# Automatic Topological Field Identification in (Historical) German Texts

**Katrin Ortmann**
Department of Linguistics
Fakultät für Philologie
Ruhr-Universität Bochum
ortmann@linguistics.rub.de

## Abstract

For the study of certain linguistic phenomena and their development over time, large amounts of textual data must be enriched with relevant annotations. Since the manual creation of such annotations requires a lot of effort, automating the process with NLP methods would be convenient. But the required amounts of training data are usually not available for non-standard or historical language. The present study investigates whether models trained on modern newspaper text can be used to automatically identify topological fields, i.e. syntactic structures, in different modern and historical German texts. The evaluation shows that, in general, it is possible to transfer a parser model to other registers or time periods with overall $F_1$-scores $>92\%$. However, an error analysis makes clear that additional rules and domain-specific training data would be beneficial if sentence structures differ significantly from the training data, e.g. in the case of Early New High German.

## 1 Introduction

To study the development of language over time, sufficient amounts of textual data from different time periods need to be enriched with linguistic annotations. For example, to investigate the historical development of certain syntactic phenomena like extraposition or object order in the middle field of the German sentence, annotated corpora from all relevant language stages, e.g. Middle High German, Early New High German, and modern German, would be needed. However, since the creation of annotations requires a lot of manual effort, historical corpora are rarely annotated with linguistic information beyond the morpho-syntactic level like sentence or clause structure. This limits investigations of syntactic change to qualitative studies on small data sets, often with limited statistical significance. Complementing the manual approaches with quantitative studies on large amounts of annotated texts could validate their results as well as unveil new patterns in the data. To reduce the annotation effort required for the application of quantitative methods, there is a growing interest in the use of NLP methods to automate the annotation task. But the necessary amounts of training data usually do not exist for non-standard or historical language. The present study investigates whether modern newspaper training data can be used to automatically identify topological fields, i.e. syntactic structures, in various modern and historical German texts.

The remainder of this paper is structured as follows: Section 2 covers the theoretical background of the study and gives a short introduction to the topological field model before Section 3 summarizes previous approaches to the automatic identification of topological fields. Section 4 describes the data sets used in this study and Section 5 explains the selected approach for the automatic topological field identification. In Section 6 the evaluation results are presented, including a detailed error analysis, followed by a conclusion in Section 7.

## 2 Topological Field Model

The topological field model (Höhle, 2019) is a widely used theory-neutral framework for the description of syntactic structures in German sentences. While German is considered to have a relatively free word order, the topological fields provide a clear structure on the clause level. In German, there are three different clause

*Proceedings of LaTeCH-CLfL 2020*, pages 10–18
Barcelona, Spain (Online), December 12, 2020.

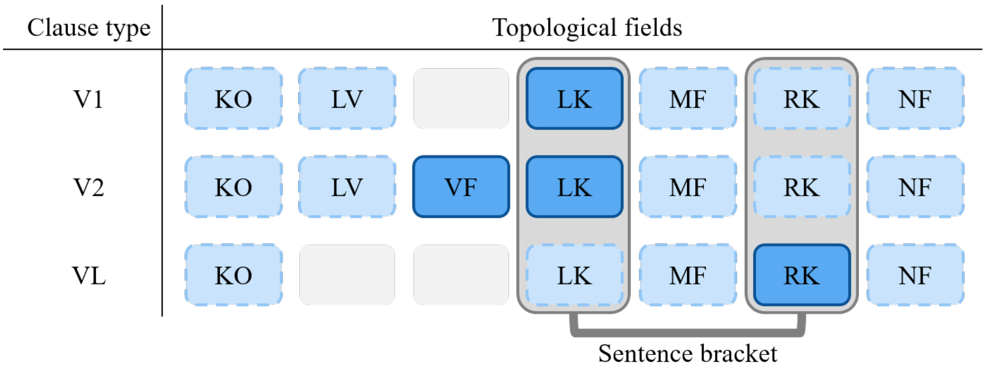


Figure 1: Simplified topological field model for verb-first (V1), verb-second (V2), and verb-last (VL) clauses with mandatory (*blue*) and optional fields (*light blue*, dashed lines). Positions that are never occupied are colored in *light gray*.

types, which are characterized by the position of the finite verb. Figure 1 illustrates the linear order of fields for verb-first (V1), verb-second (V2), and verb-last (VL) clauses. In the present study, a simplified version of the annotation scheme suggested by Telljohann et al. (2015) is used. The following fields are considered:

**VF** The pre-field (*Vorfeld*) of the sentence is obligatory in V2 clauses and always consists of exactly one constituent. Often this is the subject, but it can also be almost any other, possibly complex constituent, e.g. conditional clauses.

