A Logistic Regression Model of Determiner Omission in PPs

Tibor Kiss **Katja Keßelmeier** **Antje Müller** **Claudia Roch** **Tobias Stadtfeld** **Jan Strunk**

Sprachwissenschaftliches Institut,
Ruhr-Universität Bochum

{tibor, kesselmeier, mueller, roch, stadtfeld, strunk}@linguistics.rub.de

Abstract

The realization of singular count nouns without an accompanying determiner inside a PP (determinerless PP, bare PP, Preposition-Noun Combination) has recently attracted some interest in computational linguistics. Yet, the relevant factors for determiner omission remain unclear, and conditions for determiner omission vary from language to language. We present a logistic regression model of determiner omission in German based on data obtained by applying annotation mining to a large, automatically and manually annotated corpus.

1 The problem and how to deal with it

Preposition-Noun Combinations (PNCs, sometimes called determinerless PPs or bare PPs) minimally consist of a preposition and a count noun in the singular that – despite requirements formulated elsewhere in the grammar of the respective language – appears without a determiner. The noun in a PNC can be extended through prenominal modification (1) and postnominal complementation (2). Still, a determiner is missing. The following examples are given from German.

(1) *auf parlamentarische Anfrage* ('after being asked in parliament'), *mit beladenem Rucksack* ('with loaded backpack'), *unter sanfter Androhung* ('under gentle threat')

(2) *Er wehrt sich gegen die Forderung nach*
he defies REFL against the demand for

Stilllegung einer Verbrennungsanlage.
closedown an incineration plant
'He defies the demand for closing an incineration plant.'

PNCs occur in a wide range of languages (Himmelmann, 1998); the conditions for determiner omission, however, have not been detected yet, and conditions applying to one language do not carry over to other languages. In addition, speakers only reluctantly judge the acceptability of newly coined PNCs, so that reliance to introspective judgments cannot be assumed.

For English, Stvan (1998) and Baldwin et al. (2006) have claimed that either the semantics of the preposition or of the noun play a major role in determining whether a singular count noun may appear without a determiner in a PNC. Stvan (1998) assumes that nouns determine the well-formedness of PNCs (3) if the denotation of the noun occurs in a particular semantic field, while Baldwin et al. (2006) assume that certain prepositions impose selection restrictions on their nominal complements that allow for determiner omission (4).

(3) *from school, at school, in jail, from jail, …*

(4) *by train, by plane, by bus, by pogo stick, by hydro-foil …*

Interestingly, Le Bruyn et al. (2009) have observed that basic assumptions of Stvan's analysis do not apply to Dutch, French, or Norwegian. With regard to German, we observe that neither the pattern in (3) nor in (4) is productive. Constructions like (4) cannot be realized as PNCs in German, but require full PPs.

In the following, we propose an analysis of PNCs that combines corpus annotation, annotation mining (Chiarcos et al., 2008), and logistic regression modeling (Harrell, 2001). Annotation mining assumes that linguistically relevant generalizations can be derived in a bottom-up fashion from a suitably annotated corpus. Relevant hits in the corpus are mapped into a feature vector that serves as input for logistic regression classification. In the present case, the input con-

sists of sentences containing either PNCs or PPs. Binary logistic regression suggests itself as a classification method since the problem of PNCs can be rephrased as the following question: Under which conditions can an otherwise obligatory determiner be omitted?

The majority of required annotations can be derived automatically, but there are no available systems for the automatic determination of preposition senses in German, so preposition sense annotation has to be carried out manually and requires a language-specific tagset for preposition senses.

While our initial analysis is based on German data, the general methodology can be applied to other languages, provided that corpora receive proper annotation.

2 Corpus annotation

2.1 General characteristics

The present analysis is based on a newspaper corpus of the Swiss-German newspaper *Neue Zürcher Zeitung* from 1993 to 1999, comprising approx. 230 million words. The annotation is based on an XML-stand-off format. MMAX2 (Müller and Strube, 2006) is used for manual annotation. Annotations are carried out both for PNCs and for full-fledged PPs. For each preposition, the following data is considered: PNCs, where N is a count noun; corresponding PPs with the same count noun; and PPs containing count nouns not appearing inside PNCs.

