Extracting a PP Attachment Data Set from a German Dependency Treebank Using Topological Fields

Daniël de Kok, Corina Dima, Jianqiang Ma and Erhard Hinrichs

SFB 833 and Seminar für Sprachwissenschaft
University of Tübingen, Germany
{daniel.de-kok,corina.dima,jianqiang.ma,erhard.hinrichs}
@uni-tuebingen.de

Abstract

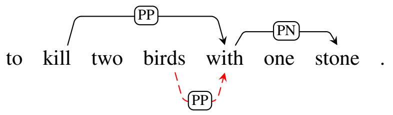
PP-attachment has traditionally been tackled as a binary classification task where a preposition is attached to the immediately preceding noun or to the main verb. In this paper, we provide an analysis of PP-attachment in German to show that the assumption that prepositions have only two head candidates does not hold. We propose a realistic PP-attachment data set, in which each preposition has multiple head candidates. The data set is extracted automatically from a dependency treebank with topological field annotations. Finally, we show that the task of PP-attachment is substantially more difficult with this realistic data set than with a binary classification data set.

1 Introduction

Treebanks are constructed to provide empirical data of language use and their syntactic structure. Apart from being of importance for linguistic research, the availability of treebanks has also opened the possibility to train statistical models that select the most plausible parse of a sentence from the (typically) exponential number of available parses.

Prepositional phrase attachment (PP-attachment) is known to be one of the difficult problems in parse selection [7]. In dependency parsing this problem manifests itself in that the preposition of a PP can have a variety of tokens as its head. This variety concerns both the category and the position of the head. The correct attachment of the preposition is typically dictated by semantics. In the example below the preposition *with* can be attached syntactically either to the verb *kill* or to the noun *birds*. In this case, the verbal attachment is the only one that makes sense semantically. However, the syntactically correct nominal attachment can also provide valuable information about potential, yet semantically incompatible heads.

(1)



A treebank containing gold syntax trees can only provide examples of correct attachments. In order to train a good model for parse selection, examples of bad attachments are necessary as well. Stochastic rule-based parsers solve this problem by producing a parse forest with all the analyses for a sentence that are generated by the grammar and lexicon. The parse selection model is then trained on all parses [1, 9] or a representative sample thereof [12] and learns to discriminate between correct and incorrect attachments. A transition-based dependency parser [11] cannot apply the same strategy: since it is optimized for finding the most plausible reading, it cannot be used to enumerate the total range of syntactically correct variations.

Our goal in this paper is to enrich the training material available to transition-based dependency parsers by approximating the range of candidates for potential PP attachments that rule-based parsers can create.[1] Moreover, our approach is more robust in the presence of out-of-vocabulary words since it does not use fine-grained syntactic analysis. The necessary information, namely the positions in a sentence where the potential heads of a preposition can reside, is already implicitly available in the treebank.

PP attachment disambiguation has generally been framed as a binary decision task [7, 10, 13, 16] where the head of the preposition is either the noun[2] immediately preceding the preposition or the verb. A more realistic setup is to decide on a PP attachment only after considering every potential attachment point in the sentence. While considering all the nouns and verbs in the sentence as possible heads for the preposition has been suggested [5], this approach overestimates the set of candidate heads and fails to rule out candidates that can be eliminated on the basis of structural, syntactic information. An illustration is example (2) from section 2.2, where, for reasons described in detail in the next section, *Goetsch* is a highly unlikely candidate head for the preposition *von*.

In this paper, we report on the creation of a new PP-attachment data set for German that is extracted from the dependency version of the TüBa-D/Z [14, 15] and includes multiple candidate heads for each preposition. We use the topological field model to investigate the distribution of PPs and their heads in order to select only those candidate heads that are plausible competition during parsing, thus approximating the candidate set that a rule-based parser would produce. Although the data set concerns German PP-attachment, the techniques that are presented in this paper are generally applicable for Germanic languages.

[1] One could argue that a rule-based parser should be used to extract all possible attachments. However, hand-crafted grammar rules for wide-coverage parsers are only available for a small number of languages.

[2] In this paper, we use the terms *noun* and *nominal* for nouns and proper names.

2 Analysis of ambiguous PP attachments

In this Section, we will provide an analysis of PP-attachment in terms of the topological field model of German clause structure. This allows us to determine the head candidates in a more fine-grained manner than simply including any noun or verb within the clause or within a certain window of words.