**LK** The left sentence bracket (*Linke Klammer*) is obligatory in V1 and V2 clauses and optional in VL clauses. In V1 and V2 clauses, it contains a single finite verb, whereas in VL clauses the position can, instead, be filled with a complementizer and, hence, is often also referred to as C. Following Telljohann et al. (2015), it can be occupied by subordinating conjunctions and relative and interrogative pronouns or phrases.

**MF** The middle field (*Mittelfeld*) is surrounded by the LK to the left and/or the RK to the right and can contain any number of constituents.

**RK** The right sentence bracket (*Rechte Klammer*) is also often referred to as verb complex VC (Telljohann et al., 2015). It contains the non-finite verbs, verb particles, and in VL clauses also the finite verb.

**NF** The post-field (*Nachfeld*) is located to the right of the (possibly empty) RK and can contain any number of constituents. While it is the default position for certain types of subclauses, it often also comprises other ‘heavy’ elements like relative clauses that are extraposed from the middle field.

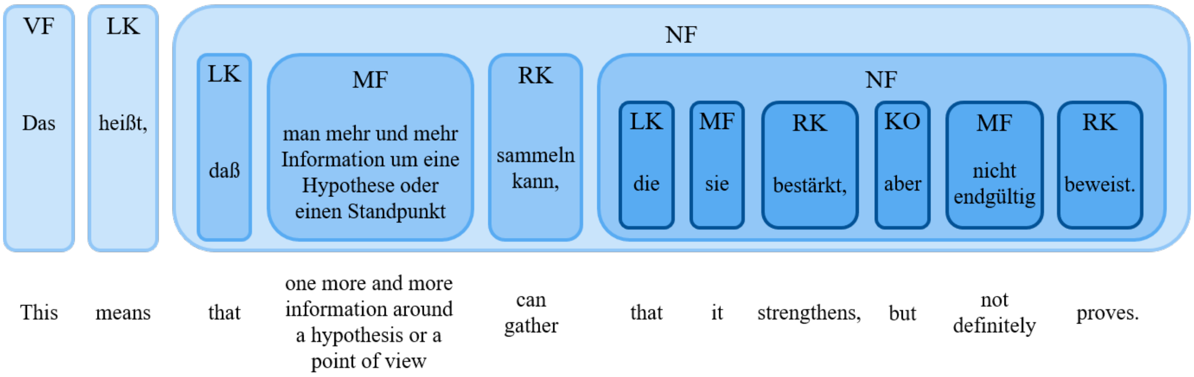
**KO** The coordination field (*Koordinationsfeld*) subsumes the KOORD and PARORD fields from Telljohann et al. (2015) and contains all conjunctions that coordinate sentences, clauses, or fields. The conjuncts themselves are not evaluated here.

**LV** Left dislocations (*Linksversetzung*) contain material that is moved in front of the pre-field.

Except for the sentence brackets and the coordination field, all fields may contain embedded clauses. Figure 2 shows an example annotation with nested topological fields from the data set of this study.

## 3 Related Work

There has been a number of different approaches to the automatic identification of topological fields in German. The first studies (Neumann et al., 2000; Müller and Ule, 2002; Hinrichs et al., 2002) used rule-based approaches, implemented with finite-state cascades, to identify the sentence brackets and, based on this, the other topological fields. For this rule-based approach, Neumann et al. (2000) report an overall $F_1$-score of about 87%. Veenstra et al. (2002) show that for sentence brackets, i.e. fields that contain a very restricted



*'This means that one can gather more and more information around a hypothesis or a point of view that strengthens it, but does not definitely prove it.'*

Figure 2: Example sentence from the present study with nested topological fields.