The following annotations are provided for each dataset in the corpus:

Lexical level: part-of-speech, inflectional morphology, derivational morphology of nouns, count/mass distinction of nouns, interpretation of nouns, interpretation of prepositions, noun compounding.

Syntactic level: mode of embedding of the phrase (adjunct or complement), syntactic dependents of the noun, modification of the noun.

Global level: Is the phrase contained in a headline, title, or quotation? Is the phrase idiomatic? Headlines, titles, and quotations are particularly prone to text truncation and PNCs occurring here might not be the result of syntactic operations. Similarly, idiomatic PNCs *and* PPs might follow combination rules that differ from the general modes of combination. Hence, the

annotations may serve to exclude these cases from general classification.

2.2 Automatic annotation

The following tools are employed for automatic annotation: Regression Forest Tagger (Schmid and Laws, 2008) for POS tagging and morphological analysis (the tagger contains the SMOR component for morphological analysis, cf. Schmid, 2004), and Tree Tagger (Schmid, 1995) for chunk parsing.

To determine noun meanings, we make use of two resources. The first resource is GermaNet (Kunze and Lemnitzer, 2002), the German version of WordNet. We employ 23 top-level categories, and each noun is annotated with every top-level category it belongs to.[1] Secondly, we use the computer lexicon HaGenLex (Hartrumpf et al., 2003), which offers specific sortal information derived from a formal ontology for each noun. Finally, we employ a classifier for the count/mass distinction. The classifier combines lexical statistics, expressed in terms of a decision tree classifier, with contextual information, which is handled by Naïve Bayes classification (cf. Stadtfeld 2010). The classification is based on the fine-grained distinctions first introduced in Allan (1980), but we employ a reduced set of five instead of eight classes. The classifier is type-based as it makes use of the relation between singular and plural realizations of noun lemmas, but takes the immediate context of the lemma into account.

Nouns are only assigned to a particular class if both classifiers come to the same result w.r.t. this class assignment. While this leads to some nouns being excluded from the count/mass distinction, the resulting classes show a high degree of precision.

2.3 Manual annotation of preposition senses

Prepositions are highly polysemous. What is more, the relation between a preposition and its senses has to be determined in a language-

---

[1] Nouns that are assigned to more than one top-level category are presumably homonymous or polysemous. We do not disambiguate the nouns. The reason is that individual features will be evaluated for their effect in a logistic model, and an ambiguous noun will receive a value in each feature. Hence, we can be sure that a significant semantic feature will be included in the classification.

specific manner. While the *Preposition Project* forms a basis for preposition sense annotation in English (cf. Litkowski and Hargraves 2005, 2007), little attention has been paid to specialized annotation schemata for preposition senses in German, which form the first prerequisite for a classification of preposition senses.

Based on four usage-based grammars and dictionaries of German (Duden 2002, Helbig and Buscha 2001, Durrell and Brée 1993, Schröder 1986), we have developed an annotation schema with a hierarchical structure, allowing for subtrees of preposition senses that require a fine-grained classification (such as TEMPORAL, SPATIAL, CAUSAL, and PRESENCE). For temporal and spatial interpretations, the annotation is further facilitated by the use of decision trees.[2]

Altogether, the annotation schema includes the following list of top-level categories: MODAL, CAUSAL, PRESENCE, SPATIAL, TEMPORAL, STATE, COMITATIVE, AGENT, REDUCTION/EXTENSION, PARTICIPATION, SUBORDINATION, RECIPIENT, AFFILIATION, CORRELATION/INTERACTION, TRANSGRESSION, ORDER, THEME, SUBSTITUTE, EXCHANGE, COMPARISON, RESTRICTIVE, COPULATIVE, ADVERSATIVE, DISTRIBUTIVE, STATEMENT/OPINION, CENTRE OF REFERENCE, and REALISATION.

