nonverbal predicate in the Fragments is reported in (99). In this case, the head is the noun *Fabrikant* ('builder').

- (99) Aber **soll** der wahre Autor nicht auch **Fabrikant** sein?²⁷⁵
 But should the real author not be a builder as well?

As for *dürfen*, the frequency of this modal verb as child of nonverbal predicates is higher with respect to *können*, and none of the heads is occupied by nouns. An example of *sollen* in a nonverbal predicate in the Fragments is reported in (100), where the head is the adjective *intolerant* ('intolerant').

- (100) Denn wenn man nicht **intolerant sein dürfte**, wäre die Toleranz nichts.²⁷⁶
 When one **may** not **be intolerant**, there would be no tolerance.

In the end, as for *mögen*, the frequency of this modal in nominal predicates is higher with respect to *können*, and most of the heads in nominal predicates are occupied by adjectives. An example of *mögen* used in a nominal predicate in the Fragments is reported in (58), in which the head of the predicate is played by the adjective *größer*:

- (101) Sie **mag größer sein** als alle andern [...].²⁷⁷
 She like to be bigger than all the others [...].

²⁷⁴ bluethenstaub-f74-s3.

²⁷⁵ athenaeum-f367-s2.


²⁷⁶ athenaeum-f349-s4.

²⁷⁷ athenaeum-f324-s3.


In conclusion, modal verbs are mainly used with verbal predicates rather than with nonverbal predicates in Fragments. However, the distribution of verbal and nonverbal heads varies for each modal verb. The modal verb showing the highest use with nonverbal predicate is *sollen* (30% nonverbal), which also shows an equal distribution of heads between adjective and nouns. The second most frequent modal verb occurring with nonverbal predicate is *müssen*, (20% nonverbal), which tends to prefer adjectives to nouns, even if the gap between the two lexical classes in the role of heads is slight. *Mögen*, which is also the less frequent modal verb in this genre, is the third most frequent modal verb occurring with nonverbal predicates (18% nonverbal), and it seems to clearly prefer adjectives to nouns in the role of heads. On the contrary, *können*, which is the most frequent modal verb in the Fragments, mostly occurs with verbal predicates (93% verbal).

4.5.4 Position of Subordinate Clauses

Subordinate clauses can either precede or follow the higher clause that they modify. The sentence in (102) from the GSD dataset exemplifies a subordinate clause preceding the higher clause, which is the main clause of the sentence. In this case, the subordinate clause is an adverbial clause introduced by the marker *Obwohl* ('although'), while the subordinate predicate is a nonverbal predicate whose nonverbal part is the adjective *abstrakt* ('abstract'). The subject of the adverbial clause is the noun *Definition* ('definition'). The main predicate is the verb *steckt* ('stands'), therefore the *advl* relation spans from right to left, i.e. from the main predicate (head) to the subordinate predicate (dependent).

- (102) Obwohl die Definition der Bettzahlen sehr **abstrakt** ist, **steckt** hinter dieser eine Anschauung.²⁷⁸
Although the definition of Betti numbers is very **abstract**, an idea **stands** behind them.
- 

By contrast, (103) exemplifies a subordinate clause occurring after the higher clause in Fragments. In this case, the subordinate clause is an adverbial clause introduced by the marker *wenn* ('when'), and the subordinate predicate is the finite verb *weiß* ('knows'), which occupies the last position of the clause. In this case, the predicate of the adverbial clause, in turn, governs a non-finite predicate, which is the verb *erregen* ('stimulate').

- (103) Mich **interessirt** etwas, wenn es mich zu der Theilnahme zu erregen **weiß**.²⁷⁹
Something **interests** me, when it knows how to **stimulate** me to participation.
- 

One may ask how the subordinate clauses are syntactically dislocated in each genre. In other words, we are asking whether the genres differ with respect to each other not only for the frequency of subordinate predicates, which we already observed in Chart 8, but also for their position within the sentence. In terms of dependency syntax, we are asking, limited to those relations governing subordinate predicates, whether right-to-left dependencies (the subordinate predicate precedes the higher clause) are more frequent than left-to-right dependencies (the subordinate predicate follows the higher clause). I define this property of

²⁷⁸ sent_id = train-s2294.

²⁷⁹ sent_id = bluethenstaub-f35-s2.

the dependency relations as *orientation*. Thanks to SETS, I investigated the orientation of all those dependency relations encoding subordinate predicates across the three genres. For each of relation, I extracted any token governed by another token through that relation, first from right to left (the head is on the right), and then from left to right (the head is on the left). To extract the clausal subjects (*csubj* relation), I ran the following queries:


_ <csubj@R _²⁸⁰
_ <csubj@L _

Table 70 summarizes the orientation of the clausal subjects in each dataset.

DEPREL	ORIENTATION	RF LIT	RF GSD	RF HDT
csubj	right to left (pre)	57.4	29.8	40.2
	left to right (post)	42.6	70.2	59.8

Table 70 Orientation of clausal subjects (*csubj*).


As shown in Table 70, the orientation of clausal subjects clearly varies across the datasets. As for the Fragments, there is a rather similar distribution between the clausal subjects preceding the higher clause (right to left), and those following the higher clause (left to right), even if those preceding the higher clause are slightly more frequent. On the contrary, there is a considerable gap between them in GSD, where almost two clausal subjects out of three go from left to right. As for HDT, the discrepancy between the two groups clearly decreases, making the orientation of this relation in this genre rather similar to that observed in Fragments. (104) exemplifies a right-to-left clausal subject in Fragments. The clausal subject is the verb *verrät* ('betrays'), whose nominal subject is the pronoun *Wer* ('who'). The clausal subject depends forward on *verrät* in the main clause. In this case, the clausal subject occupies the *Vorfled* of the main clause, which is the position usually filled by the nominal subject in a verb-second clause. As a result, the main clause begins with the second position, which is in fact occupied by the verbal predicate.

(104)  Wer die Wahrheit **verrät**, **verrät** sich selbst.²⁸¹
Who the truth betrays, betrays himself.

²⁸⁰ This query should be read as follows: return any token which is governed from the right through *csubj* relation.

²⁸¹ sent_id = bluethenstaub-f39-s3.

Intuitively, I would expect a higher frequency of the clausal subject preceding the higher clause with respect to those following the higher clause. In fact, the *Vorfeld* is usually occupied by the nominal subject in verb-second clauses. By contrast, Fragments show a very high RF of clausal subjects occurring after the higher clause (left to right). (105) exemplifies a left-to-right clausal subject in the Fragments.

(105)  Es ist **unmöglich**, jemanden ein Ärgernis zu **geben**, wenn er 's nicht nehmen will.
It is impossible, to give someone an offense, when he does not want to take it.

In (105), the predicate of the clausal subject is the non-finite verb *geben* ('to give'), which is preceded by the infinitival marker *zu*. The main clause precedes the clausal subject, and the main predicate is a nonverbal predicate, i.e. the adjective *unmöglich* ('impossible'). In this case, the personal pronoun *es* plays the role of expletive element of the nonverbal predicate in the main clause. Therefore, the role of subject must necessarily be played by a clause that occurs after the predicate. Such structure, in which an expletive element in the main clause depends on a nonverbal predicate, could be very common in those sentences with clausal subjects, causing the high RF of left-to-right *csubj* relation in this genre. To test whether this structure is actually responsible for the high RF left-to-right clausal subjects in Fragments (or whether this is caused, for instance, by errors in the annotation), I ran the following query, in order to retrieve all those clausal subjects that are governed by a token on the left, which, in turn, governs an expletive element:


`_ <csubj@L (_ >expl _)`

The result is that 41% of clausal subjects in Fragments matched this query²⁸², which is almost exactly the RF of left-to-right clausal subjects (42%). Consequently, it seems that all the clausal subjects with the head on the right in Fragments are due to the [es + nonverbal predicate] construction in the higher clause.


In GSD, the frequency of clausal subjects preceding the higher clause decreases a lot, and the RF of those postponed with respect to the higher clause increases a lot. (106) exemplifies the use of a clausal subject preceding the main clause in this dataset. In this case, the clausal subject consists in a nonverbal predicate, whose predicative element is the noun *Skandal* ('scandal'), while the pronoun *was* ('what') is the nominal subject of *Skandal*. The main predicate is the noun *Anekdote* ('anecdote'), therefore the *csubj* relation spans from *Anekdote* to *Skandal*. As observed also in (104), the clausal subject occupies the *Vorfeld* of the main clause, therefore, in this case, the main clause begins with the auxiliary of the nonverbal predicate, i.e.

²⁸² RF of the hittokens returned by the query was calculated with respect to the total occurrences of the *csubj* relation.

war, which usually occupies the second position in declarative clauses, while the nonverbal part occupies the last position.


- (106) Was heutzutage ein **Skandal** wäre, war damals nur eine witzige **Anekdote**.²⁸³
 What today would be a **Scandal**, was then just a witty **anecdote**.
- 

By contrast, (107) shows a clausal subject that follows the higher clause from GSD, i.e. with the head on the left. In this case, the clausal subject is introduced by the complementizer *dass* ('that'), the predicate of the clausal subject is the verb *habe* ('have'), which depends back on the main predicate, which is the noun *Ende* ('end'). The nominal subject of the verb *habe* is the pronoun *ich* ('I').

- (107) Das **Ende** vom Lied ist, dass ich eine Gehörgangentzündung auf beiden Ohren **habe**.²⁸⁴
 The end of the story is that I have an inflammation of the ear canal in both ears.
- 

I repeated the same test performed for Fragments, therefore I tested how many of the clausal subjects occurring after the higher clause are due to a construction with an expletive element in the higher clause. The result is that 65% of them are due to this reason. It seems therefore that a significant part of clausal subjects in this dataset are due to constructions similar to that reported in (107).

In HDT, the frequency of left-to-right clausal subjects increases decreases respect to GSD, but the frequency of clausal subjects with the head on the left remains higher with respect to those with the head on the left. (108) illustrates the use of a left-to-right clausal subject in this dataset. As already observed in GSD, the postponed clausal subject is not generated by the structure [es + nonverbal predicate] in the main clause. In this case, the main predicate the verb *bleibt* ('remains'), modified the adverb *Fraglich* ('doubtful'), and by an oblique modifier, i.e. the noun *Prozeß* ('litigation'). The expletive element is omitted instead. In the clausal subject, the subordinate nonverbal predicate is the adjective *tätig* ('active').


- (108) Fraglich **bleibt** im Prozeß, ob CompuServe als Service-Provider oder Access-Provider **tätig** war.
 Remains **doubtful** in the Litigation, whether was active as CompuServe or as Service-Provider.
- 

I repeated on HDT the same test previously run on both LIT and GSD, to check to what extent the structure [es + nonverbal predicate] contribute to the RF of left-to-right clausal subjects. I therefore run

²⁸³ sent_id = train-s383.