set of elements, such rule-based systems can yield competitive results. For the identification of more complex topological fields and embedded clauses, though, using (probabilistic) parsers seems more promising: Becker and Frank (2002) train a non-lexicalized chart parser on a probabilistic context-free grammar and achieve labeled recall and precision values of about 93%. Klatt (2004) describes a bi-directional bottom-up parsing approach for non-recursive topological field recognition, resulting in an overall $F_1$-score of about 95%. de Kok and Hinrichs (2016) treat topological field annotation as a sequence labeling task. They use a bi-directional LSTM and achieve an overall accuracy of 97% for non-recursive topological field identification. For recursive topological field annotation, Cheung and Penn (2009) apply the Berkeley parser (Petrov et al., 2006) and report an $F_1$-score of 95% on the Tüba-D/Z corpus and 91% on the NEGRA corpus. They observe the best results for sentence brackets with $F_1$-scores >98%. $F_1$-scores of about 95% or more are also achieved for coordinations and the pre- and middle field. The post-field is recognized less reliably with about 83% and left dislocation with only 7%. All of these approaches focus on standard German (newspaper) text.

To date, there has only been one attempt to automatically identify topological fields in historical data. Using CoNLL-RDF and SPARQL, Chiarcos et al. (2018) implement a deterministic rule-based parser for topological field identification in Middle High German. It relies on grammars and expert knowledge and makes use of the manual annotations provided in the Reference Corpus of Middle High German (ReM). However, in the absence of a manual gold standard annotation, the accuracy of the parser is not evaluated and thus remains unclear.

## 4 Data

Although the topological field model is widely used for the description of German syntactic structures, only few corpora actually provide topological field annotations. The Tüba-D/Z corpus (Telljohann et al., 2015)[1] is the largest available data set, consisting of 3,816 German newspaper articles that are manually annotated with POS tags and topological fields. Discounting headlines and other fragments, which do not receive a topological field annotation, it contains 92,505 sentences with 606,755 fields. For this study, the corpus is split into a training (80%), development (10%), and test set (10%). Most of the studies described in Section 3 use previous versions of this corpus for training and/or evaluation.

To investigate how well the automatic identification of topological fields can be transferred to other domains, the present study includes two additional data sets for modern German. The Tüba-D/S corpus (Hinrichs et al., 2000) consists of 14 spontaneous speech dialogues from a business context, which were manually transcribed and annotated with POS tags and topological fields. Discounting fragments, the data set comprises 19,523 sentences with 107,432 fields. The data set of Ortmann et al. (2019)[2] contains a

[1] Release 11.0 in CoNLL-U v2 format, `http://www.sfs.uni-tuebingen.de/ascl/ressourcen/corpora/tueba-dz.html`

[2] `https://github.com/rubcompling/konvens2019`

| | **Newspaper** | | | **Modern** | | **Historical** | |
|---|---|---|---|---|---|---|---|
| | *Train* | *Dev* | *Test* | *Spoken* | *Written* | *HIPKON* | *DTA* |
| #Docs | 3,075 | 377 | 364 | 14 | 78 | 53 | 29 |
| #Sents | 73,884 | 9,345 | 9,276 | 19,523 | 462 | 342 | 414 |
| #Toks | 1,534,476 | 190,794 | 192,156 | 263,303 | 7,224 | 4,210 | 16,251 |
| **Fields** | | | | | | | |
| KO | 11,195 | 1,521 | 1,458 | 3,274 | 123 | 66 | 252 |
| LV | 1,080 | 159 | 138 | 477 | 41 | 20 | 64 |
| VF | 86,923 | 10,804 | 10,875 | 21,982 | 514 | 290 | 441 |
| LK | 130,321 | 16,322 | 16,345 | 31,013 | 819 | 398 | 1,081 |
| MF | 138,756 | 17,390 | 17,449 | 30,375 | 819 | 327 | 1,356 |
| RK | 88,455 | 10,912 | 11,087 | 14,211 | 493 | 406 | 1,156 |
| NF | 35,076 | 4,404 | 4,427 | 6,170 | 245 | 350 | 478 |
| *Total* | 491,806 | 61,512 | 61,779 | 107,502 | 3,054 | 1,857 | 4,828 |

Table 1: Overview of the data sets. Only sentences with a gold standard annotation are considered.

collection of five different written registers: Wikipedia articles, fiction texts, Christian sermons, TED talk subtitles, and movie subtitles. The data is provided with manually annotated POS tags and was enriched with topological fields for this study. Without fragments, it consists of 462 sentences with 3,054 fields.