2.1 The topological field model of German

The topological field model can be used to account for regularities in word order across different clause types in German [3, 4, 6, 8]. This model postulates that each clause type has a *left bracket* (*LK*) and a *right bracket* (*RK*), which appear left and right of the *middle field* (*MF*). In verb-second declarative clauses, the LK is preceded by an *initial field* (*VF*), while the RK can optionally be followed by a *final field* (*NF*). Table 1 illustrates the topological field annotation for different types of clauses. As can be seen in these examples, the RK contains the verb cluster. In main clauses the finite verb moves to the LK, whereas in subordinate clauses the LK holds the complementizer(s).

	VF	**LK**	**MF**	**RK**	**NF**
MC:	Gestern	hat	er häufiger	angerufen	als heute
	Yesterday	has	he more-often	called	than today
MC:	Er	ruft	häufig	an	
	He	calls	frequently	up	
SC:		der	noch häufiger	anruft	als er
		who	more often	calls	than him

Table 1: Topological field structure of a main clause with an auxiliary verb, a main clause without an auxiliary/modal verb, and a subordinate clause.

It has been shown that the distribution of the fields, wherein the heads and dependents of a particular dependency relation lie, provides information that can improve dependency parsing [2]. For example, it is very likely that the subject is in the VF of a main clause and highly unlikely that it is in the NF. Therefore, an analysis that attaches an NP in the NF as the subject is probably incorrect. We expect PP-attachment to have similar properties. For example, it seems similarly implausible that a preposition in the NF attaches to a head in the VF. The scope of head candidate selection can thus be constrained by using the distribution of the PP-attachment relation in combination with topological fields. The use of the topological fields model also has practical benefits since topological fields can be predicted accurately using words and part-of-speech tags [2].

Table 2 shows the distribution of head fields of prepositions in the VF, MF, and NF in TüBa-D/Z release 10. Two overarching properties can be observed. (1) Prepositions in all fields can attach to verbs in both brackets. This is due to the fact that prepositions in the dependency version of the TüBa-D/Z are usually attached to the main (non-auxiliary/modal) verb, with one exception that we discuss later. (2) If the head is not in one of the brackets it is most likely to be in the same field as the preposition. In the next section, we will explore the per-field attachment properties of PPs in more detail.

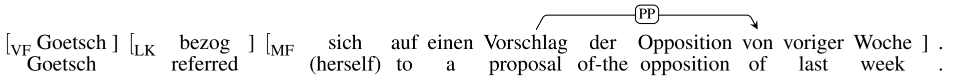
Prep. head field			**Preposition field**		
			VF	MF	NF
	nominal	VF	41.16	0.24	0.57
		MF	1.73	33.47	6.15
		NF	0.00	0.05	35.74
	verbal	LK	55.24	22.19	18.17
		RK	1.87	44.05	39.37

Table 2: Distribution of prepositions and their heads.

2.2 Analysis per topological field

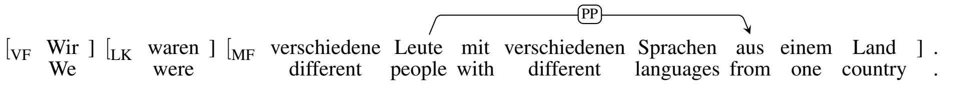
Mittelfeld The distribution of the heads of prepositions in the MF (Table 2) reveals an interesting property: they are very rarely in the VF or NF. More specifically, the verbal candidate head for a MF preposition is in the LK or RK, while the nominal candidate heads are mostly in the MF. The set of candidate heads in the MF can be further restricted, given that German prepositions with nominal heads typically attach leftwards [16].[3] Another assumption that is sometimes made [10, 16] is that only the noun that immediately precedes the preposition is a nominal head candidate. However, in 12.41% of the preposition-noun attachments in the TüBa-D/Z there is another noun between the head and the preposition. Typically, the interspersed word is a genitive modifier (*der Opposition* in Example 2) or a prepositional phrase (*mit verschiedenen Sprachen* in Example 3) that attach to the same head as the preposition. There are, however, also cases when the interspersed material attaches to another head (*auf der Kundgebung* in Example 4).

(2)



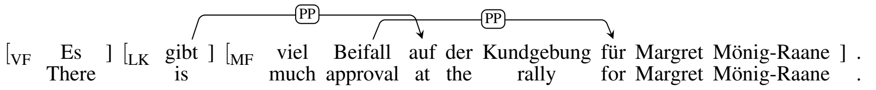
$[_{VF}$ Goetsch]	$[_{LK}$ bezog]	$[_{MF}$ sich	auf	einen	Vorschlag	der	Opposition	von	voriger	Woche]	.
Goetsch	referred	(herself)	to	a	proposal	of-the	opposition	of	last	week	.