²⁸⁴ sent_id = hdt-s1096.

the same query on HDT as well. It turned out that 34.6% of clausal subjects are generated by this structure in the main clause. The syntax of clausal subjects postponed to the higher clause seems therefore more variable in HDT and in GSD with respect to LIT²⁸⁵. An example of the use of the structure [es + nonverbal predicate] to postpone the clausal subject in HDT is shown in (109). The clausal subject is introduced by the marker *wann*. The predicate of the clausal subject is the verb *behoben* (‘removed’), whose nominal subject is the pronoun *diese* (‘these’). The main predicate is a nonverbal predicate, whose nonverbal part is the adjective *absehbar* (‘predictable’).

(109)  Es sei noch nicht **absehbar**, wann diese **behoben** werden könne.²⁸⁶
It would not be still predictable, when these could be removed.

DEPREL	ORIENTATION	RF LIT	RF GSD	RF HDT
ccomp	right to left	3.8	17.8	13.6
	left to right	96.2	82.2	86.4

Table 71 Orientation of clausal complements.

Table 71 shows the orientation of clausal complements in each dataset (*ccomp* relation), i.e. those clauses with the role of direct object of a higher clause. I extracted both right-to-left and left-to-right clausal complements through the following queries:

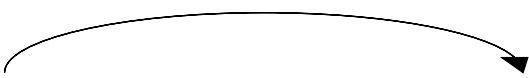
_ <ccomp@R _
_ <ccomp@L _

As shown in Table 71, most clausal complements have the head on their left in each dataset, especially in Fragments, i.e. clausal complements are mostly postponed with respect to the higher clause. This is the result that I would intuitively expect from the distribution of this relation, since direct objects in German declarative clauses (v2 clauses) usually occur right after the predicate. An example of a clausal complement postponed with respect to the main clause in Fragments is reported in (110). The clausal complement is introduced by the complementizer *daß* (‘that’), and the nonverbal predicate of the clausal complement is the noun *Vermögen* (‘ability’), followed by the copula *ist* in last position. The main clause


²⁸⁵ However, we cannot absolutely exclude the presence of some errors in the annotation of the *csubj* relation either, in this dataset as well as in GSD.

²⁸⁶ sent_id = hdt-s815.

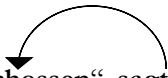
precedes the clausal complement, and the main predicate is the verb *wußten* ('knew'), while the nominal subject is the noun *Römer* ('Romans').


- (110) Die Römer **wußten** daß der Witz ein prophetisches **Vermögen** ist;²⁸⁷
 The Romans **knew**, that the wit was a prophetic **ability**.
- 

However, almost 20% of clausal complements in GSD comes before the higher clause, i.e. their head is on the right. (111) exemplifies this use of a clausal complement in GSD. In this case, the first clause reports a quoted speech, and plays the role of clausal object of the main clause, which follows. The predicate of the clausal object is the verb *zahlen* ('pay'), modified by the auxiliary *werden* to build the future form, while the main predicate is the *Praeteritum* verb *sagte* ('said'). The nominal subject of the clausal object is the personal pronoun *sie* ('they'), and the subject of the main clause is the proper noun *Mubarak*, which occurs in postverbal position.

- (111) "Eines Tages werden sie einen hohen Preis dafür **zahlen**", **sagte** Mubarak der Pariser Zeitung Le Monde
 "One day they will **pay** a very high price for this", Mubarak **said** to the Parisian newspaper Le Monde.
- 

Such structure, in which a quoted speech is the first clause and it is followed by a main clause introduced by a speech verb, such as *sagen*, occurs rather frequently in the news style, since it allows to put the focus on the quoted information. I provide other examples of the same structure retrieved from GSD in both (112) and (113).

- (112) "Die Krankenhäuser wurden systematisch **beschossen**", **sagte** Ljubic.²⁸⁸
 "Hospitals were systematically **closed**", Ljubic **said**.
- 

- (113) "Die Lage ist völlig außer Kontrolle, es herrscht das totale Chaos", **sagte** ein Vollzugsbeamter.²⁸⁹
 "The position is totally out of control, the complete chaos dominates", a guardian **said**.
- 

²⁸⁷ sent_id = lyceum-f126-s1.

²⁸⁸ sent_id = train-s1629.

²⁸⁹ sent_id = train-s1586.


As a consequence, the news texts collected in the GSD dataset must play a role in the rather considerable frequency of the clausal complements occurring before the main clause.

It is worth noting that the status of the relation spanning from a speech verb to the predicate of a quoted speech that stands before the declarative clause is actually debated in UD. In fact, according to the UD 2 scheme, when a structure similar to that shown in (111), (112), and (113) occurs, the predicate of the quoted speech and the speech verb in the higher clause should actually be attached through *parataxis* relation, and not through *ccomp* relation. Moreover, this relation should span from the predicate of the reported speech to the speech verb, i.e. the reported speech should be considered as the main clause. For instance, in (111), the parataxis relation should span from *zahlen* to *sagte*, while here happens the opposite. I here report the passage from the UD official guidelines:

When the reported speech follows the speech verb and is separated by a colon, the reported speech forms a main clause that attaches to the preceding main clause with a parataxis relation, hence with the speech verb as its head. However, when the speech verb occurs as a medial or final parenthetical, the relation is reversed and the speech verb is treated as a parataxis of the reported speech. This analysis is not uncontroversial but follows many authorities, such as Huddleston and Pullum (2002), *The Cambridge Grammar of the English Language* (see chapter 11, section 9).²⁹⁰

As for HDT, most of the clausal objects occurs after the main clause in this genre too. By contrast, (114) exemplifies the use of a clausal complement preceding the higher clause, therefore generating a right-to-left *ccomp* relation. The clausal object is introduced by the adverb *wie* ('how'), here in the role of complementizer, while the predicate of the clausal object is the verb *lassen* ('leave'), which, in turn, governs a secondary predicate, i.e. the verb *ausführen* ('carry out'). The main predicate is the verb *verraten* ('reveal'), while the noun *Außenminister* ('Foreign Minister') is the nominal subject.

(114) Wie sich diese Zensur ausführen **lassen** soll, hat der Außenminister nicht **verraten**.²⁹¹
How this censorship should be carried out, the Foreign Minister has not revealed.



²⁹⁰ Cf. <https://universaldependencies.org/u/dep/parataxis.html> (last access on 25th June 2020).

²⁹¹ sent_id = hdt-s102331.

DEPREL	ORIENTATION	RF LIT	RF GSD	RF HDT
xcomp	right to left	47.9	34.6	25.2
	left to right	52.1	65.4	74.8


Table 72 Orientation of open clausal complements (xcomp).

Table 72 shows the orientation of the open clausal complements in each dataset. I extracted them through the following queries:

```
_ <xcomp@R _
_ <xcomp@L _
```


Let us focus on Fragments. In this genre, open clausal complements are almost equally divided between those preceding and those following the higher clause. (115) exemplifies the use of an open clausal complement occurring before the higher clause. In this case, the open clausal complement is a final clause introduced by the marker *Um*. The non-finite predicate of the final clause is the verb *schreiben* ('write'), while the predicate of the main clause is the verb *interessieren* ('interests'), which occupies the last position of the clause, since the second position right after the final clause is filled by the auxiliary verb *muß* ('must').

(115) **Um** über einen Gegenstand gut **schreiben zu** können, muß man sich nicht mehr für ihn **interessieren**.²⁹²
 In order to be able **to write** good about an object, we must no longer **interest** in it.



The sentence in (116) exemplifies an open clausal complement occurring before its head in the role of secondary predicate. In this case, the main predicate is the past participle *geworden* ('become') at the passive form, while the noun *Fragmente* ('Fragments') has the role of secondary predicate, therefore it depends on *geworden* through *xcomp* relation.

(116) Viele Werke der Alten sind **Fragmente geworden**.²⁹³
 A lot of works by the ancients have become fragments.



²⁹² sent_id = lyceum-f37-s1.

²⁹³ sent_id = athenaeum-f24-s1.

To check how many of the right-to-left *xcomp* relations are caused by on-finite clauses and how many by secondary predicates instead, I ran the following queries:


```
_ >mark _ <xcomp@R _
_!>mark _ <xcomp@R _
```

Through the first query, I searched for all those right-to-left open clausal complements that also govern a subordination marker (non-finite clauses). Through the second query, I searched for all those right-to-left open clausal complements that are not governed by any subordination marker (secondary predicates). The result is that 80% of right-to-left open clausal complements are caused by secondary predication, and only 20% by non-finite clauses. Moreover, I verified how many of the right-to-left non-finite clauses are actually final clauses. I therefore searched for all those right-to-left open clausal complements that also govern the subordination marker *um*, which introduces this of non-finite clauses. I ran the following query:

```
_ >mark um <xcomp@R _
```

The result is that only 15% of the right-to-left non-finite clauses are actually final clauses introduced by the marker *um*. An example was already shown in (115). We can therefore conclude that the anticipation of non-finite clauses with respect to the main clause (or the higher clause) and, in particular, the anticipation of the final clauses, are rather rare in Fragments. On the contrary, open clausal complements occurring before the head are rather frequent, such as that shown in (115).


As for left-to-right open clausal complement, an example is reported in (117) from Fragments. In this case, there is a non-finite clause occurring after the main clause, even if there is a parenthetical clause between them. The main predicate is a nonverbal predicate, whose nonverbal part is the adjective *fein* ('refined'). The open clausal complement is a final clause introduced by *um*, in which the non-finite verb is *wegzuläugnen* ('avoid'), which includes the final marker *zu*, because it is a separable verb.

- (117) Sie ist fein **genug, um** alles Große **wegzuläugnen**.²⁹⁴
 She is **refined** enough **to avoid** everything rough.
- 

²⁹⁴ sent_id = bluethenstaub-f106-s3.

Example (118) shows the use of a left-to-right *xcomp* relation attaching a secondary predicate to the main predicate, within a main clause. The main predicate is the verb *wird* ('becomes'), while the secondary predicate is the adjective *sophistisch* ('sophistic').

(118) Auch die Universalhistorie **wird** **sophistisch**, [...].²⁹⁵
 Also the universal history **becomes** **sophistic**, [...].



As done before, I then investigated how many right-to-left open clausal complements are actually non-finite clauses, and how many are secondary predicates instead. I therefore run the following queries:


_>mark _ <xcomp@L _
 _!>mark _ <xcomp@L _

The result is that 45% of left-to-right *xcomp* relations governs non-finite clauses, while 45% governs secondary predicates. The distribution is therefore very different with respect to what observed for right-to-left open clausal complements. In this case, the two functions seem almost equally distributed in those open clausal complements postponed with respect to the higher clause. I also tested how many of these non-finite clauses are actually final clauses. I ran the following query:

_>mark um <xcomp@L _

The result is that only 9% of the returned left-to-right non-finite clauses are actually final clauses introduced by the marker *um*, while most of them are simply infinitival clauses without their own subject. (125) exemplifies the use of this kind of open clausal complement. In this case, the non-finite verb is *erwarten* ('expect') introduced by the marker *zu* and by the subordinating conjunction *ohne* ('without').