Besides the modern data, the present study also includes two historical German corpora to assess whether topological fields can be identified automatically in texts from different time periods, without any historical training data available. The HIPKON corpus (Coniglio et al., 2014) contains sermons from the 12th to the 18th century and offers manual annotations for 342 sentences from the entire time span (except 15th century). Because HIPKON was created for the investigation of post-fields, only sentences with a post-field are annotated. For the present study, these sentences were manually enriched with topological fields, yielding a total amount of 1,857 fields. As HIPKON is the only corpus annotated with a custom POS tagset specifically for historical data, for this study it was mapped to the German standard tagset STTS (Schiller et al., 1999). The second historical corpus, the German Text Archive DTA (BBAW, 2019), is provided with automatically generated linguistic annotations, including sentence boundaries and POS tags. For this study, 414 sentences from 29 texts published between 1562 and 1912 were selected and annotated with a total number of 4,828 topological fields. The DTA sample includes texts from a variety of genres: five newspaper texts and three texts each from the genres funeral sermon, language science, medicine, gardening, theology, chemistry, law, and prose. For every genre, the texts were randomly selected from three (five) different centuries. Since the POS tagging and sentence segmentation accuracies in the sample were considered too low to use them as an evaluation basis, POS tags and sentence boundaries were manually corrected during topological field annotation.[3] Table 1 gives an overview of the data used in the study. The manually annotated data sets and additional resources can be found in this paper's repository at `https://github.com/rubcompling/latech2020`.

## 5 Topological Field Identification

The best results for recursive topological field identification are, so far, reported by Cheung and Penn (2009), who apply the unlexicalized latent variable-based Berkeley parser (Petrov et al., 2006)[4] to the identification of topological fields in German newspaper text. In the present study, their approach is transferred to different data sets, including modern non-standard and historical German texts. To train the Berkeley parser, the Tüba-D/Z training data is converted to a treebank format. Only sentences with a gold standard annotation are used for training. To ensure the applicability to different data sets, the model cannot be based on word forms, which differ significantly between modern and historical writings. Instead, the POS tags, which are consistent across data sets, are taken to form the basic text. To meet the required input format of the parser without supplying word forms, the topological field annotations must be modified. A top-level sentence node is added and artificial pre-terminal nodes are inserted where necessary so that each pre-terminal corresponds to exactly one terminal symbol as it would be the case for words and POS tags in a

[3] The POS error rate in the DTA texts from the sample ranges between 1.3% and 15% (avg: 6.3%). The sentence $F_1$-score for the sample lies between 54.1% and 100.0% (avg: 86.7%).

[4] `https://github.com/slavpetrov/berkeleyparser`

| | **Modern** | | | | | | | | | **Historical** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Newspaper* | | | *Spoken* | | | *Written* | | | *HIPKON* | | | *DTA* | | |
| Field | Prec | Rec | $F_1$ | Prec | Rec | $F_1$ | Prec | Rec | $F_1$ | Prec | Rec | $F_1$ | Prec | Rec | $F_1$ |
| KO | 92.69 | 87.79 | 90.17 | 84.94 | 52.77 | 65.10 | 100.00 | 56.10 | 71.88 | 93.94 | 93.94 | 93.94 | 85.63 | 56.75 | 68.26 |
| LV | 68.57 | 53.73 | 60.25 | 22.15 | 15.95 | 18.54 | 60.00 | 7.50 | 13.33 | 0.00 | 0.00 | 0.00 | 69.57 | 26.23 | 38.10 |
| VF | 95.49 | 97.58 | 96.53 | 87.73 | 97.57 | 92.39 | 96.37 | 99.41 | 97.87 | 88.78 | 99.26 | 93.73 | 78.13 | 94.93 | 85.71 |
| LK | 98.78 | 99.75 | 99.27 | 97.81 | 99.00 | 98.40 | 97.95 | 99.88 | 98.90 | 96.00 | 97.96 | 96.97 | 89.87 | 91.83 | 90.84 |
| MF | 94.61 | 97.74 | 96.15 | 89.45 | 98.20 | 93.62 | 95.87 | 99.37 | 97.59 | 85.80 | 97.69 | 91.36 | 74.24 | 92.18 | 82.24 |
| RK | 99.05 | 99.52 | 99.29 | 97.40 | 99.60 | 98.49 | 98.99 | 99.59 | 99.29 | 94.74 | 95.94 | 95.33 | 88.36 | 97.04 | 92.50 |
| NF | 83.11 | 86.09 | 84.57 | 63.26 | 67.18 | 65.16 | 82.95 | 80.00 | 81.45 | 86.22 | 84.59 | 85.40 | 53.22 | 65.75 | 58.75 |
| *Overall* | 95.80 | 97.45 | 96.61 | 90.87 | 95.02 | 92.90 | 96.15 | 95.09 | 95.62 | 90.85 | 93.99 | 92.39 | 80.05 | 88.34 | 83.99 |