Based on an extension of the weighted kappa statistic we have reached an overall kappa value (κ_w) of 0.657 and values between 0.551 and 0.860 for individual features (cf. Müller et al. 2010a). Two properties of the annotation schema prohibit the application of a standard kappa statistic: First, the schema allows subsorts, and secondly, a preposition may receive more than one annotation if its sense cannot be fully disambiguated. The values reported in Müller et al. (2010) for maximal subtypes such as SPATIAL (κ_w = 0.709) and TEMPORAL (κ_w = 0.860) can be equated to aggregate values in standard kappa statistics.

In the models presented below, we employ top-level categories only and have aggregated more specific sense annotations.

---

3 Preparing logistic regression models for *ohne* ('without') and *unter* ('under', 'below')

The problem of PNCs, i.e. why a determiner is omitted in a construction which otherwise requires the realization of the determiner, can be rephrased as a problem for binary logistic regression and classification.

While binary logistic regression does not prohibit monocausal explanations, typical models for binary logistic regression employ more than one factor, and the value of the coefficients models the relative influence of the individual factors. Logistic regression thus does not only help to identify factors for determiner omission, but also reveals the interplay of multiple licensing conditions – thus possibly accounting for the relative difficulty to distinguish acceptable from inacceptable PNCs.

We are aiming at a description of PNCs in German for the 22 prepositions listed in (5).

(5) *an, auf, bei, dank, durch, für, gegen, gemäß, hinter, in, mit, mittels, nach, neben, ohne, seit, über, um, unter, vor, während, wegen*

These prepositions have been chosen on the basis of the following two assumptions: a) they appear in PNCs and PPs, and b) their 'typical' object is an NP.

We present logistic regression models of determiner realization for two prepositions: *ohne* ('without') and *unter* ('under', 'below'). The first preposition, *ohne*, is the only preposition that appears more often in PNCs than in PPs. The second preposition, *unter*, belongs to the class of highly polysemous prepositions. In fact, it is the preposition with the second largest number of senses (10 senses), only surpassed by *mit* ('with') (11 senses)*, which however appears much more often than *unter* in the corpus and thus requires further annotation. The following table summarizes the distribution of PNCs and PPs for both prepositions, after tokens that had been annotated as belonging to *headlines, quotations, telegram style sentences,* or as being idiomatic were excluded from the data. With regard to the first group (headlines etc.), the elimination mostly applies to PNCs, but among the PPs we found many idiomatic expressions and fixed phrases, which have also been excluded from modeling.

---

[2] The schema does not directly distinguish between local and directional senses, but makes use of cross-classification to deal with the distinction. Cf. Müller et al. (2010b).

Preposition	Σ	PP	PNC
ohne	3,750	591	3,159
unter	5,181	4,334	857

Table 1. Data Distribution of PNCs and PPs

The analysis has been carried out in R (R Development Core Team, 2010) and makes extensive use of Harrell's DESIGN package (Harrell, 2001).

The feature vector consists of the dependent variable – the factor DET with its levels *no* and *yes* – and of relevant classificatory features representing the interpretation of the preposition (in terms of the features presented in section 2), the internal syntactic structure of the nominal projection (prenominal modification of N, syntactic arguments of N, internal structure of N as a compound, derivational status of N), the external syntactic embedding of the PNC or PP, and the interpretation of the noun.

Features starting with DEP signify syntactic arguments of the noun (DEP-S a sentential complement, DEP-NP an NP complement, etc.); the feature ADJA signifies the presence of one or more modifying adjectives; the feature COMPOUND indicates whether the noun in question is a compound. The feature GOVERNED indicates whether a noun or a verb governs the phrase. The feature NOMINALIZATION provides information about the derivational structure of the noun, in particular it indicates whether a noun is derived from a verb by use of the suffix *–ung*.

Features starting with GN are GermaNet top-level categories, features starting with HL are HaGenLex ontological sorts; both describe the interpretation of the noun.