(3)



$[_{VF}$ Wir]	$[_{LK}$ waren]	$[_{MF}$ verschiedene	Leute	mit	verschiedenen	Sprachen	aus	einem	Land]	.
We	were	different	people	with	different	languages	from	one	country	.

(4)

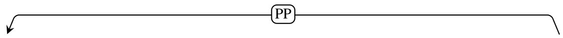


$[_{VF}$ Es]	$[_{LK}$ gibt]	$[_{MF}$ viel	Beifall	auf	der	Kundgebung	für	Margret	Mönig-Raane]	.
There	is	much	approval	at	the	rally	for	Margret	Mönig-Raane	.

As mentioned before, verbal attachments of the preposition always attach to the main verb in the TüBa-D/Z dependency scheme. The main verb is in the LK if the clause is verb-second declarative without a auxiliary/modal verb (Example 2). Otherwise, it is in the RK (Example 5).

[3] In TüBa-D/Z 97.51% of the preposition-noun attachments in the MF are leftward.

Vorfeld Two things stand out in the distribution of heads of VF prepositions in Table 2: (1) in contrast to prepositions in the MF, VF prepositions can have nominal heads that lie outside their field, namely in the MF and (2) the vast majority of verbal heads are in the LK. It should not be surprising that nominal heads can be in the MF, since German permits topicalization of PPs. In Example 5 the PP *Für diese Behauptung* is topicalized. Additionally, this example shows that in the case of noun-attachment, the head is not necessarily the first noun of the MF. Instead, the preposition *für* attaches to the direct object *Nachweis* and not to the first MF noun, *Beckmeyer*. The consequence for candidate extraction is that this generates a lot of candidates since nominal heads of a preposition in topicalized PPs can lie anywhere in the MF.



(5)

[$_{VF}$	Für	diese	Behauptung	]	[$_{LK}$	hat	]	[$_{MF}$	Beckmeyer	keinen	Nachweis	]	[$_{RK}$	geliefert	]	.
	For	this	assertion			has			Beckmeyer	no	proof			provided		.

When the PP in the VF is not topicalized, it becomes very likely that the preposition attaches to a noun in the VF. Typical for this case is that the preposition is immediately preceded by a noun. Table 3 gives the distribution of such prepositions and shows that all nominal attachments are now in the VF. The immediately preceding noun is the head of the preposition in 88.65% of these cases, while in the other 11.35% of cases the head is another preceding VF noun.

			Preposition field	
			VF	NF
Prep. head field	**nominal**	VF	98.34	0.07
		MF	0.01	0.47
		NF	0.00	94.85
	verbal	LK	1.65	1.74
		RK	0.00	2.88

Table 3: Distribution of prepositions and their heads *when a preposition is in the VF or NF, and is immediately preceded by a noun.*

Zooming back out on the overall distribution in Table 2, we see that the overwhelming majority (55.24% versus 1.87%) of verbal heads is in the LK. This is an artifact of the dependency conversion of the TüBa-D/Z — which attaches PPs in the VF to the LK. For consistency, we reattach the preposition to the main verb when the main verb is not in the LK.

Nachfeld Prepositions in the NF regularly attach to heads in every bracket or field, except in the VF. NF preposition with nominal heads show again a marked preference for attachments to heads in the same field (35.74% nominal NF attachments in Table 2), while at the same time allowing for the most nominal attachments to another field, namely to the MF (6.15%). Similarly to the VF, if the preposition is immediately preceded by a noun in the NF, the head virtually always lies in the NF or brackets (Table 3).

3 Data set construction

The PP-attachment data set is constructed using a set of rules based on the insights of the previous section. Using this set of rules, which is summarized in Appendix A, we extracted 72,878 prepositions with at least two candidate heads from TüBa-D/Z release 10.

In Gloss 6 we show an example from our PP attachment data set. All the words bounded by boxes would be considered possible heads for the underlined preposition *an*[4]. However, using the insights from the previous section, we can remove the seven words highlighted in red from the candidate set and retain only the other four candidates. The correct candidate head *geflossen* is highlighted in green, while the incorrect candidates are blue. Compared to a crude extraction procedure, seven of the eleven original candidates are eliminated. The remaining four candidates are included in the dataset, specifying in each case whether the candidate is the actual head or not.

(6) [$_{\text{VF}}$ 165.000 Mark aus der bundesweiten Geldsammlung für die Flutopfer in Südpolen]
165,000 Mark from the nation-wide fundraiser for the flood-victims in south-Poland

sind [$_{\text{MF}}$ über das Konto des Bremer Landesverbandes der AWO [$_{\text{P}}$ an] die Caritas in
are via the account of-the of-Bremen state-chapter of-the AWO to the Caritas in

Danzig] geflossen .
Danzig flowed .