Table 2: Evaluation results for all fields and data sets. The numbers for Precision, Recall, and $F_1$-score are given in percent.

standard syntax tree. The parser is trained with default options[5] using the larger topological field tagset of the Tüba-D/Z corpus (Telljohann et al., 2015) during training, which is then mapped to the simple scheme as described in Section 2 for evaluation. Becker and Frank (2002) note that this strategy of training on more fine-grained categories and evaluating on a coarser tagset can improve the accuracy of topological parsing. To run the Java-based Berkeley parser, it is invoked in interactive mode via the command line and always returns the single best parse.

## 6 Evaluation and Results

For the evaluation of the automatic topological field identification, the parser output is compared to the gold standard annotation and labeled precision and recall are calculated. Here, true positives are fields that cover the correct span of tokens and are labeled with the correct field tag. The evaluation only considers the token span covered by a field, independently of possibly intermediate embedded fields. Punctuation at the edge of fields is removed before evaluation. If fields in the parser output have incorrect boundaries or do not exist at all in the gold standard, they are counted as false positives. If a field is present in the gold standard, but there is no corresponding field in the parser output, this is counted as a false negative. Only sentences for which there is a gold annotation are evaluated.

Table 2 gives an overview of the results for all data sets and fields. As could be expected, the parser achieves the best results on the Tüba-D/Z test data, i.e. the type of data it was trained on, with an $F_1$-score of 96.6%. This is comparable to the results of Cheung and Penn (2009), who report an $F_1$-score of 95.2% for a (much smaller) part of the same corpus. For the two other modern data sets, the parser reaches an overall $F_1$-score of 95.6% (written) and 92.9% (spoken). For the historical data, accuracies differ between data sets. While the results for the HIPKON corpus are comparable to the modern spoken data, the overall $F_1$-score for the DTA is much lower with about 84.0%. Like in previous studies, the sentence brackets are annotated with the highest accuracy in all data sets, followed by the pre- and middle fields, while the results for post-fields are worse for all data sets. Left dislocations are recognized even more rarely by the parser. The results for the coordination field vary between data sets, as well as the proportion of sentences the parser can analyze without errors (31%–79%). In general, correctly analyzed sentences are on average shorter and contain fewer fields. For some fields, it makes a difference if they are embedded in other fields or contain embedded fields themselves. For example, post-fields and left dislocations are recognized less often and less accurately if they do not contain other fields. This can be explained by the characteristics of the training data: Post-fields and left dislocations are rare in newspaper texts and mostly contain 'heavy' elements, i.e. longer clauses. Besides those general observations, every corpus poses different challenges to the parser. To better understand the differences between data sets and the causes of errors, in the following, the results for the different corpora are analyzed in more detail and a qualitative error analysis is carried out.

[5] Training options: `java -Xmx1024m -cp BerkeleyParser-1.7.jar edu.berkeley.nlp.PCFGLA.GrammarTrainer -treebank SINGLEFILE -out grammar.gr -path treebank.txt`

**Newspaper** Except for post-fields and left dislocations, all fields in the Tüba-D/Z test data are recognized with $F_1$-scores between 90% and 99%. The sentence brackets are identified with the highest accuracy, followed by the pre- and middle field and coordinations. For all fields (except KO and LV), at least 40% of the false positives have incorrect boundaries but overlap with the gold annotation. This value is highest for the middle field, where 82% of the false positives only have incorrect boundaries. This can, for example, be the case if the right sentence bracket is empty and the parser regards the middle and post-field as a single field, resulting in a false positive middle field and a false negative post-field. In total, four out of five sentences from this data set are analyzed without any error. On average, those sentences are ten words shorter than sentences containing errors. Errors mostly occur with elliptical constructions, fragments, and parenthetical phrases as well as sentence structures that are uncommon in standard written German and therefore rare in the training data. This observation is in alignment with Cheung and Penn (2009), who also identify parentheticals as the main error cause in their study. Further error sources are quotes and reported (direct) speech, as well as left dislocations and post-fields without internal structure. Overall, the newspaper data is annotated with high accuracy, reproducing the results of prior studies.