(119) Man soll von jedermann Genie **fordern**, aber ohne es **zu erwarten**.²⁹⁶
 One should request Genius from everyone, without expecting it.



As a result, we can say that Fragments are overall marked by a very low frequency of final clauses. The orientation of the open clausal complements is almost equally distributed between those following and those preceding the higher clause. In the first case, most of the complements are secondary predicates, while the frequency of infinitival clauses in this position is rather low. By contrast, in the second case,

²⁹⁵ sent_id = athenaeum-f223-s3.

²⁹⁶ sent_id = lyceum-f16-s2.


there is almost an equal distribution between infinitival clauses and secondary predicates, therefore the frequency of infinitival clauses postponed with respect to the higher clause is significantly higher with respect to that of those preceding the higher clause.

Let us now consider GSD. As showed in Table (72), the frequency of those open clausal complements preceding the higher clause decreases with respect to Fragments, and most open clausal complements have the head on their left. I verified how many of these clausal complements are actually non-finite clauses, how many of them final clauses, and how many secondary complements instead. I run the same queries previously run on LIT, i.e.:

```
_>mark _ <xcomp@R _
_!>mark _ <xcomp@R _
_>mark um <xcomp@R _
```

Interestingly, the first query did not return any token, therefore it seems that none of the right-to-left open clausal complements in GSD is a non-finite clause (or a final clause). On the contrary, 92% of the right-to-left open clausal complements were returned by the second query. It seems therefore that almost all the right-to-left *xcomp* relations in this dataset encode secondary predicates. (126) exemplifies the use of one of these relations in GSD. In this case, the main predicate is the verb *sagt* ('said'), while the subordinate predicate is *werden* ('become'), which governs a secondary predicate, i.e. the adjective *direkter* ('more direct').

(120) Die Demokratie müsse **direkter werden**, sagte sie.²⁹⁷
 Democracy should **become more direct**, she said.



I repeated the three queries for the left-to-right open clausal complements:

```
_>mark _ <xcomp@L _
_!>mark _ <xcomp@L _
_>mark um <xcomp@L _
```

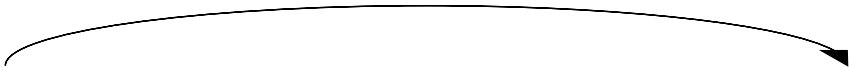
The result was that 27% of the left-to-right open clausal complements, i.e. those coming after the higher clause, are non-finite clauses. However, the last query did not return any token, therefore none of these

²⁹⁷ sent_id = test-s371.

non-finite clauses is actually a final clause²⁹⁸. Most of the left-to-right open clausal complements should therefore be secondary predicates. (121) exemplifies a non-finite clause occurring after the higher clause. The main predicate is *verspricht* (‘promises’), which generates a non-finite clause, whose predicate is *machen* (‘make’), introduced by the infinitival marker *zu*. It is worth noting that, in this case, the clausal complement depends on the main predicate through *xcomp* relation and not through *ccomp* relation, even if it should be considered as clausal object. I here report the explanation from the UD official guidelines (version 2):


If the subject of the clausal complement is controlled (that is, must be the same as the higher subject or object, with no other possible interpretation) the appropriate relation is *xcomp*.²⁹⁹

(121) Reilly **verspricht**, in dem nächsten Jahr eine neue gewinnbringende Erfindung zu **machen**.³⁰⁰
 Reilly promises to make e new profitable discovering next year.



The sentence in (128) exemplifies a secondary predicate postponed with respect to the first predicate. In this case, the main predicate is the *Preteritum* verb *blieb* (‘remained’) in second position, while the secondary predicate is the adjective *geöffnet* (‘open’)³⁰¹ in last position.

(122) Die Schule **blieb** bis 1939 und noch einmal zwischen 1946 und 1950 **geöffnet**.³⁰²
 The school remained open until 1939, and then once again from 1964 to 1950.



In the end, as for HDT, the vast majority of *xcomp* relations spans right-to-left. I performed the same test done on the other datasets, to check the distribution of non-finite clauses among the right-to-left open clausal complements³⁰³. I used the same queries run in the previous tests. The result was that that 26% of these complements do not govern any subordination marker, while 27 % of them subordination govern a marker. As a result, about 50% of the right-to-left open clausal complements were not returned by any of the following queries: `_>mark _<xcomp@L _; _!>mark _<xcomp@L _`. An example of a right-to-left open clausal complement in HDT is reported in (129). In this case, it is a non-finite clause whose predicate is the verb *erzielen* (‘gain’), which is introduced by the marker *zu*. The main predicate is the verb *befinde* (‘find’).

²⁹⁸ To test whether the result was caused by errors in the annotation of the marker *um*, the query was also repeated as follows, i.e. removing the type of the relation: `_> um <xcomp@L _`. No relevant token was returned either.

²⁹⁹ Cf. <https://universaldependencies.org/u/dep/ccomp.html>.

³⁰⁰ sent_id = train-s4490.

³⁰¹ It is a past participle used as adjective.

³⁰² sent_id = train-s2677.

³⁰³ I run the queries on right-to-left *xcomp* relations only, since they are the vast majority of *xcomp* relations in this dataset.

- (123) Der Unternehmensbereich **befinde** sich auf dem besten Wege, nächstes Jahr Gewinn zu **erzielen**.³⁰⁴
 The corporate division **find** itself on the best way next to **gain** profit next year.

DEPREL	ORIENTATION	RF LIT	RF GSD	RF HDT
advcl	right to left	27.6	29.9	30.9
	left to right	72.4	70.1	69.1

Table 73 Orientation of adverbial clauses (advcl).

Table 73 shows the orientation of the *advcl* relations in each dataset, i.e. those relations encoding adverbial clause. I extracted the two groups of relations though the following queries:

_ <advcl@R _
 _ <advcl@L _

In this case, the distribution is rather similar across all the datasets: about 30% of adverbial clauses occurs before the higher clause (right-to-left), and almost 70% of them occurs after the higher clause. Overall, we can therefore observe that the position of adverbial clauses seems to be more conservative across all the three genres with respect to the position of the other clausal complements. I here report two examples of the orientation of the *advcl* relation from the Fragments. (124) exemplifies a right-to-left adverbial clause. In this case, the adverbial clause is introduced by the subordinating conjunction *Wenn* (‘When’), and the subordinate predicate is the verb *lernt* (‘meet’). The main predicate is the verb *wirken* (‘affect’).


- (124) Wenn der Künstler dann auf Reisen romantische Szenen kennen **lernt**, so **wirken** sie desto mächtiger auf ihn.³⁰⁵
 When the artist **meet** romantic scenes on travel, they **affect** him even more powerfully.

The sentence in (125) illustrates an example of a left-to-right adverbial clause in Fragments. The main predicate is the verb *konstruiert* (‘builds’), which, in turn, is coordinated to the verb *schafft* (‘makes’). The adverbial clause is introduced by the adverb *wie* (‘how’), which here works as subordinating conjunction.

³⁰⁴ sent_id = hdt-s3608.

³⁰⁵ sent_id = athenaeum-f190-s4.

The predicate of the adverbial clause is the verb *sein* ('to be'), which is here not used as a copula, but as a verbal predicate.

- (125) Der synthetische Schriftsteller **konstruiert** und schafft sich einen Leser, wie er **sein** soll.³⁰⁶
The synthetic writer **builds** and shape a reader for him, as he should **be**.
- 

³⁰⁶ sent_id = lyceum-f112-s3.

5 Conclusions

In this thesis, I presented and analysed a new historical dependency treebank for the German language. At the same time, I attempted to show the benefits offered by this resource to the linguistic analysis of the literary language. Since the treebank was specifically designed to collect literary texts, I defined it a literary treebank. In particular, the treebank focuses on a specific literary genre, i.e. the Fragments of the early German Romanticism, which were published in the last decade of 18th century in two important literary magazines of that age, i.e. *Lyceum der schönen Künste* and, most importantly, *Athenaeum*, which is considered the reference magazine of the early Romantic movement. Fragments are very short texts, often in aphorism-like form, which mainly deal with aesthetic issues, i.e. concerning art, poetry, beauty, and related topics. Most of them were written by Friedrich Schlegel, which is widely considered the founding father of this literary genre. The Fragments are considered a revolutionary genre in the German literary history, not only for the witty, and often cryptic, way to address philosophical issues, but also for their concise style, which deliberately clashes with the long and elaborated prose of the neoclassical authors of the same age, such as Goethe and Schiller. Their style perfectly embodied the spirit and the values of the new-born early Romanticism.³⁰⁷ Both the cultural and the linguistic importance of the Fragments made this textual genre absolutely worthy of being represented for the first time in a dependency treebank. In fact, such a resource allows a wide spectrum of empirical linguistic investigations, which can also, and especially, exploit the dependency annotation for analysing a series of linguistic features of the genre, especially on the syntactic level, which can hardly be investigated through traditional corpus-based approaches, whose annotation do not go beyond the level of parts of speech. Besides discussing the development of the resource and the linguistic annotation of the Fragments, this thesis attempted to demonstrate the benefits offered by a treebank-based approach, and specifically a dependency-based approach, to the linguistic analysis of a literary genre, which is still a mostly unexplored area in the field of linguistic approaches to literature. The Fragments were therefore taken as testbed for this approach, and different aspects of their language were empirically analysed exploiting all the levels of annotation, but especially the dependency relations.

The treebank was developed within the Universal Dependencies (UD) framework, i.e. an international research project which aims to develop a consistent grammar annotation across different human languages, as well as a huge online repository of multilingual treebanks freely available for research purposes. Therefore, the treebank implements the UD annotation scheme version 2, and it is encoded in the CoNLL-U data format, which is the official format adopted to encode treebanks in the UD. The

³⁰⁷ The first definition of Romantic Poetry was written in the *Athenaeum Fragmente* (Fragment 116).

treebank implements four levels of metadata: lemma (in the field LEMMA), coarse-grained part of speech (UPOS, through the UTS tagset), fine-grained part of speech (XPOS, through the STTS tagset), and syntactic information, which consists in the syntactic head of the token (in the field HEAD), and the syntactic function played by the dependent with respect to the head (in the field DEPREL). After passing the official UD validation test, which automatically checks the consistency of the annotation of a treebank with the official UD guidelines, the literary treebank was published in the UD 2.4 release under CC BY-NC-SA 4.0 license. Currently, it is available in UD 2.6 as well, which is the latest UD release. Table 74 summarizes the treebank in its current state, while Table 75 reports the texts that are collected.