The statistical modeling started with the assumption that each feature is relevant, so that an initial feature set of 92 features was considered. Feature elimination took place through *fast backwards elimination* (Lawless and Singhal, 1978) and manual inspection. The results of fast backwards elimination were not followed blindly. Following Harrell's (2001:56) suggestion, we have kept factors despite their low significance levels. In most cases, however, manual inspection and fast backwards elimination suggested the same results. The resulting models were subjected to *bootstrap validation* to identify possible overfitting (cf. section 5.1).

The value DET = *no* is taken to be the default value in the following models. As a consequence, negative values for coefficients indicate rising probability for an omission of a determiner, while positive coefficients shift odds in favor of a realization of the determiner.

4 Logistic models for the omission of a determiner with *ohne* and *unter*

The logistic regression models developed for the prepositions *ohne* and *unter* make use of 13 and 22 features, respectively. In each case, we have started with a full model fit (Harrell, 2001:58f.), evaluated the full model and eliminated factors through manual inspection and fast backwards elimination. The coefficients for the models for *ohne* and *unter* are reported in tables 2 and 3.

	Coef.	S.E.	Wald Z	p
INTERCEPT	-2.4024	0.1109	-21.66	0.000
NOMINAL.	-1.3579	0.1870	-7.26	0.000
ADJA	1.1360	0.1188	9.57	0.000
CAUSAL	1.2063	0.1302	9.26	0.000
COMITAT.	2.2821	0.5201	4.39	0.000
PARTICIP.	3.4027	0.4895	6.95	0.000
PRESENCE	-0.7780	0.1463	-5.32	0.000
DEP-S	5.0797	1.0542	4.82	0.000
DEP-NP	2.9752	0.1718	17.32	0.000
DEP-PP	2.1978	0.1487	14.78	0.000
GN-RELAT.	-1.0292	0.4072	-2.53	0.011
GN-ATTR.	-1.3528	0.3038	-4.45	0.000
GN-EVENT	-0.8431	0.1431	-5.89	0.000
GN-ARTE.	-0.4117	0.1564	-2.63	0.008

Table 2. Coefficients for a logistic regression model of determiner omission with *ohne*.[3]

---

[3] In the following tables, S.E. stands for standard error Wald Z reports the Z-score of the Wald statistic, which is determined by divided the value of the coefficient through its standard error. The squared Wald Z statistic is χ^2-distributed and thus indicates the goodness of fit for the coefficients of the model.

	Coef.	S.E.	Wald Z	p
INTERCEPT	-0.4379	0.1657	-2.64	0.008
NOMINAL.	-0.8346	0.2259	-3.70	0.000
ADJA	-1.0177	0.1432	-7.11	0.000
COMPOUND	2.1719	0.2538	8.56	0.000
GOVERNED	1.9894	0.3017	6.59	0.000
SPATIAL	2.3237	0.2044	11.37	0.000
CAUSAL	1.3047	0.2272	5.74	0.000
SUBORD.	3.0529	0.2559	11.93	0.000
ORDER	3.4228	0.1861	18.40	0.000
TRANSGR.	4.4186	0.3677	12.02	0.000
DEP-S	8.4717	4.0734	2.08	0.037
DEP-NP	0.8551	0.1436	5.95	0.000
DEP-PP	0.3043	0.2170	1.40	0.161
GN-GROUP	0.5241	0.2563	2.04	0.041
GN-COMM.	-0.9149	0.1443	-6.34	0.000
GN-LOC.	2.2704	0.6208	3.66	0.000
GN-REL.	-2.1161	0.6022	-3.51	0.000
GN-POSS.	-0.8482	0.3665	-2.31	0.021
GN-ATTR.	-2.2847	0.2741	-8.33	0.000
GN-ARTE.	0.4169	0.1601	2.60	0.009
GN-HUM.	1.8870	0.4999	3.77	0.000
HL-AD	-1.0253	0.1888	-5.43	0.000
HL-AS	-1.4214	0.3804	-3.74	0.000

Table 3. Coefficients for a logistic model of determiner omission with *unter*.