**Spoken** While the sentence brackets and pre- and middle fields are recognized with $F_1$-scores $>$92%, only two thirds of the coordinations and post-fields are identified correctly in the spoken data. Left dislocations are recognized with an $F_1$-score of only 18.5%. Again, many false positives overlap with the gold standard annotation, especially in the case of pre- (59%) and middle fields (80%), which often erroneously stretch across left dislocations or post-fields, in turn leading to low recall values for the latter fields. Almost two thirds of all sentences in the spoken data set are analyzed without errors. On average, these sentences contain nine words less than incorrect sentences. Errors mostly result from the divergence between spoken and written language structures, for instance incomplete utterances, repeated words, or unrelated clauses and fragments in a single sentence. Still, it can be stated that, despite the differences between written training and spoken test data, the majority of the fields is recognized with fairly high accuracy and, if similar data should be processed automatically, using part of the spoken data as additional training resource could further improve the results for this text type.

**Written** Looking at the modern written data set, the evaluation shows that texts from different registers can be analyzed with comparable accuracy as newspaper data. The parser performs best on the Wikipedia articles ($F_1$: 99%) while for the other registers the $F_1$-score ranges between 94% and 96%. Although the data shows a slightly different distribution of fields with more left dislocations, post-fields, and coordinations, the parser still recognizes most fields with high $F_1$-scores. Also, half of the false positives overlap with the gold standard annotation: More than two thirds of the false middle fields and more than half of the false post-fields only have incorrect boundaries. 58% of the sentences are analyzed completely correctly. For many sentences, missing coordination fields are the only error. Since coordinating conjunctions are not always annotated in the training data, the parser often does not recognize them in the test data, leading to low recall for the KO field. Using simple rules to add missing coordination fields, the recall for the KO field in this data set can be raised from 56% to 97% while keeping the precision at 100%, thus improving the $F_1$-score of this field to 98%. Further common causes of errors are direct and reported speech, especially in sermons and fiction texts, and the higher proportion of left dislocations and post-fields in informal, spoken-like language.

**HIPKON** The results for the first historical corpus are comparable to those of modern spoken data. For most fields, the $F_1$-score is $>$90% and, despite the higher proportion of post-fields resulting from the corpus design, post-fields are analyzed with a higher $F_1$-score in this historical text sample than in the other corpora. For left dislocations, the opposite is true: Although they are more frequent in the data set, no LV field is recognized in the HIPKON sample. Either the corresponding tokens are not analyzed at all or they are analyzed as part of the pre-field, which is also reflected in the high percentage of pre-fields with incorrect boundaries. In general, more than half of all false positives overlap with the gold standard annotation. The proportion is highest for the post-field with 74% and ranges between 38% and 57% for the other fields. About two thirds of the sentences from this data set are analyzed without errors by the parser. While the recall only shows minor changes with respect to the age of the text, the precision decreases for older texts, reflecting their increasing divergence from the modern training data. Common error causes for this data set include empty middle fields like in (1), which are relatively frequent in the HIPKON corpus due to its specific focus on the post-field.

(1) *vn̄ [LK woͤlte] [RK gan] zů ſínem vatt' vnd ſprechē.*
'And wanted to go to his father and speak.'

Adding historical training data or implementing simple rules, in these cases, could prevent the wrong identification of a middle field if, for example, it is preceded by a right bracket or starting with verbal elements. Additional rules could also improve the identification of post-fields, which are often not recognized by the parser. By simply labeling non-analyzed tokens following a post-field or right bracket as post-field, the recall for this field can be increased by six percentage points to over 90%. Another common cause for errors in this historical data set are left brackets like relative adverbs and particles that no longer exist in modern German, e.g. as in (2):

(2) *nach mittē tage [LK do] er hat geſclâfen*
'after the middle of the day where he had slept'

While these tokens were annotated as relative adverbs or particles with the original custom POS tagset, the information about their relative function was lost during conversion to the modern STTS tagset, preventing the parser from identifying them. Since one missing bracket can easily change the complete analysis of a sentence, the explicit marking of these tokens as left brackets results in improvements of all fields from pre- to post-field. If older historical data should be analyzed reliably, available information about the relative function of tokens must somehow be transferred to the modern tagset, e.g. by adding a special tag and corresponding training data or by (mis-)using an existing tag for relativizers. Overall, the evaluation of the HIPKON data shows that, by using the POS tags as input, it is generally possible to transfer a model from modern to historical data although some special adjustments and/or historical training data would be beneficial to further improve the reliability of the automatic analysis.