UD ID	Genre	Historical Variety	Sentences	Tokens
UD-German-LIT	Fragments	Modern German (end of 18 th century)	1,922	40,545

Table 74 Portrait of the Literary Treebank in the latest release (UD2.6).

Author	Work	Metadata	Raw-Text Source
F. Schlegel	<i>Lyceum Fragmente</i> (<i>Kritische Fragmente</i>) (entire collection)	LEMMA UPOS	zeno.org
F. Schlegel et al.	<i>Athenäums-Fragmente</i> (fragments from 1 to 421)	XPOS HEAD DEPREL	
Novalis	<i>Blüthenstaub</i> (entire collection)		

Table 75 Texts collected in the literary treebank (UD 2.6).

The development of the treebank was dealt with in Chapter 2. The Fragments were annotated through a semi-automatic method using data-driven NLP tools. First of all, I search for a source of the raw texts of the Fragments. I found them freely available as plain texts on the website *zeno.org*. I therefore collected them to build the source corpus of the treebank. I then manually annotated a small test set (about 7.000 tokens) from the source corpus, in order to test a set of data-driven NLP tools and find the best performing ones on this genre. All the candidate tools were trained (or pre trained) on data of contemporary German treebanks, which were the only ones available to train the models, and which collect contemporary news texts and web texts. I tested one lemmatizer, two POS-taggers and four different dependency parsers (see Chapter 2). In the end, I chose the best NLP pipeline, which was the Mate Tools’ pipeline in Anna 3.6 implementation, which consists in a lemmatizer and a POS-tagger both based on Support Vector Machines (SVM), and a graph-based dependency parser. This pipeline requires the files to be in CoNLL-2009 format, therefore the initial test set, as well as all the Fragments that were successfully processed, were converted into this format through a Python script. Once the semi-automatic annotation was terminated, the final file was converted into CoNLL-U format through another Python script. The tools of the Mate Tool’s pipeline used to process the Fragments, as well as the models on which they were trained, and their overall accuracy attained on the initial test set are summarized in Table 76, while the flow of the pipeline is sketched in Figure 42.

Task	Model	Tool	Test Set	Accuracy %
Lemmatization	Pre-trained on <i>Tiger Corpus</i>		frag1.conll09	97.6
POS-tagging (STTS)	Pre-trained on <i>Tiger Corpus</i>	Mate Tools –	frag1_lem.conll09	97.3
Dependency Parsing	Trained on <i>de-ud2.0- train.conllu</i>	Anna 3.6	frag1_xpos.conll09	67.2 (LAS)

Table 76 NLP Tools and models used to process the Fragments, and their accuracy on the initial test set.



Figure 41 NLP pipeline used to process the Fragments.

As shown in Table 76, the accuracy on Fragments for both lemmatization and POS-tagging based on pretrained models was rather high. As for POS-tagging, it is worth noting that I opted for using the Stuttgart-Tübingen-Tagset (STTS) rather than the Universal Tagset (UTS), since the accuracy by the Mate Tools' POS-tagger was significantly higher using the STTS with respect to that attained using the UTS. Also, the use of STTS in POS-tagging caused a slight improvement in dependency-parsing performance. In the development phase, I therefore processed all the other Fragments by implementing the STTS in the POS field of the CoNLL-2009 format. At the end of the annotation, the STTS was automatically converted into the UTS to obtain the UPOS. As for dependency parsing, as expected, the accuracy dropped with respect to the previous tasks, attaining a Labeled Attachment Score (LAS) of 67.2% on the initial test set. The in-depth evaluation revealed that all the clausal dependents encoding subordination, but especially adverbial clauses (advcl), clausal subjects (csubj), clausal complement (ccomp), and open clausal complements (xcomp) caused the accuracy of parsing to drop, being more frequent in Fragments rather than in the variety of the training set. I observed how the low accuracy on these relations was correlated to the dependency length, since subordinate clauses are verb-final clauses in German, therefore they generated long-standing dependency relations. Moreover, direct objects, oblique modifiers and nominal modifiers, whose frequency in the test set was significantly higher with respect to that of the clausal modifiers, overall attained a rather low accuracy, therefore contributing to the degradation of the parsing performance. I observed how a different distribution of the positions of nominal subjects (nsubj) and direct objects (obj) with respect to the predicate between the training set (UD) and the test set (Fragments) could have influenced the accuracy on these relations. The accuracy on single dependency relations by the Mate Tool's graph-based parser on the initial test set of Fragments, as well as the RF of these relations in the test set, are shown in Figure 43. However, now that the first version of the literary treebank is completed, the Anna 3.6 graph-based parser should be retrained on a training set of Fragments and re-tested on the initial test set, in order to evaluate the increase in accuracy, if any, when trained on in-domain data with respect to the model trained on UD. A high-performing dependency parsing model would be very helpful to extend the amount of syntactically annotated data of the genre of Fragments in future, for example to terminate the annotation of the source corpus.

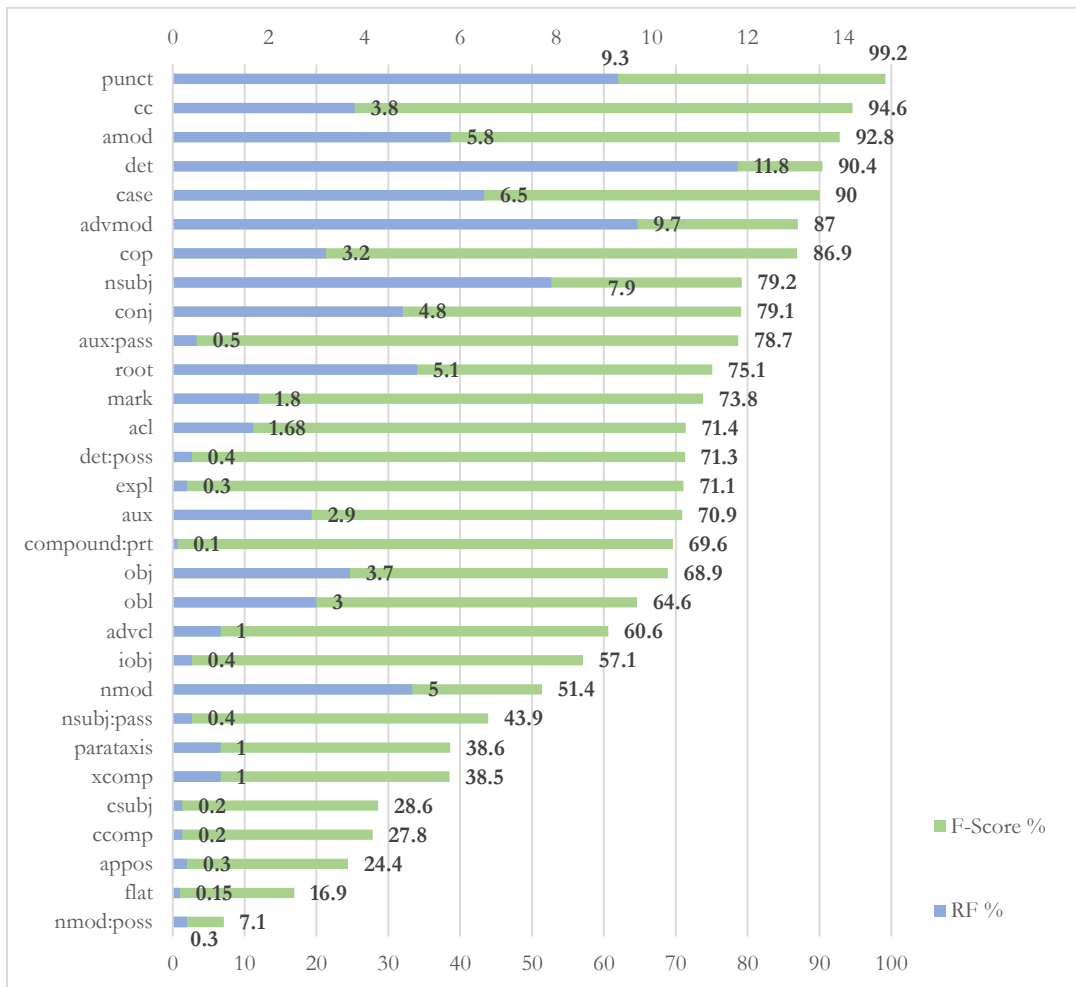


Figure 42 Accuracy by Anna 3.6 graph-based dependency parser on single dependency relations.

The annotation of the Fragments was illustrated in Chapter 3. I applied the UD scheme version 2. The UD standard is designed as a mixed functional-structural system, i.e. basic dependencies encode both the grammar function and the structural category of the dependent. The grammar function is the syntactic function played by the dependent with respect to the head. The structural category is the role of the dependent in the structure of the sentence. For instance, the functional category changes according to whether it generates a dependency within a clause or, on the contrary, whether it introduces a dependency that spans into a new clause. An example of the annotation through UD version 2 is shown in Figure 44, encoded in the CoNLL-U file of the treebank. The same sentence is illustrated in tree-like form in Figure 45. The sentence is also reported in linear unannotated form in (1), followed by a translation. As for the syntactic annotation of this sentence, the main predicate is the higher node of the sentence, i.e. the root node, whose head is therefore 0. In this case, the main predicate is a nonverbal predicate, therefore the root node is the noun *Feindin* ('enemy'). All the other elements are either direct or indirect child nodes, i.e. dependents, of the predicate. Each dependency relation spans from the governor, i.e. the head, to the

dependent, and encodes the syntactic function that the dependent has with respect to the governor. For instance, the copula *ist* depends on the nonverbal predicate through *cop* relations. In the CoNLL-U file, each sentence is preceded by a series of fields introduced by ‘#’. Some of them are mandatory, such as the univocal ID in the treebank (*sent_id*). However, I added optional fields to preserve the information concerning the work form which each sentence comes from, as well as the author and the textual genre. Moreover, the ID of each sentence (*sent_id*) incorporates the name of the work to which the sentence belongs to, as well as the exact position occupied by the sentence in that work. For instance, *sent_id = bluethestaub-f13-s1* in Figure 44 means that the sentence is the first sentence (s1) of the Fragment 13 (f-13) in the collection named *Bluethenstaub*. This solution aims to conceive the treebanked data as the exact annotated counterpart of the original literary texts, therefore maintaining a parallelism with respect to the source texts. In a long term-perspective, this allows the extraction of linguistic information concerning specific authors, works and genres for specific linguistic analysis.