General measures of the two models are reported in table 4. Somers' D_{xy} describes the proportion of observations, for which the model provides an appropriate class probability. D_{xy} can be derived from C, the corresponding receiver operating characteristic curve area, since $D_{xy} = 2 \times (C-0.5)$. Model L.R. (likelihood ratio) indicates the improvement reached by including the predictors. Degrees of freedom (d.f.) have been omitted from table 4, as they correspond to the number of predictors, i.e. 12 in the case of *ohne* and 23 in the case of *unter*. The high figures for Somers' D_{xy} are reassuring.

	Model L.R.	p	C	D_{xy}
ohne	1,063.5	0	0.876	0.753
unter	2,245.6	0	0.937	0.874

Table 4. Model Quality.

4.1 The model for *ohne*

Starting with the model in table 2, we can identify several groups of factors:

The first group comprises the interpretation of the preposition. The group discriminates between determiner omission and realization. The semantic features CAUSAL, COMITATIVE, and PARTICIPATION show positive coefficients, suggesting that prepositions receiving the aforementioned interpretations tend to favor an 'ordinary' NP including a determiner. The interpretation PRESENCE, on the other hand, shows a negative coefficient and thus suggests the omission of a determiner. There are further senses of *ohne,* which do not have a significant effect on determiner omission/realization.

Turning to the representation of syntactic argument structure of the noun, we find that the coefficients of DEP-S, DEP-NP, and DEP-PP receive positive values throughout. The presence of syntactic complements thus shifts odds in favor of determiner realization. There is a strong preference against determiner omission with DEP-S, and somewhat weaker values for Dep-NP and Dep-PP, respectively. A comparison of interpretation and complement realization offers a general assessment of PNCs. As *ohne* and *unter* share only a few senses, we do not necessarily expect that the discerning senses relevant for a realization of a PNC with *ohne* carry over to *unter;* but we do expect that features pertaining to the syntactic structure of the nominal complement play a role not only for *ohne,* but for *unter* (or for prepositions admitting PNCs in general) as well. And this prediction is actually borne out in the model for *unter.* The model thus already offers interesting insights not only w.r.t. the realization conditions of PNCs and PPs headed by *ohne,* but for broader analyses of PNCs as well.

We will return to the role and value of the features ADJA and NOMINALIZATION in section 4.3.

The last group comprises the semantic characteristics of nouns derived from GermaNet. If a noun is classified as belonging to the relevant GermaNet top-level categories, determiner omission is favored.

4.2 The model for *unter*

A first glance at the model for *unter* shows that it requires a larger set of predictors than the model for *ohne.* In part, this is due to the higher degree of polysemy of *unter:* with more senses, we expect more semantic predictors to enter the discrimination. In addition, a wider range of senses

also allows for a wider range of selection restrictions, and hence for a larger number of different sortal specifications for selected nouns. The higher complexity of the model, however, should not conceal a peculiarity of this model that casts serious doubt on the idea that PNCs are monocausally licensed by particular senses of a preposition: the model selects five senses from the ten top level interpretations of *unter,* but the coefficients are unsigned. Thus, the model indicates that the senses SPATIAL, CAUSAL, SUBORDINATION, ORDER, and TRANSGRESSION *block* the omission of a determiner. What we do not find are senses that favor the omission of a determiner.

The features DEP-S, DEP-NP, and DEP-PP again favor the realization of a determiner. A comparison of the coefficient of DEP-S to the coefficients of DEP-NP and DEP-PP shows, however, that the presence of a sentential complement has a strong influence on determiner realization, while NP- and PP-complements may still occur in PNCs, as their coefficients are relatively low (also in comparison to the coefficients of these values for *ohne).[4]*

In more general terms, we suspect a general mechanism relating sentential complementation to the realization of the determiner, a topic to be addressed in future research.