**DTA** The results for the second historical corpus are substantially worse than for the other data sets. Only the sentence brackets are identified with $F_1$-scores $>$90%, while the other fields range only between 38.1% and 85.7%. Like for the other corpora, the results are worst for left dislocations: Only a quarter of them is recognized, while the rest is mostly skipped by the parser, especially if they do not contain embedded fields. Coordination fields are also often not recognized, but adding the same simple rules as for the modern written data can increase the recall for the KO field from 56.8% to 90.1%, improving the $F_1$-score of this field by 20 percentage points.

Again, half of all false positives result from incorrect field boundaries. Two thirds of the false middle fields and more than half of the false right brackets and post-fields overlap with the corresponding gold standard annotation. But only 30% of all sentences are analyzed without errors. Those sentences are on average 26.5 words shorter and contain on average 6 fewer fields than sentences with one or more errors. This already indicates that the sentences in the DTA are very long and complex. The average sentence length in the sample is 39 words, compared to 19 words in the modern newspaper texts (spoken: 10, written: 14, HIPKON: 12), with a maximum embedding depth of 10 fields, i.e. one field containing nine other nested topological fields, compared to a maximum depth of 6 fields in the newspaper data (spoken: 5, written: 4, HIPKON: 3). Long and complex left dislocations and deeply embedded post-fields are very common in the data set, as well as embedded structures within the middle field, which are infrequent in modern German. Furthermore, the data contains many parenthetical constructions that, even for human annotators, are hard to process and understand.

The often extreme sentence length and complexity and the deep embedding of fields is a typical characteristic of the Early New High German data and not covered by the modern training data, which explains the high amount of errors. While the parser is mostly able to recognize local, clause-internal structures, e.g. left and right brackets surrounding a middle field, it often fails to identify larger structures, especially in complex constructions, e.g. with several embedded post-fields. The different historical use of punctuation further exacerbates the problems, for example with reported speech and parenthetical constructions. The same can be said about the fact that writers during this time period commonly left out right sentence brackets, which makes embedded clauses even harder to recognize and analyze correctly, for example in (3):

(3) *Ob dieſes wol eine lo̊bliche Sache / wodurch vielmal ſolche Seuche abzuhalten [...]: So bezeuget doch die tågliche Erfahrung / daß [...]*
'Although this (is) a laudable thing, whereby often such an epidemic can be prevented [...], daily experience shows that [...]'

Also, similar to the HIPKON corpus, the DTA sample contains many adverbial left brackets that the parser cannot recognize, leading not only to missing left brackets but also to incorrect surrounding fields. Since these error sources become less frequent over time, there is a clear relationship between the age of the text and how well the parser performs: precision and recall both decrease with increasing age of the text, with the effect being stronger for precision. This observation holds for all genres in the sample, except funeral sermons, which are only available for earlier time periods. The highest $F_1$-scores are reached for the most recent newspaper and chemistry texts, the lowest for the oldest texts from the genres of language science, law, and newspaper.

It has to be kept in mind, though, that the texts in this study are already corrected for sentence boundaries and POS tags. Using the original annotations, the results would be even worse, especially for older texts where POS error rates are high. When the parser is supplied with the original POS tags (and gold sentence boundaries for evaluation purposes), the overall $F_1$-score decreases by almost 10 percentage points to 75.6%. For many older texts, there is an even larger reduction in $F_1$-score of 20 or more percentage points. Using the original sentence segmentation can be expected to further reduce the accuracy. While missing sentence boundaries do not necessarily cause problems, the low precision values (avg: 83%) would lead to many incomplete fields crossing sentence boundaries. This highlights the importance of reliable basic annotations like sentence and token boundaries or POS tags.

Overall, the evaluation of this data set shows that texts from the Early New High German period, which were written by skilled writers or scientists like it is the case for the DTA sample, can only unsatisfactorily be analyzed with models purely trained on modern German. While additional rules could certainly improve the automatic field identification to a certain extent, it is unlikely that a parser will be able to reliably analyze such complex sentences without sufficient similar training data.