- (1) Die Natur ist Feindin der ewiger Besitzungen.
The Nature is Enemy of the eternal property.

```
# newpar id = bluethenstaub-f13
# genre = fragments
# author = Novalis
# work = Blütenstaub
# sent_id = bluethenstaub-f13-s1
# text = Die Natur ist Feindin ewiger Besitzungen.
1  Die der DET ART _ 2 det _ _
2  Natur Natur NOUN NN _ 4 nsubj _ _
3  ist sein AUX VAFIN _ 4 cop _ _
4  Feindin Feindin NOUN NN _ 0 root _ _
5  ewiger ewig ADJ ADJA _ 6 amod _ _
6  Besitzungen Besitzzung NOUN NN _ 4 nmod _
7  . -- PUNCT $. _ 4 punct _ _
```

Figure 43 Representation in CoNLL-U format of the sentence “Die Natur ist Feindin der ewiger Besitzungen.” according to UD scheme version 2.

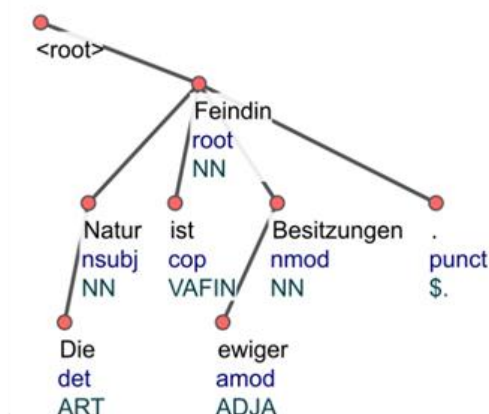


Figure 44 Dependency representation of the sentence “Die Natur ist Feindin der ewiger Besetzung” according to UD scheme version 2.

The treebank-based analysis of Fragments was dealt with in Chapter 4. The aim of this analysis was twofold. On the one hand, this analysis attempted to demonstrate what can be detected in the language of a literary genre through a treebank-based approach, especially in syntax, which cannot be seen at naked eyes or through the common corpus-based approaches, which cannot exploit any dependency information. On the other hand, the linguistic features of Fragments were not investigated in absolute terms, but with respect to two different textual genres. I chose the two genres that are currently represented in the two main UD treebanks for German, i.e. web texts from the GSD Treebank, which mainly come from Wikipedia, and web-news texts from the HDT treebank, which all came from the website *heise.de*, and therefore belong to a subdomain of web news, i.e. news from the field of technology. Therefore, the analysis also provided empirical evidence on the differences between the new literary variety that was here represented in a dependency treebank for the first time, and those that are usually considered the *de facto* standard variety to work with (dependency) treebanks, not only in UD but in treebanking in general.

For the analysis, I selected a dataset in CoNLL-U format for each of the three treebanks (UD 2.5), and I selected a tool that allows to retrieve specific evidence from the datasets through formal queries, especially through dependency relations. I opted for SETS, a free online tool integrated in the UD infrastructure, which is maintained by the Turku NLP group. The query language implemented in this tool is loosely inspired by TGrep2 and TRegex, but it is specifically designed for querying general dependency graphs. I then began the analysis. In each part of the analysis, I always reported the relative frequency (RF) of the specific parameters under investigation, and I discussed the results, by providing different examples from the datasets. The RF was always calculated through a script written in R and run in R Studio. The analysis was structured as follows. I first provided an insight into the distribution of parts of speech (POS), showing how the dependency information can help read the distribution of POS

more in detail. I then provided an overview of the distribution of syntactic relations in each dataset, which offered an insight into the most frequent syntactic functions in each genre. I then focused on predicates, which are the core of dependency-based syntax. I therefore exploited the dependency annotation to investigate different features of predicates and of some of their direct dependents as well. I first considered the overall distribution of nonverbal and verbal predicates, what are the syntactic function in which nonverbal predication is used most, and the distribution of the parts of speech in the role of nonverbal predicates. I then moved to the distribution of the verbal forms, showing a correlation between the textual genres and distribution of finite forms, non-finite forms and past participles. Given the very high frequency of the verbal form *gibt* in all the datasets, I then considered the distribution of existential clauses, which are based on the impersonal use of the verb *geben* in the existential construction *es gibt*. I analysed different aspects of the existential constructions, i.e. the position of the pronoun *es* with respect to the verb, the role of the preverbal entity when the pronoun *es* occupies the postverbal position, as well as the capacity to generate subordination. I then focused on modal verbs, of which I investigated different properties, such as the dependency length, the distribution of forms, among the others. Finally, I focused on the position of the subordinate clauses within the sentence, and I investigated the orientation of those dependency relations encoding subordination.

The treebank-based analysis revealed different hidden features of the language of Fragments, as well as both parallelisms and differences with respect to the other two contemporary genres. In the distribution of parts of speech, I exploited the dependency relations to show how nouns, i.e. the most frequent lexical class in all the datasets (LIT RF nouns = 21.0; GSD RF nouns = 23.9; HDT RF nouns = 23.9), are distributed across different syntactic functions. Overall, I demonstrated how nouns tend to fulfil the role of nominal dependents, i.e. nominal modifiers (nmod) and oblique dependent (obl), more frequently than that of core-arguments, i.e. nominal subject (nsubj) and objects (obj/iobj), in all the genres, but especially in GSD (LIT RF nd = 42.6/RF ca = 33.3; GSD RF nd = 51.3/RF ca = 30.4; HDT RF nd = 48.6/ RF ca = 37.1)³⁰⁸. The most common function in all the datasets for this word class is in fact that of nominal modifier (nmod). This trend was observed in all the three genres, with a rather similar RF (LIT RF nmod = 24.1; GSD RF nmod = 25.4; HDT RF nmod = 23.4). By contrast, Fragments showed a lower RF of nouns in the role of oblique arguments (obl) with respect to both the other two genres (LIT RF obl = 17.7; GSD RF obl = 23.5; HDT RF obl = 25.2). This result is caused by the low necessity to encode locative and temporal information in this genre, which prefers to convey messages with a universal scope, without limiting them to a particular time or context. Also, many nouns have the role of appositions (appos) in both GSD and HDT (GSD RF appos = 2.3; HDT RF appos = 3.6), while this function is much rarer in Fragments (RF appos = 0.9). The RF of nouns in the role of nominal subjects of passive

³⁰⁸ *Nd* stands for nominal dependents, while *ca* stands for core arguments.

verbs (nsubj;pass) is also very low in Fragments (LIT RF nsubj;pass = 0.7), while it is much higher in both the other genres, especially in GSD (GSD RF nsubj;pass = 4.0; HDT nsubj;pass = 3.1). This anticipated a syntactic feature later confirmed by the distribution of dependency relations, i.e. a much lower use of the passive voice in Fragments with respect to the other genres. Conversely, Fragments show a significantly higher use of both adverbiation (LIT RF_{ADV} = 10.2; GSD RF_{ADV} = 5.3; HDT RF_{ADV} = 6.4) and pronominalization (LIT RF_{PRON} = 9.1; GSD RF_{PRON} = 5.5, HDT RF_{PRON} = 4.1) with respect to the other genres. As for adverbs, the high frequency of this word class in Fragments is mainly caused by a high frequency of different types of adverbs, especially *auch* ('also'; RF = 8.4), but also *nur* ('only'; RF = 3.7), *noch* ('again'; RF = 3.3), and *sehr* ('very'; RF = 2.6). Interestingly, none of the most frequent adverbs encodes spatial or time information in this genre. As for pronouns, the most frequent pronominal lemma in Fragments is the personal pronoun *der* (RF = 14.0), followed by the neuter pronoun *es* (RF = 12.5), the reflexive pronoun *sich* (RF = 10.7), the personal pronoun *sie* (RF = 10.2), and the impersonal pronoun *man* (RF = 8.8). Moreover, I observed how pronouns tend to have a very different distribution across functional categories in this genre with respect to nouns. In fact, most of them have the role of core-arguments (RF = 78.0), 54% of which are nominal subject (nsubj) and 20.4% direct objects (obj), while only 15.1% of them have the function of nominal dependents, precisely 6.8% as oblique argument (obl) and only 3.3% as nominal modifiers (nmod).

The distribution of dependency relations confirmed the assumptions made when analysing the distribution of parts of speech, and provided further information concerning the distribution of the syntactic functions in the genres as well. The most frequent dependency relations in all the datasets is the *det* relation, which is clearly correlated to the fact that nouns are the most frequent word class in all the datasets. *case* relation is the second most frequent relation in both GSD and HDT, due to the high frequency of nouns in the role of oblique dependents observed in these datasets, while its frequency decreases in Fragments, in which specification through oblique arguments was demonstrated lower. To this respect, the *obl* relation is obviously much more frequent in both GSD and HDT than in Fragments. The distribution of *deprel* confirmed the high use of adverbs (*advmod*) as well as the low use of passive voice (nsubj;pass) in Fragments. Moreover, this distribution shed light on a much frequent use of both coordination and subordination in Fragments with respect to the other genres. As for subordination, the frequency of both clausal core arguments (*csubj*, *ccomp*, and *xcomp*) and clausal dependents (*advcl*, *acl*) was clearly higher in this genre, in particular the frequency of adverbial clauses (*advcl*) and open clausal complements (*xcomp*). The distribution of the subordination markers (*mark*) also confirmed the role of subordination in this genre. Fragments showed a high RF of auxiliaries (*aux*) as well, which was later demonstrated to be due to a frequent use of modal verbs, and a high RF of *cop* relations, which anticipated the distribution of nonverbal predicates in this genre. In addition, the distribution showed how the Fragments have some cases of *orphan* relation, which was mainly due to cases of *gapping*, while this relation

was not present at all in the other genres. Chart 15 reports the relations whose distribution was more different between the Fragments and the other genres.