It should also be noted that the external syntactic realization of the phrase plays a role for *unter.* The feature GOVERNED did not play a role for *ohne,* but suggests the realization of a determiner for *unter.* The reason might be that few verbs or nouns govern the preposition *ohne.* Prepositional objects headed by *unter,* however, are more common. Prepositional objects headed by *ohne* make up only 1.2 % of the occurrences of *ohne* in the present corpus, while the share of prepositional objects headed by *unter* is three times larger: 3.6 %.

Finally, we note that a variety of sortal classifications for nouns suggest either an omission or realization of the determiner, supporting the assumption that in addition to the preposition's meaning, the meaning of the noun plays a role. GermaNet top-level categories were already discriminating in the model for *ohne;* but the model for *unter* also makes use of HaGenLex sortal categories (HL-AD and HL-AS). The predictors stand for dynamic and static concepts that both receive an abstract interpretation. Their inclusion is particularly interesting, as it is sometimes claimed (e.g. Bale and Barner, 2009) that 'abstract' nouns are never to be classified as count nouns.

4.3 General assessment of the models

Both models show that the realization of syntactic complements, of sentential complements in particular, seems to impede determiner omission. That syntactic complexity does not seem to play a role per se, can be deduced from the coefficients for the factor ADJA: While ADJA favors determiner realization with *ohne,* it prohibits determiner realization with *unter.*

The role of morphological derivation through *–ung,* as represented by the factor NOMINALIZATION, is the same in both models: derived nominals shift odds in favor of determiner omission. While the derivational structure might be considered a formal property of the construction, it might also reflect an underlying denotational distinction between events and objects, which has to be clarified in future work.

It is a striking feature of the model for *unter that* we do not find interpretational features of the preposition *unter* that favor determiner omission. Taken together with the other factors in the two models presented, the analysis suggests a picture rather different from the (more or less) monocausal analyses of Stvan (1998) and Baldwin et al. (2006). With regard to *unter* a model in the sense of Baldwin et al. (2006) could only provide negative rules of the form *"if P does not mean this, its nominal complement may be realized without a determiner",* but such a model would lead to less precision than the multicausal model presented here.

5 Validation of the models

5.1 Bootstrap validation

Logistic regression models may suffer from overfitting the data. We have thus carried out a

---

[4] One could argue against the inclusion of the coefficient for DEP-PP altogether, as it does not seem to be significant (p > 0.05) in the first place. However, we have followed Harrell's (2001) advice that blind exclusion of seemingly insignificant factors may not lead to model improvement. In fact, models for *unter* including Dep-PP outperform models excluding this feature.

bootstrap validation of both models and applied penalized maximum likelihood estimation (Harrell, 2001) to the models. The results of the initial (non-penalized) models are reported in Table 5 and Table 6, where we report values for D_{xy} and the average maximal error of the model. Bootstrap validation makes use of sampling with replacement. The training samples for evaluation thus may contain certain instances many times, but some original data will never be sampled and can thus be used for testing the models. Bootstrap validation is carried out 200 times, the results being averaged. The overfitting of the models is determined by the optimism derived from the bootstrap evaluation.

	D_{xy}	E_{max}
Original Index	0.7525	0.0000
Training	0.7578	0.0000
Test	0.7497	0.0123
Optimism	0.0080	0.0123
Corrected Index	0.7445	0.0123

Table 5. Bootstrap validation of model for *ohne*.[5]

	D_{xy}	E_{max}
Original Index	0.8737	0.0000
Training	0.8741	0.0000
Test	0.8690	0.0072
Optimism	0.0051	0.0072
Corrected Index	0.8685	0.0072

Table 6. Bootstrap validation of model for *unter*.

Penalized maximum likelihood estimation (Harrell, 2001:207) for both models resulted in penalties of 0.3 and 0.8, respectively, based on Akaike's AIC. The updated models have again been bootstrap validated, resulting in the improved values presented in table 7 and table 8.