165,000 Mark from the nation-wide fundraiser for flood victims in south Poland flowed to Caritas in Danzig through the account of the Bremen state chapter of the AWO.

This selection of heads is representative for our data set. Table 4 shows the average number of possible heads before and after our candidate selection. On average, the number of candidates is reduced from 10.34 to 3.15. Moreover, the thesis that PP-attachment is not a binary classification task is confirmed by the average number of candidates in our data set. In the next section, we will explore the ramifications of the average number of candidates further.

Prep. field	Instances	Possible heads	Candidate heads
VF	21560	9.42	3.42
MF	48250	10.67	3.04
NF	3068	11.55	3.11
All	72878	10.34	3.15

Table 4: Average number of candidate heads for prepositions before and after selection. Only instances with at least two candidates are counted.

4 Consequences for the PP-attachment task

As discussed in Section 1, PP-attachment is typically treated as a binary classification task, where a preposition can be attached to the main verb or the noun that

[4] Of course, the noun *Caritas* is the complement of the preposition; if the PP is already bracketed this noun can be immediately removed from the candidate list.

immediately precedes the preposition. We have shown that this approach to PP-attachment is unsound, both because the preposition is not necessarily preceded by a noun in German (Section 2.2) and because the average number of heads is larger than two (Section 3). This raises two interesting questions: how many prepositions in ambiguous positions were missed because they were not immediately preceded by a noun; and is the task of preposition attachment more difficult when there are more than two candidate heads?

In order to answer these questions, we extract a data set from the PP-attachment set that was described in Section 3 where each preposition only has two candidate heads (binary data set). First, we remove all instances where either a preposition is not preceded by a noun or the head is a noun that is not the immediately preceding noun. From the remaining instances we remove all noun candidates, except for the noun that immediately precedes the head.

The binary data set contains 32.8% fewer training instances than the full data set. This answers our first question — in treating preposition attachment as a binary classification task, almost one third of the ambiguous prepositions are missed because they have no immediately preceding noun or incorrectly treated because another noun than the immediately preceding noun was the head.

To answer the second question, we train feed-forward neural networks that estimate attachment probabilities for each candidate, on the binary data set and on the data set with multiple candidate heads, respectively.[5] In the evaluation of both networks, the attachment with the highest probability is regarded as the attachment chosen by the network. The networks use a hidden ReLU layer, a sigmoid output layer, and the feature set proposed by Kübler et al. [10]: the preposition, the object of the preposition, the candidate head, and their part-of-speech tags (where each word or tag with one of these three relations is encoded as a binary feature). In addition, the absolute and relative distances between the candidate head and the preposition are added as integer features.

Table 5 shows the results on the two data sets (with binary v.s. multiple candidates) using an 80/20% split for training and evaluation data. We can clearly see that the realistic task with multiple head candidates is considerably more difficult than the binary classification task.

Data set	Train samples	Eval samples	Accuracy (%)
Binary	39179	9795	78.00
Multiple	39179	9795	68.79

Table 5: Preposition attachment accuracy with only two head candidates (*binary*) and multiple candidates (*multiple*).

[5] Since the binary data set is the smaller of the two sets, we first take a random sample of the set with multiple candidates so that both data sets have the same size.

5 Conclusion

Most previous work in German PP-attachment has assumed that a preposition attaches either to the immediately preceding noun or the main verb. However, the qualitative analysis in the present work provides evidence that prepositions do not only attach to immediately neighboring nouns (Examples 2, 3, and 5). The quantitative analysis shows that such contexts, where there are more than two candidates, are indeed very common. Consequently, PP-attachment should rather be considered to be a ranking task, supporting the thesis of Foth and Menzel [5].

Based on these insights, we have constructed a PP-attachment data set for German that includes all the realistic attachment points for a preposition, using topological field analyses. We expect that this data set can facilitate future research in PP-attachment, since our preliminary analysis in Section 4 has shown that the task is considerably harder under the presence of multiple head candidates.

The data set is provided as a stand-off annotation for the TüBa-D/Z treebank. This allows users of the data set to extract the information that is relevant to the task at hand from the annotation layers provided by TüBa-D/Z. This stand-off annotation will be made available to licensees of the TüBa-D/Z.[6]

Acknowledgments

Financial support for the research reported in this paper was provided by the German Research Foundation (DFG) as part of the Collaborative Research Center "The Construction of Meaning" (SFB 833), project A3.