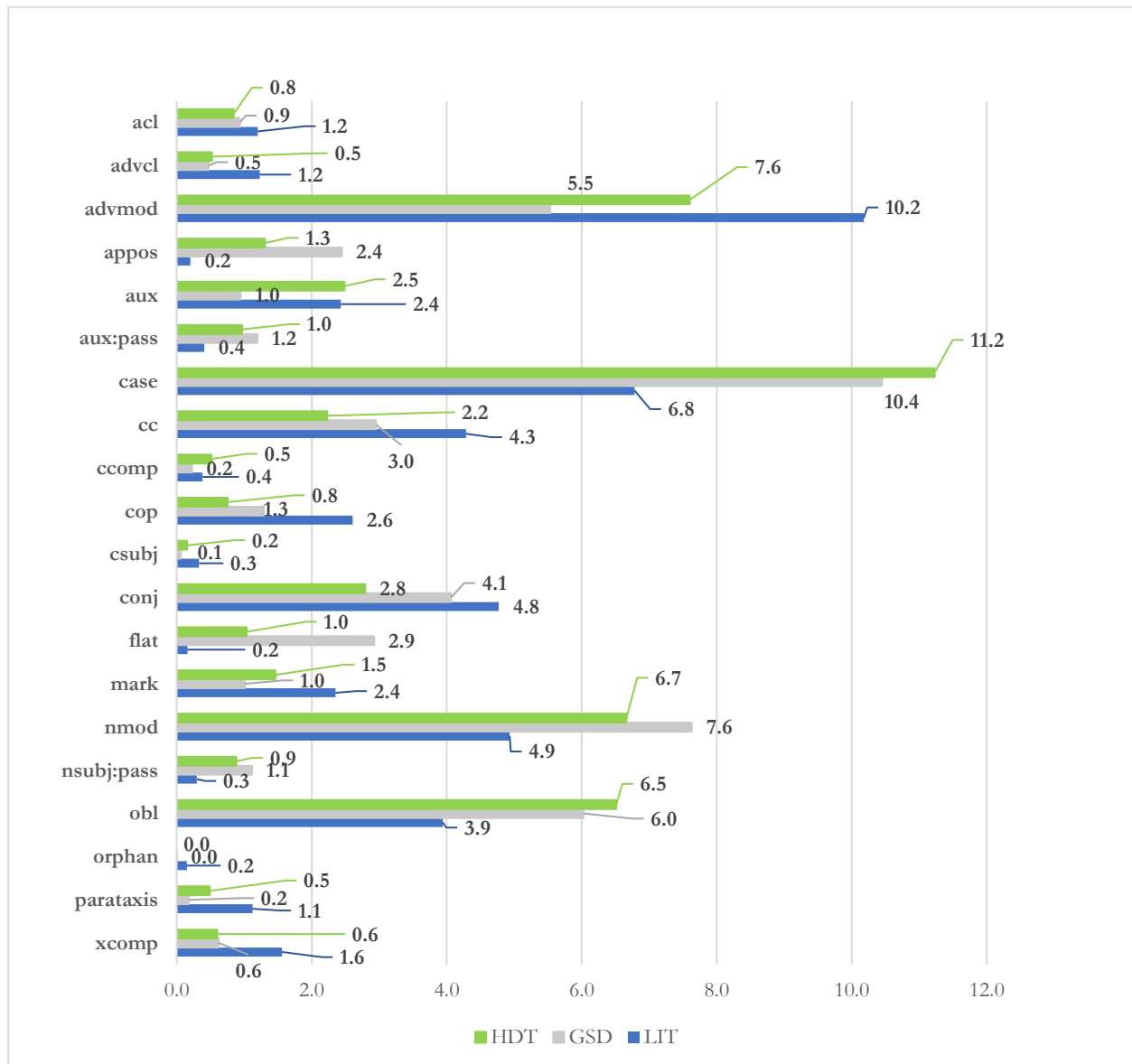


Chart 15 Dependency relations showing the higher different distribution across the three genres.

As for the investigation of predicates, I first considered the distribution of nonverbal and verbal predicates across the data sets. Fragments clearly showed a more frequent use of nonverbal predication with respect to the other genres (LIT RF nvp = 27.1; GSD RF nvp = 18.7; HDT RF nvp = 13.8)³⁰⁹. In

³⁰⁹ *nvp* stands for nonverbal predicates.

this respect, I reported some occurrences of nonverbal predicates in Fragments, and I assumed that the rather frequent use of nonverbal predication was mainly due to two functions: attribution and equation, of which I reported some examples. Both the functions seem to match the communicative purpose of Fragments, which are a judgment-oriented genre. Conversely, the other genres are more informative and explanatory, therefore the use of nonverbal predication appears to be mainly confined to one main function, i.e. inclusion. I then considered the distribution of nonverbal predication across different syntactic functions. In Fragments, nonverbal predication is especially used in the role of open clausal complement (xcomp; 15.5% nouns, 23.8% adjectives), main predicate (root; 23.7% nouns, 11.6% adjectives), and clausal complement (ccomp; 12.2% nouns, 12.8% adjectives). The use decreases in adjectival clauses (acl) and adverbial clauses (advcl), but also in the predicates of coordinate clauses (conj and parataxis). As for the open clausal complements, I demonstrated how the high frequency of this function in the genre is due to a large use of secondary predication (RF = 61.4), and how 60.4% of the secondary predicates embody nonverbal predicates.

The investigation of the verbal forms revealed how Fragments shows a much higher frequency of non-finite forms with respect to the other genres, as well as a lower frequency of past-participle forms. I assumed that this could be especially due to the fact that Fragments tend to show a frequent use of modal verbs, non-finite clauses and, as already stated, nonverbal predicates, especially at present tense. By contrast, I assumed that these phenomena are rarer in the other genres. In addition, I assumed that especially GSD but also HDT are more past-oriented genres, even if in two different ways: GSD's texts prefers to use *Präteritum* (the simple perfective form of the past), while HDT's texts the *Perfekt* (the compound perfective form of the past), which respectively serve two distinct communicative purposes. The first one is mainly used as historical tense, i.e. for narrating past actions in biographical texts, while the second one to report about recently happened past events. The distribution of the most frequent verbal forms in each dataset confirmed these assumptions. Moreover, it overall shed more light on the use of verbs. It showed how the third-person singular is the most frequent form in both LIT and GSD. It showed that the form *gibt*, i.e. the third-person singular of *geben* ("to give"), is the most frequent form in both LIT (RF = 4.0) and GSD (RF = 1.5), and the second-most frequent form in HDT (RF = 1.2). It showed how the most frequent form in HDT is the *Präteritum* speech verb *sagte* (RF = 9), i.e. the third-person singular of *sagen* ("to say"), and therefore how this genre is overall influenced by the reported speech. I then considered the existential clauses, and I investigated the *es-gibt* construction. In fact, 84% of occurrences of the form *gibt* in Fragments are due to this construction, in which the entity is in accusative case, and the pronoun *es* is used as expletive element. In GSD, 77.6% of the occurrences of *gibt* are due to this form. However, I also showed how the annotation of the relation spanning from *gibt* to the entity in accusative case is inconsistency across the three UD German Treebanks, since it is sometimes considered a nominal subject (nsubj), and other times a direct object (obj). I also investigated

the syntax of *es-gibt* constructions. I found that the pronoun *es* occurs in preverbal position most of the times in Fragments (RF pre = 81.1), while it occurs much more frequently in postverbal position in GSD (RF post = 61.9). I showed that this feature of the texts in GSD is mainly due to the necessity to encode specification through oblique modifiers in preverbal position. Finally, I investigated the capacity of the existential clauses to generate complex sentences, i.e. to produce subordinate or coordinate clauses. In this respect, there was a clear difference in the use of existential clauses between LIT and GSD. In the Fragments, 64.2% of them generate subordinate clauses, while 28.3% are used in simple sentences. The most frequent subordinate clause generated by existential constructions in Fragments is the relative clause (RF acl = 56.6%). In GSD, only 23% of the *es-gibt* constructions generate subordinate clauses, while 68.6% of them are used in simple sentences. However, only 7% of the existential constructions generate coordinate clauses in both the datasets.

I then focused on modal verbs. The assumption about the high frequency of modal verbs causing the high distribution of non-finite verbs in Fragments was proved true: in this genre, 28.7% of the tokens tagged as AUX are actually modal verbs, while only 16.2% in GSD and 14.9% in HDT. Modal verbs can cause long-standing relations, since the modal usually occupies the second position in verb-second clauses (main clauses), while the non-finite verb occupies the last position. By contrast, modal verbs generate short relations when used in subordinate clauses, which are conversely verb-final clauses, therefore the modal verb immediately follows the subordinate predicate, in last position. I therefore checked the Average Dependency Length (ADL) of modal verbs in each dataset. It turned out to be rather low in Fragments (3.8 tokens), since 37.3% of modal verbs occur in subordinate clauses in this genre, while ADL increases in the other genres, where the RF of modal verbs in subordinate clauses clearly decreases (GSD ADL = 5.2/RF sub = 29.1%; HDT ADL = 6/RF sub = 12.2%)³¹⁰. I then tested the distribution of modal verbs in each dataset, which showed that the most frequent modal verb (lemma) is *können* in all the datasets (LIT RF = 36.1; GSD RF = 49.2; HDT = 31.4), while the Fragments show a higher frequency of *müssen* with respect to the other genres (LIT RF = 19.1; GSD RF = 16.9; HDT RF = 12.5). I also analysed some real contexts in which the modal verbs are used, showing how the nature of the textual genres and their communicative purposes can influence the semantic scope of the modal verbs. In Fragments, they overall appear to be a useful linguistic mean to encode considerations about the limits and the power of art and artists. Finally, I checked the use of verbal and nonverbal predicates as heads of modal verbs in both Fragments and GSD. In Fragments, the result was that all the modal verbs are mainly used with verbal predicates, especially *können* (93% verbal predicates), even if the use of nonverbal predication clearly increases with *müssen* (19.4% nonverbal predicates) and especially with *sollen* (30% nonverbal predicates).

³¹⁰ *sub* stands for modal verb in subordinate clause.

As last parameter, I investigated the position of subordinate clauses within the sentence. In fact, subordinate clauses can either precede or follow the higher clause. Therefore, their syntactic relation, which spans from the higher predicate to the subordinate predicate, can be right-oriented (the head is on the left) or left-oriented (the head is on the right). I therefore investigated the orientation of all those dependency relations governing subordinate predicates (apart from *acl*). The result was that the position of the subordinate clauses varies a lot across the datasets, but also within the same genre, according to the syntactic function. The orientation of clausal subjects (*csubj*) clearly varies across the genres. As for the Fragments, there is a rather similar distribution between the clausal subjects preceding the higher clause (RF = 57.4), and those following the higher clause (RF = 42.6). Those clausal subjects postponed with respect to the higher clause were demonstrated due to constructions with expletive elements in the higher clause. The frequency of the clausal subjects occurring after the higher clause increases in GSD with respect to Fragments, while their distribution is rather similar in HDT. The orientation of open clausal complements (*xcomp*) also showed a high degree of variation in Fragments: 47.9% of them occur before the higher clause, and 52.1% of them occur after the clause. In detail, 80% of right-to-left open clausal complements, i.e. with the head on the right, are caused by secondary predication, and only 20% by non-finite clauses (with a very low use of final clauses introduced by the marker *um*, also). By contrast, the distribution of functions is rather different for left-to-right open clausal complements, i.e. with the head on the left, since 45% of them are non-finite clauses, while 45% secondary predicates. Overall, Fragments are characterized by a larger use of open clausal complements preceding the higher clause with respect to the other two genres, since the RF of them clearly decreases in both GSD (34.6%) and especially (25.2%). As for adverbial clauses, the distribution of the orientation is rather similar across all the datasets, with almost 70% of adverbial clauses spanning from left to right, i.e. with head on the left, in all the datasets.