	D_{xy}	E_{max}
Original Index	0.7526	0.0000
Training	0.7570	0.0000
Test	0.7500	0.0096
Optimism	0.0070	0.0096
Corrected Index	0.7456	0.0096

---

[5] E_{max} is the maximal error determined in average over the bootstrap runs.

Table 7. Bootstrap validation of penalized model for *ohne*.

	D_{xy}	E_{max}
Original Index	0.8736	0.0000
Training	0.8744	0.0000
Test	0.8692	0.0055
Optimism	0.0052	0.0055
Corrected Index	0.8684	0.0055

Table 8. Bootstrap validation of penalized model for *unter*.

5.2 Representing the influence of factors in a nomogram

The respective influence of individual factors can be read of a nomogram (Banks, 1985) derived from the models presented above (we make use of a tabular presentation for reasons of legibility). The nomogram for *ohne* consists of the tables 9 and 10. Table 9 lists the individual scores for the factors in the model for *ohne,* were 0 indicates that the pertinent property is not present and 1 indicates that the property is present. Table 10 maps the sum to probability of determiner omission.

Predictor	0	1
NOMINALIZATION	27	0
ADJA	0	22
CAUSAL	0	24
COMITATIVE	0	45
PARTICIPATION	0	67
PRESENCE	15	0
DEP-S	0	100
DEP-NP	0	59
DEP-PP	0	43
GN-RELATION	20	0
GN-ATTRIBUTE	27	0
GN-EVENT	17	0
GN-ARTEFACT	8	0

Table 9. Nomogram: individual scores of predictors for *ohne*.

Total Points	Pr("Omission of Det")
118	0.9
134	0.8
144	0.7
153	0.6

161	0.5
169	0.4
178	0.3
188	0.2
204	0.1

Table 10. Nomogram: mapping from total points to probability of determiner omission.

As an illustration, consider pairs of *ohne* and a noun with the values in (6) and (7).

(6) NOMINALIZATION = 1, ADJA = 1, COMITA-TIVE = 1, all other senses including PRESENCE = 0, all GN features = 0, DEP features = 0.

(7) NOMINALIZATION = 0, ADJA = 1, PRESENCE = 1, all other senses = 0, GN-ATTRIBUTE = 1, all other GN features = 0, DEP features = 0.

Given the individual scores for the factors in table 9, the total number of points for the combination in (6) is 144, leading to a probability of 0.7 that a determiner will be omitted in the construction. In other words, a determiner omission is likely with the feature set given in (6). In (7), we reach a total of 92 only, so that the likelihood of determiner omission rises above 0.9.

6   Summary and prospects

The models presented support the general assumption that the realization or omission of a determiner in a prepositional phrase should be analyzed as a multicausal phenomenon. The logistic regression analysis presents evidence for the assumption that the senses of the preposition and the interpretation of the noun (possibly governed by selection restrictions of the preposition) as well as the syntactic complexity of the embedded nominal projection are major factors in determining whether an article can be dropped or not.

With regard to the complexity of the nominal projection, the two models presented here indicate that it is not complexity per se, but that the realization of a complement of the noun, in particular of a sentential complement, clearly raises the probability of article realization. While this is a speculation, based on the models presented here, it might very well be that this dependency reflects a deeper referential requirement.

In developing further models for prepositions, we expect that the realization of a complement of the noun will establish itself as a common factor, but this has to await further research and model development.

References

Allan, Keith. 1980. Nouns and countability. *Language* 56(3):541-567.

Baldwin, Timothy, John Beavers, Leonoor van der Beek, Francis Bond, Dan Flickinger, and Ivan Sag. 2006. In search of a systematic treatment of determinerless PPs. In Patrick Saint-Dizier (eds.), *Syntax and Semantics of Prepositions*. Springer, Dordrecht, 163-179.

Bale, Alan, and David Barner. 2009. The interpretation of functional heads: Using comparatives to explore the mass/count distinction. *Journal of Semantics 26,* 217-252.