## 7 Conclusion

The present study has investigated the automatic identification of topological fields in different modern and historical German texts using only modern newspaper text as training data. The evaluation has shown that, in general, transferring a model from modern newspaper data to other registers or time periods is possible. Using the Berkeley parser, different non-standard and spoken modern data sets as well as sermons from the 12th to the 18th century can be analyzed automatically with overall $F_1$-scores $>92\%$. For the most common fields like sentence brackets or the middle field, the accuracy can be considered sufficient for qualitative and quantitative research based on the automatic field identification.

However, additional rules and especially additional training data for specific data sets could be very beneficial if sentence structures or the distribution of fields differ substantially from modern newspaper language. The evaluation has shown that texts from the Early New High German period, in particular, often exhibit such complex structures that they are hard to process even for human annotators. As a result, the parser only reaches an overall $F_1$-score of 84% on the DTA data set. Future work has to unveil whether time- or genre-specific training data can improve these results and enable a reliable identification of all topological fields in various text types from all time periods. Since creating such training resources is effortful and time-consuming, the presented automatic analyses could be used for pre-annotation, subsequently improving their accuracy by adding further training material.

## Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102 (Project C6). I am greatful to the student annotators Anna Maria

Schroeter, Anna Ehlert, Doreen Scholz, Larissa Weber, Sara Klein, and Jennifer Wodrich for the annotations. Also, I would like to thank the anonymous reviewers for their helpful comments.

## References

BBAW. 2019. Deutsches Textarchiv. Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache. Berlin-Brandenburgische Akademie der Wissenschaften; `http://www.deutschestextarchiv.de/`.

Markus Becker and Anette Frank. 2002. A stochastic topological parser for German. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Jackie Chi Kit Cheung and Gerald Penn. 2009. Topological field parsing of German. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, page 64–72, USA. Association for Computational Linguistics.

Christian Chiarcos, Benjamin Kosmehl, Christian Fäth, and Maria Sukhareva. 2018. Analyzing middle high German syntax with RDF and SPARQL. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Marco Coniglio, Karin Donhauser, and Eva Schlachter. 2014. HIPKON: Historisches Predigtenkorpus zum Nachfeld (Version 1.0). Humboldt-Universität zu Berlin. SFB 632 Teilprojekt B4.

Daniël de Kok and Erhard Hinrichs. 2016. Transition-based dependency parsing with topological fields. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–7.

Erhard W. Hinrichs, Julia Bartels, Yasuhiro Kawata, Valia Kordoni, and Heike Telljohann. 2000. The Tübingen Treebanks for Spoken German, English, and Japanese. In Wolfgang Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translation*, pages 550–574. Springer, Berlin.

Erhard W Hinrichs, Sandra Kübler, Frank Henrik Müller, and Tylman Ule. 2002. A hybrid architecture for robust parsing of German. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*.

Tilman N. Höhle. 2019. Topologische Felder. In Stefan Müller, Marga Reis, and Frank Richter, editors, *Beiträge zur deutschen Grammatik: Gesammelte Schriften von Tilman N. Höhle*, pages 7–89. Language Science Press, Berlin.

Stefan Klatt. 2004. Segmenting real-life sentences into topological fields - for better parsing and other nlp tasks. In *KONVENS 2004. 7. Konferenz zur Verarbeitung natürlicher Sprache*. Ernst Buchberger.

Frank Henrik Müller and Tylman Ule. 2002. Annotating topological fields and chunks - and revising POS tags at the same time. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Günter Neumann, Christian Braun, and Jakub Piskorski. 2000. A divide-and-conquer strategy for shallow parsing of German free texts. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, ANLC '00, pages 239–246, USA. Association for Computational Linguistics.

Katrin Ortmann, Adam Roussel, and Stefanie Dipper. 2019. Evaluating Off-the-Shelf NLP Tools for German. In *Proceedings of the Conference on Natural Language Processing (KONVENS)*, pages 212–222.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 433–440.

Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. 1999. *Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset)*. Retrieved from `http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf`.

Heike Telljohann, Erhard W. Hinrichs, Sandra Kübler, Heike Zinsmeister, and Kathrin Beck. 2015. *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z)*. Seminar für Sprachwissenschaft, Universität Tübingen, Tübingen, Germany.

Jorn Veenstra, Frank Henrik Müller, and Tylman Ule. 2002. Topological field chunking for German. In *Proceedings of the 6th conference on Natural language learning - Volume 20*, pages 1–7.