In conclusion, I overall attempted to demonstrate the importance of a literary dependency treebank in the wealth of the digital linguistic resources for the German language, especially for the investigation of the literary language. The analysis demonstrated that different hidden features of the language of a literary genre can be detected through a treebank-based approach, and especially through a dependency-based annotation. In this respect, the UD scheme, as well as the query language implemented in the tool SETS, were both demonstrated useful and suitable to extract different levels of linguistic information. In particular, I highlighted how the dependency relations allow to investigate a series of parameters that cannot be studied through traditional corpus-based approaches, in which dependencies are not encoded. We observed how the distribution of parts of speech can be read more in depth thanks to dependencies, which allow to investigate how the lexical classes are distributed in different syntactic roles. Since UD relations encode the function played by the dependent with respect to the head, and they directly attach to content words, the retrieval of the syntactic functions of different content words was immediate. The

results offered very interesting insights into the relation between lexical classes and syntax, as we observed about the distribution of nouns and pronouns in different functional categories. Also, we observed how the overall distribution of dependency relations can shed light on some macroscopic syntactic phenomena, for instance the use of coordination and subordination in a genre, as well as the use of the passive voice, to name but a few. In this case too, these phenomena can be easily investigated thanks to extraction of the dependency labels from the datasets. Moreover, I highlighted how the dependency annotation is the stepping stone for in-depth investigations on different aspects of predicates, which are the core of this syntactic formalism. For instance, the study of the use of nonverbal predication, the study of the position of subordinate clauses in a sentence, as well as of some properties of the modal verbs, would be hardly doable without any syntactic annotation, but especially without any dependency annotation. In this regard, the analysis that I here performed on predicates, and the issues that were here posed, can help lay the foundation for the development of a predicate-centered empirical analysis of the literary language, which is a totally unexplored area in the field of linguistic approaches to literature. For sure, the whole analysis that I here conducted cannot be considered a comprehensive stylistic dependency-based analysis of the language of the Fragments. Such an analysis would require many more parameters to be considered, and it should also be conducted with respect to a textual genre from the same age of the Fragments. In any case, the choice of data to compare should also move from very well-grounded literary criteria. Such an analysis was not possible within the scope of this thesis. However, this thesis was the necessary first step toward possible future investigations, which can take advantage of the results produced by this project. For instance, the semi-automatic syntactic annotation (with the UD scheme version 2) of the *Letters Upon The Aesthetic Education of Man* by Friedrich Schiller is under way. They were published in the same period of Fragments, and they perfectly embodied the features of the classical prose, to which Fragments deliberately and fiercely opposed their concise style. F. Schiller and F. Schlegel are widely considered two great literary opponents, since they embodied the values of the two clashing movements of their age, i.e. Classicism and early Romanticism, respectively. A treebank-based linguistic comparison of the Fragments against the Letters by Schiller would be therefore more than well-grounded, both linguistically and in terms of literary criticism. Such a study would surely dig up a lot of hidden, and still unexplored, differences (or even similarities) between these two literary languages, which are both milestones of the German literary history. As I said, such an analysis could not fall within the scope of this thesis. However, this thesis aimed to move the first necessary steps to pave the way toward this goal, which can be pursued in future work. By and large, I hopefully demonstrated how a dependency treebank can be a tool to empirically find new answers to old questions concerning the raw material of

which literature is made, which is maybe the most enigmatic form of that incredible product of the human mind³¹¹, which we call language.

³¹¹ See also (Van Peer, Hakemulder, and Zyngier 2012).

6 References

- Abeillé, Anne. 2012. *Treebanks: Building and Using Parsed Corpora*. Vol. 20. Springer Science & Business Media.
- Albert, Stefanie, Jan Anderssen, Regine Bader, Stephanie Becker, Tobias Bracht, Sabine Brants, Thorsten Brants, Vera Demberg, Stefanie Dipper, and Peter Eisenberg. 2003. 'Tiger Annotationsschema'. *Universität Des Saarlandes and Universität Stuttgart and Universität Potsdam*.
- Andrews, Avery. 2007. 'The Major Functions of the Noun Phrase'. In Timothy Shopen (ed.), *Language Typology and Syntactic Description: Clause Structure (2nd Ed)*. Cambridge: Cambridge University Press.
- Antomo, Mailin, and Markus Steinbach. 2010. 'Desintegration Und Interpretation: Weil-V2-Sätze an Der Schnittstelle Zwischen Syntax, Semantik Und Pragmatik'. *Zeitschrift Für Sprachwissenschaft* 29 (1): 1–37.
- Ballesteros, Miguel, and Joakim Nivre. n.d. 'MaltOptimizer: An Optimization Tool for MaltParser', 5.
- Baroni, Marco, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. 'The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora'. *Language Resources and Evaluation* 43 (3): 209–26.
- Behler, Ernst. 2011. *Frühromantik*. Vol. 2807. Berlin/New York: Walter de Gruyter.
- Besch, Werner, Anne Betten, Oskar Reichmann, and Stefan Sonderegger. 1998. *Sprachgeschichte: Ein Handbuch Zur Geschichte Der Deutschen Sprache Und Ihrer Erforschung*. Vol. 1. Berlin / New York: Walter de Gruyter.
- Biber, Douglas. 2010. 'What Can a Corpus Tell Us about Registers and Genres'. In Anne O'Keeffe, Michael McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics*. Abingdon and New York: Routledge, 241–54.
- Biber, Douglas E. 2012. 'Corpus-Based and Corpus-Driven Analyses of Language Variation and Use'. In Bernd Heine, Heiko Narrog (eds.), *The Oxford Handbook of Linguistic Analysis*. Oxford: Oxford University Press.
- Blanchot, Maurice, Deborah Esch, and Ian Balfour. 1983. 'The Athenaeum'. *Studies in Romanticism*, 163–72.
- Bohnet, Bernd. 2010. 'Very High Accuracy and Fast Dependency Parsing Is Not a Contradiction', In *Proceedings of the 23rd international conference on computational linguistics (coling 2010)*, 9.
- Bohnet, Bernd, Miguel Ballesteros, Ryan McDonald, and Joakim Nivre. 2016. 'Static and Dynamic Feature Selection in Morphosyntactic Analyzers'. *ArXiv Preprint ArXiv:1603.06503*. <http://arxiv.org/abs/1603.06503>.
- Bohnet, Bernd, and Joakim Nivre. 2012. 'A Transition-Based System for Joint Part-of-Speech Tagging and Labeled Non-Projective Dependency Parsing'. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 1455–1465. Jeju Island, Korea: Association for Computational Linguistics. <http://www.aclweb.org/anthology/D12-1133>.

- Brants, Sabine, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. 'The TIGER Treebank'. In *Proceedings of the workshop on treebanks and linguistic theories*, Vol. 168.
- Buchholz, Sabine, and Erwin Marsi. 2006a. 'CoNLL-X Shared Task on Multilingual Dependency Parsing'. In *Proceedings of the tenth conference on computational natural language learning (CoNLL-X)*, 149–64. Association for Computational Linguistics.
- Buckland, Michael, and Fredric Gey. 1994. 'The Relationship between Recall and Precision'. *Journal of the American Society for Information Science* 45 (1): 12–19.
- Collins, Michael. 2002. 'Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms'. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, 1–8. Association for Computational Linguistics. <https://doi.org/10.3115/1118693.1118694>.
- Cortes, Corinna, and Vladimir Vapnik. 1995. 'Support-Vector Networks'. *Machine Learning* 20 (3): 273–97. <https://doi.org/10.1007/BF00994018>.
- Crammer, Koby, and Yoram Singer. 2001. 'Ultraconservative Online Algorithms for Multiclass Problems'. In David Helmbold, Bob Williamson (eds.), *Computational Learning Theory*, 2111:99–115. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-44581-1_7.
- De Marneffe, Marie-Catherine, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D Manning. 2014. 'Universal Stanford Dependencies: A Cross-Linguistic Typology.' In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 14:4585–92.
- De Marneffe, Marie-Catherine, and Christopher D Manning. 2008. 'The Stanford Typed Dependencies Representation'. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (Vol. 14)*, 1–8. Association for Computational Linguistics.
- Demske, Ulrike. 2007. 'Das MERCURIUS-Projekt: Eine Baubank Für Das Frühneuhochdeutsche'. *Sprachkorpora: Datenmengen Und Erkenntnisfortschritt*, 91–104.
- Dipper, Sefanie. 2010. 'POS-Tagging of Historical Language Data: First Experiments'. In Manfred Pinkal (ed.), *Semantic Approaches in Natural Language Processing: Proceedings of the Conference on Natural Language Processing 2010*. Saarbrücken: Universaar.
- Dipper, Stefanie, and Sandra Kübler. 2017. 'German Treebanks: TIGER and TüBa-D/Z'. In Nancy Ide, James Pustejovsky (eds.), *Handbook of Linguistic Annotation*, 595–639. Springer.
- Dipper, Stefanie, Anke Lüdeling, and Marc Reznicek. 2013. 'NoSta-D: A Corpus of German Non-Standard Varieties'. *Non-Standard Data Sources in Corpus-Based Research* 5: 69–76.
- Dixon, Robert MW. 2012. 'Basic Linguistic Theory, Vol. 3: Further Grammatical Topics'.
- Droganova, Kira, and Daniel Zeman. 2017. 'Elliptic Constructions: Spotting Patterns in UD Treebanks'. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, 48–57.
- Dryer, Matthew S. 2007. 'Clause Types'. *Language Typology and Syntactic Description* 1: 224–75.
- Foth, Kilian, Arne Köhn, Niels Beuck, and Wolfgang Menzel. 2014. 'Because Size Does Matter: The Hamburg Dependency Treebank' In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, 2326–2333.