Banks, J. 1985. Nomograms. In S. Kotz and N.L. Johnson (eds.), *Encyclopedia of Statistical Sciences,* Vol. 6. Wiley, New York.

Chiarcos, Christian, Stefanie Dipper, Michael Götze, Ulf Leser, Anke Lüdeling, Julia Ritz, and Manfred Stede. 2008. A flexible framework for integrating annotations from different tools and tagsets. *Traitement Automatique des Langues*. Special Issue Platforms for Natural Language Processing. ATALA, 49 (2).

Duden. 2002. Duden. Deutsch als Fremdsprache. Bibliographisches Institut and F.A. Brockhaus AG, Mannheim.

Durell, Martin and David Brée. 1993. German temporal prepositions from an English perspective. In Cornelia Zelinsky-Wibbelt (ed.), The Semantics of Prepositions. From Mental Processing to atural Language Processing. De Gruyter, Berlin/New York, 295-325.

Harrell, Frank E. 2001. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer: New York.

Hartrumpf, Sven, Hermann Helbig, and Rainer Osswald. 2003. The Semantically Based Computer Lexicon HaGenLex - Structure and Technological Environment. *Traitement automatique des langues* 44(2):81-105.

Helbig, Gerhard and Joachim Buscha. 2001. Deutsche Grammatik. Ein Handbuch für den Ausländerunterricht. Leipzig, Langenscheidt.

Himmelmann, Nikolaus. 1998. Regularity in irregularity: Article use in adpositional phrases. *Linguistic Typology*, 2:315–353.

Kunze, Claudia, and Lothar Lemnitzer. 2002. GermaNet - representation, visualization, application. *Proc. LREC 2002*, main conference, Vol V., 1485-1491.

Lawless, J. and K. Singhal. 1978. Efficient screening on nonnormal regression models. *Biometrics 34*:318-327.

Le Bruyn, Bert, Henriëtte de Swart, and Joost Zwarts. 2009. *Bare PPs across languages*. Presented at the Workshop on Bare nouns, Paris.

Müller, Antje, Olaf Hülscher, Claudia Roch, Katja Keßelmeier, Tobias Stadtfeld, Jan Strunk, and Tibor Kiss. 2010. An Annotation Schema for Preposition Senses in German. Proceedings of ACL-LAW IV, Uppsala, Sweden.

Müller, Antje, Katja Keßelmeier, Claudia Roch, Jan Strunk, Tobias Stadtfeld and Tibor Kiss. 2010. Creating a Feature Space for the Annotation of Preposition Senses in German. Linguistic Evidence,TÜbingen 2010.

Müller, Christoph, and Michael Strube. 2006. Multi-level annotation of linguistic data with MMAX2. In Sabine Braun, Kurt Kohn, Joybrato Mukherjee, (eds.), *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*. Peter Lang, Frankfurt a.M., 197-214.

Nivre, Joakim. 2006. *Inductive Dependency Parsing (Text, Speech, and Language Technology)*. New York: Springer.

R Development Core Team. 2010. R: *A language and environment for statistical computing*. Foundation for Statistical Computing, Vienna, Austria. http://www.rproject.org.

Stadtfeld, Tobias. 2010. Determining the Countability of English and German Nouns. Ms. Ruhr-University Bochum.

Schmid, Helmut. 1995. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the EACL SIGDAT Workshop*, Dublin, Ireland, March.

Schmid, Helmut, Arne Fitschen, and Ulrich Heid. 2004. SMOR: A German computational morphology covering derivation, composition, and inflection. In *Proceedings of LREC 2004*, 1263-1266, Lisbon, Portugal.

Schmid, Helmut, and Florian Laws. 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In *Proceedings of COLING 2008*, Manchester, UK.

Schröder, Jochen. 1986. Lexikon deutscher Präpositionen. Leipzig, VEB Verlag Enzyklopädie.

Stvan, Laurel S. 1998. *The Semantics and Pragmatics of Bare Singular Noun Phrases*. Ph.D. thesis, Northwestern University, Evanston/ Chicago, IL.