References

[1] Steven P Abney. Stochastic attribute-value grammars. *Computational Linguistics*, 23(4):597–618, 1997.

[2] Daniël de Kok and Erhard W. Hinrichs. Transition-based dependency parsing with topological fields. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*, 2016.

[3] Erich Drach. *Grundgedanken der Deutschen Satzlehre*. Frankfurt/Main, 1937.

[4] Oskar Erdmann. *Grundzüge der deutschen Syntax nach ihrer geschichtlichen Entwicklung dargestellt*. Stuttgart: Cotta, 1886. Erste Abteilung.

[5] Kilian A. Foth and Wolfgang Menzel. The benefit of stochastic PP attachment to a rule-based parser. In *Proceedings of the COLING/ACL on Main*

[6] `http://www.sfs.uni-tuebingen.de/ascl/ressourcen/corpora/tueba-dz.html`

conference poster sessions, pages 223–230. Association for Computational Linguistics, 2006.

[6] Simon Herling. Über die Topik der deutschen Sprache. In *Abhandlungen des frankfurterischen Gelehrtenvereins für deutsche Sprache*, pages 296–362, 394. Frankfurt/Main, 1821. Drittes Stück.

[7] Donald Hindle and Mats Rooth. Structural ambiguity and lexical relations. *Computational linguistics*, 19(1):103–120, 1993.

[8] Tilman Höhle. Der Begriff 'Mittelfeld'. Anmerkungen über die Theorie der topologischen Felder. In A. Schöne, editor, *Kontroversen alte und neue. Akten des 7. Internationalen Germanistenkongresses Göttingen*, pages 329–340. Tübingen: Niemeyer, 1986.

[9] Mark Johnson, Stuart Geman, Stephen Canon, Zhiyi Chi, and Stefan Riezler. Estimators for stochastic unification-based grammars. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 535–541. Association for Computational Linguistics, 1999.

[10] Sandra Kübler, Steliana Ivanova, and Eva Klett. Combining Dependency Parsing with PP Attachment. In *Fourth Midwest Computational Linguistics Colloquium*, 2007.

[11] Joakim Nivre. An efficient algorithm for projective dependency parsing. In *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT)*, pages 149–160, 2003.

[12] Miles Osborne. Estimation of stochastic attribute-value grammars using an informative sample. In *Proceedings of the 18th conference on Computational linguistics-Volume 1*, pages 586–592. Association for Computational Linguistics, 2000.

[13] Adwait Ratnaparkhi. Statistical models for unsupervised prepositional phrase attachment. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, pages 1079–1085. Association for Computational Linguistics, 1998.

[14] Heike Telljohann, Erhard W. Hinrichs, Sandra Kübler, Heike Zinsmeister, and Kathrin Beck. Stylebook for the Tübingen treebank of written German (TüBa-D/Z). In *Seminar fur Sprachwissenschaft, Universitat Tubingen, Tubingen, Germany*, 2006.

[15] Yannick Versley. Parser evaluation across text types. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005)*, 2005.

[16] Martin Volk. Exploiting the WWW as a corpus to resolve PP attachment ambiguities. In *Proceedings of Corpus Linguistics*, volume 200, 2001.

A Extraction rules

In this appendix, we give a per-field description of these extraction rules. For all three fields, the verb candidate is the main verb. The main verb is found by (transitively) resolving the *AUX* relation (which is used to attach verbs to an auxiliary or modal) until we encounter a verb that is not an auxiliary or modal verb.

Mittelfeld To find the set of candidate heads in the MF, we scan backwards from a preposition until we find a token that forms the LK. Every noun on this path is marked as a candidate head. If the clause under consideration is a main clause, the finite verb is in the LK. We resolve for the main verb using the LK and add it to the candidate set. If the clause is a subordinate clause, the LK is normally a complementizer, which has an attachment to the finite verb in the RK. We use this verb to find the main verb and add it to the candidate set.

Vorfeld While extracting from the VF, we should take two different scenarios into account: (1) the preposition is immediately preceded by a noun in the VF or (2) the preposition is not immediately preceded by a noun in the VF. In the former case, we only add nominal candidates in the VF that precede the preposition. In the latter case, nouns in the MF are added as candidates as well. The verb candidate is found by scanning rightward from the preposition until we find the LK. The verb in the LK is then used to find the main verb.

Nachfeld Processing of the NF is similar to the VF: when the preposition is immediately preceded by a noun, nouns in the NF immediately preceding the preposition are added as candidates. If the preposition is not preceded by a noun, nouns in the MF are added as well. To find the main verb candidate, we scan leftward until we find a bracket and resolve for the main verb.