- Gaumann, Ulrike. 1983. *Weil Die Machen Jetzt Bald Zu': Angabe-Und Junktivsatz in Der Deutschen Gegenwartssprache*. Kümmerle.
- Geyken, Alexander. 2013. 'Wege Zu Einem Historischen Referenzkorpus Des Deutschen: Das Projekt Deutsches Textarchiv', In Ingelore Hafemann (ed.), *Perspektiven einer corpusbasierten historischen Linguistik und Philologie. Internationale Tagung des Akademienvorhabens „Altägyptisches Wörterbuch“ an der BBAW* (Thesaurus Linguae Aegyptiae ; 4). Berlin, S. 221-234.
- Gildea, Daniel. 2001. 'Corpus Variation and Parser Performance'. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*.
- Hajič, Jan, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, and Jan Štěpánek. 2009. 'The CoNLL-2009 Shared Task: Syntactic and Semantic Dependencies in Multiple Languages'. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, 1–18. Association for Computational Linguistics.
- Hirschmann, Hagen, and Sonja Linde. 2011. 'Annotationsguidelines Zur Deutschen Diachronen Baumbank'. Tech. rep. Humboldt-Universität zu Berlin.
- Horsmann, Tobias, and Torsten Zesch. 2016. 'Assigning Fine-Grained PoS Tags Based on High-Precision Coarse-Grained Tagging'. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 328–336. Osaka, Japan: The COLING 2016 Organizing Committee. <http://aclweb.org/anthology/C16-1032>.
- Ide, Nancy, Nicoletta Calzolari, Judith Ecker-Köhler, Dafydd Gibbon, Sebastian Hellmann, Kiyong Lee, Joakim Nivre, and Laurent Romary. 2017. 'Community Standards for Linguistically-Annotated Resources'. In Nancy Ide, James Pustejovsky (eds.), *Handbook of Linguistic Annotation*, 113–65. Springer.
- Ide, Nancy, and James Pustejovsky. 2017. *Handbook of Linguistic Annotation*. Springer.
- Imrényi, András, and Nicolas Mazziotta. 2020. *Chapters of Dependency Grammar: A Historical Survey from Antiquity to Tesnière*. Vol. 212. John Benjamins Publishing Company.
- Jurafsky, Dan. 2000. *Speech & Language Processing*. Pearson Education India.
- Jurafsky, Dan, and James H Martin. 2014. *Speech and Language Processing*. Vol. 3. Pearson London.
- Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubiček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. 'The Sketch Engine: Ten Years On'. *Lexicography* 1 (1): 7–36.
- Kübler, Sandra, Ryan McDonald, and Joakim Nivre. 2009. 'Dependency Parsing'. *Synthesis Lectures on Human Language Technologies* 1 (1): 1–127.
- Kübler, Sandra, Jelena Prokic, and Rijksuniversiteit Groningen. 2006. 'Why Is German Dependency Parsing More Reliable than Constituent Parsing'. In *Proceedings of the Fifth International Workshop on Treebanks and Linguistic Theories*, Prague, Czech Republic, December.
- Kübler, Sandra, and Heike Zinsmeister. 2015. *Corpus Linguistics and Linguistically Annotated Corpora*. Bloomsbury Publishing.
- Maier, Wolfgang, Sandra Kübler, Daniel Dakota, and Daniel Whyatt. 2014. 'Parsing German: How Much Morphology Do We Need?' In *Proceedings of the First Joint Workshop on Statistical Parsing of*

- Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages*, 1–14. Dublin, Ireland: Dublin City University. <http://www.aclweb.org/anthology/W14-6101>.
- Marcus, Solomon. 1965. 'Sur La Notion de Projectivité'. *Zeitschr. f. Math. Logik Und Grundlagen d. Math.*, 11: 181–92.
- McDonald, Ryan, Joakim Nivre, Yvonne Quirnbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, and T Oscar. 2013a. 'Universal Dependency Annotation for Multilingual Parsing'. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 92–97.
- Mel'čuk, Igor. 2009. 'Dependency in Natural Language'. *Dependency in Linguistic Description* 111: 1.
- Mel'čuk, Igor Aleksandrovič. 1988. *Dependency Syntax: Theory and Practice*. SUNY press.
- Mukherjee, Atreyee, Sandra Kübler, and Matthias Scheutz. 2017. 'Creating POS Tagging and Dependency Parsing Experts via Topic Modeling'. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1*, 347–55.
- Nivre, Joakim. 2005. 'Dependency Grammar and Dependency Parsing'. *MSI Report 5133 (1959)*: 1–32.
- Nivre, Joakim. 2009. 'Treebanks'. In Anke Lüdeling, Merja Kytö, *Corpus Linguistics. An International Handbook.*, Vol. 1. Handbooks of Linguistics and Communication Science. Berlin, Germany: Mouton De Gruyter.
- Nivre, Joakim, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Gabrielė Aleksandravičiūtė, Lene Antonsen, Katya Aplonova, et al. 2019. 'Universal Dependencies 2.4'. <http://universaldependencies.org/>, May. <https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2988>.
- Nivre, Joakim, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, and Natalia Silveira. 2016. 'Universal Dependencies v1: A Multilingual Treebank Collection'. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 1659–66.
- Nivre, Joakim, Johan Hall, and Jens Nilsson. n.d. 'MaltParser: A Data-Driven Parser-Generator for Dependency Parsing', In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, 4.
- Öhl, Peter. 2009. 'Anne Breitbarth, Live Fast, Die Young. The Short Life of the Early Modern German Auxiliary Ellipsis'. *Beiträge Zur Geschichte Der Deutschen Sprache Und Literatur (PBB)* 131 (1): 145–52.
- O'Keefe, Anne, and Michael McCarthy. 2010. *The Routledge Handbook of Corpus Linguistics*. Routledge.
- Paroubek, Patrick. 2007. 'Evaluating Part-of-Speech Tagging and Parsing Patrick Paroubek'. In Laila Dybkjær, Holmer Hemsén, Wolfgang Minker (eds.), *Evaluation of Text and Speech Systems*, 99–124. Springer.
- Payne, Thomas E. 1997. *Describing Morphosyntax: A Guide for Field Linguists*. Cambridge: Cambridge University Press.
- Petran, Florian, Marcel Bollmann, Stefanie Dipper, and Thomas Klein. 2016. 'ReM: A Reference Corpus of Middle High German—Corpus Compilation, Annotation, and Access'. *Corpora and Resources for (Historical) Low Resource Languages*, 1.

- Petrov, Slav, Dipanjan Das, and Ryan McDonald. 2012. 'A Universal Part-of-Speech Tagset'. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*.
- Petrov, Slav, and Ryan McDonald. 2012. 'Overview of the 2012 Shared Task on Parsing the Web'.
- Pittner, Karin, and Judith Berman. 2015. *Deutsche Syntax: Ein Arbeitsbuch*. Narr Francke Attempto Verlag.
- Robinson, Jane J. 1970. 'Dependency Structures and Transformational Rules'. *Language*, 259–85.
- Rothstein, Susan Deborah. 2013. *Secondary Predication and Aspectual Structure*. Universitätsbibliothek Johann Christian Senckenberg.
- Roy, Isabelle. 2013. *Nonverbal Predication: Copular Sentences at the Syntax-Semantics Interface*. Oxford University Press.
- Sag, Ivan. 1976. 'Jorge Hankamer Deep and Surface Anaphora'. *Linguistic Inquiry* 7 (3): 391–428.
- Salomoni, Alessio. 2017a. 'Dependency Parsing on Late-18th-Century German Aesthetic Writings: A Preliminary Inquiry into Schiller and F. Schlegel'. In *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage*, 47–52. ACM.
- Sampson, Geoffrey. 2003. 'Thoughts on Two Decades of Drawing Trees'. In *Treebanks*, 23–41. Springer.
- Scheible, Silke, Sabine Schulte im Walde, Marion Weller, and Max Kisselew. 2013. 'A Compact but Linguistically Detailed Database for German Verb Subcategorisation Relying on Dependency Parses from Web Corpora: Tool, Guidelines and Resource'. In .
- Scherer, Wilhelm. 2014. *Geschichte Der Deutschen Literatur*. BoD–Books on Demand.
- Schiller, Anne, Simone Teufel, and Christine Thielen. 1995. 'Guidelines Für Das Tagging Deutscher Textcorpora Mit STTS'. *Manuscript, Universities of Stuttgart and Tübingen* 66.
- Schlegel, Friedrich. 1971. 'Lucinde and the Fragments'. *Trans. Peter Firchow. Minneapolis: U of Minnesota P*.
- Schuster, Sebastian, Matthew Lamm, and Christopher D Manning. 2017. 'Gapping Constructions in Universal Dependencies V2'. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, 123–32.
- Silveira, Natalia G. 2016. 'Designing Syntactic Representations for NLP: An Empirical Investigation', (Doctoral dissertation, PhD Thesis. Stanford University).
- Štěpánek, Jan, and Petr Pajas. 2010. 'Querying Diverse Treebanks in a Uniform Way'. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, 1828–35.
- Stoljar, Margaret Mahony. 1997. *Novalis: Philosophical Writings*. SUNY Press.
- Straka, Milan, and Jana Straková. 2017. 'Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe'. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 88–99. Vancouver, Canada: Association for Computational Linguistics. <https://doi.org/10.18653/v1/K17-3009>.
- Straková, Jana, Milan Straka, and Jan Hajič. 2014. 'Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition'. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 13–18. Baltimore, Maryland: Association for Computational Linguistics. <http://www.aclweb.org/anthology/P14-5003>.

- Telljohann, Heike, Erhard Hinrichs, Kübler, Sandra. 2004. 'The TüBa-D/Z Treebank: Annotating German with a Context-Free Backbone'. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*.
- Tesnière, Lucien. 1959. 'Eléments de Syntaxe Structurale'.
- Thompson, Sandra A. 1997. 'Discourse Motivations for the Core-Oblique Distinction as a Language Universal'. *Directions in Functional Linguistics* 36: 59–82.
- Toutanova, Kristina, and Christopher D. Manning. 2000. 'Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger'. In *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics* -, 13:63–70. Hong Kong: Association for Computational Linguistics. <https://doi.org/10.3115/1117794.1117802>.
- Van Peer, Willie, Jemeljan Hakemulder, and Sonia Zyngier. 2012. *Scientific Methods for the Humanities*. Benjamins Amsterdam.
- Van Rijsbergen, Cornelis Joost. 1974. 'Foundation of Evaluation?'. *Journal of Documentation* 30 (4): 365–73.
- Völker, Emanuel Borges, Maximilian Wendt, Felix Hennig, and Arne Köhn. 2019. 'HDT-UD: A Very Large Universal Dependencies Treebank for German'. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, 46–57.
- Wermke, Matthias, Kathrin Kunkel-Razum, and Werner Scholze-Stubenrecht. 2005. 'Duden: Die Grammatik [Duden: The Grammar]'.
- Zeman, Daniel. 2008. 'Reusable Tagset Conversion Using Tagset Drivers.' In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, 2008:28–30.
- Zeman, Daniel. 2017. 'Core Arguments in Universal Dependencies'. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*. Association for Computational Linguistics. <http://aclweb.org/anthology/W/W17/W17-6532>.
- Zeman, Daniel. 2018. *The World of Tokens, Tags and Trees*. Ústav formální a aplikované lingvistiky.
- Zeman, Daniel, and Philip Resnik. 2008. 'Cross-Language Parser Adaptation between Related Languages'. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*.
- Zeman, Daniel; Nivre, Joakim; Abrams, Mitchell; et al., 2019, Universal Dependencies 2.5, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, <http://hdl.handle.net/11234/1-3105>.
- Zeman, Daniel; Nivre, Joakim; Abrams, Mitchell; et al. 2020. Universal Dependencies 2.6, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, <http://hdl.handle.net/11234/1-3226>.

