On the Role of Explicit Morphological Feature Representation in Syntactic Dependency Parsing for German

Wolfgang Seeker and Jonas Kuhn
Institut für Maschinelle Sprachverarbeitung
University of Stuttgart
{seeker,jonas}@ims.uni-stuttgart.de

Abstract

We investigate the question whether an explicit feature representation for morphological features is necessary when parsing German with a fully lexicalized, statistical dependency parser. We use two morphosyntactic phenomena of German to show that while lexicalization does indeed suffice to a large extent when recovering the internal structure of noun phrases, an accurate explicit representation can support the correct selection of its grammatical function.

1 Introduction

German is usually considered a border case between morphologically rich languages like Czech and morphologically poor languages like English. It does show phenomena that are typical for morphologically rich languages, e. g. a rich nominal and verbal inflection system and hence a relatively free word order. However, compared to Czech or other morphologically rich languages, the morphological system is less elaborate and characterized by a large amount of form syncretism, which introduces a lot of ambiguity.

A lot of work investigated the best way to utilize morphological information in statistical PCFG parsers for German, mostly by transforming the treebank making morphological information more accessible (Schiehlen, 2004; Dubey, 2005). Lexicalization of PCFGs has been a controversial subject of research in German, where some found no effect (Dubey and Keller, 2003) while others did (Kübler et al., 2006; Rafferty and Manning, 2008). However, this work concentrated on constituent parsing. While there are many parsing results of dependency parsers on German (Buchholz and Marsi, 2006; Kübler, 2008; Hajič et al., 2009), the investigation of morphological representations and their interplay with dependency parsing algorithms has been started only recently (cf. Tsarfaty et al. (2010)). In this paper, we pursue the question of how important it is to mark morphological information explicitly for a data-driven lexicalized dependency parser when applied to German. We therefore investigate the performance of the parser on two morphosyntactic phenomena of German, namely the agreement within a noun phrase[1] and the recognition of the grammatical function of a noun that is marked by its case value.

2 Morphology of German Noun Phrases

Three morphological categories participate in the agreement of a German noun phrase: gender, number, and case.[2] Number and gender values are governed by the noun, and the case value is determined by the grammatical function of the noun. The dependents of a noun (determiners, adjectives, attributive pronouns) need to agree with their head noun in these three categories.

(1)

die	*Öl*
ART+nom/acc.sg.fem	NN+acc.sg.neut
the	oil
verarbeitende	*Industrie*
ADJ+nom/acc.sg.fem	NN+nom/acc.sg.fem
processing	industry

'the oil processing industry'

Example 1 shows a German noun phrase consisting of a determiner (*die*), an adjective (*verarbeitende*), and a noun (*Industrie*). Additionally, the noun *Öl* is an argument of the adjective. Without morphological information we might in principle be dealing with two separate noun phrases here, which just happen to appear in adjacent position. However, agreement tells us that the determiner is not depending on the first noun because morphologically it marks either singular feminine

[1]We use the term noun phrase to denote a noun and all its direct dependents although strictly speaking there are no phrases in dependency syntax.

[2]German has three gender values, two number values, and four case values.

Proceedings of the 12th International Conference on Parsing Technologies, pages 58–62,
October 5-7, 2011, Dublin City University.

or plural for all genders, but it cannot mark singular neuter. It is thus morphologically incompatible with the first noun.

According to Eisenberg (2006, 142), the inflectional paradigms of the parts of the German noun phrase have developed such that they mark morphological information with diverging explicitness. The declension patterns of nouns tend to use different forms to mark number but show form syncretism in marking case while determiners (and adjectives) show more form syncretism for marking number than for marking case. Eisenberg therefore argues for what he calls function sharing (*Funktionsteilung*) in the German noun phrase. In Example 1, the determiner by itself can mark nominative or accusative case for feminine singular nouns, or for plural nouns of every gender. The second noun, to which the determiner is attached, is ambiguous for all four cases but cannot be plural. This shows the importance of the agreement relation because only by agreement, determiner and noun disambiguate each other for nominative or accusative feminine singular.

Example 1 also demonstrates an inherent problem of the German nominal paradigms as a whole. Because of the vast amount of syncretism in the system, there are some ambiguities that can never be resolved by formal means, but need to be disambiguated by their semantic context. This affects the distinction between nominative and accusative for all feminine, neuter, and plural nouns as well as the distinction between genitive and dative case for feminine singular nouns. In Example 1, we therefore cannot tell without further context which one of the two possible case values is correct.

(2)	*den Löwen*	*sieht*	*der Hund*
	OBJ+acc		SUBJ+nom
	the lion	see	the dog
	'the dog sees the lion'		

While morphological information by agreement helps us recover the internal structure of a noun phrase, it also plays a role when determining the grammatical function of the whole phrase. German uses its four case values to mark the arguments of verbs, adpositions, and adjectives. Example 2 shows a transitive sentence where the subject is marked by nominative case and the object is marked by accusative case. In German, the subject of a sentence will always be in the nominative case, while the structural object receives accusative case. In ditransitive sentences, the direct object gets accusative case while the indirect object receives dative case.[3]

The relation between a case system and the grammatical functions in a language is usually not a one-to-one mapping (Blake, 2001, 48ff). In German, nominative encodes subjects and predicates, accusative mostly marks objects and some adjuncts, dative also marks objects, and genitive mostly marks possessive constructions but can also mark objects and some adjuncts. Since the mapping is not one-to-one, a certain amount of ambiguity remains (e. g. both subject and predicate are marked by nominative case), but it also restricts the choice for a lot of nouns. A noun in accusative case cannot end up being subject and a noun in dative case cannot mark a possessor.

To summarize, we deal with three kinds of ambiguity: the first one is the diverging explicitness of feature marking in different nominal inflectional paradigms, as discussed for determiners and nouns. This ambiguity can often be resolved by taking agreement into account, which then leads to mutual disambiguation. The second kind of ambiguity is inherent to the morphological system and affects all paradigms alike. I. e. certain distinctions simply cannot be made in the system, e. g. the distinction between genitive and dative feminine singular. The third ambiguity concerns the mapping between case values and grammatical functions. Since a particular case value can signal more than one grammatical function, the final decision between those functions must be made using non-morphological information.

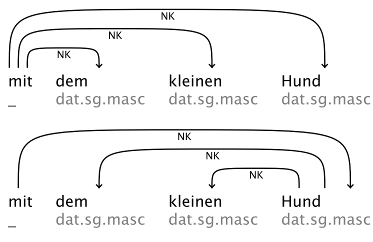

Figure 1: A prepositional phrase in the CoNLL 2009 data and in our version for the phrase *mit dem kleinen Hund* (with the little dog)

3 Data

For our experiments, we use the CoNLL 2009 Shared Task data (Hajič et al., 2009), which has been derived automatically from the German TiGer treebank (Brants et al., 2002). In order to

[3]However, a big group of transitive verbs assigns lexical dative or genitive case to its direct object.

get more consistent data we semi-automatically changed the annotation of prepositional phrases: in the original treebank, prepositional phrases have been analysed as flat structures without an embedded noun phrase. We introduced additional structure as shown in Figure 1 to achieve a consistent annotation for all noun phrases where agreement is represented by direct links labelled by *NK* (noun kernel element). This excludes any effects caused by the otherwise inconsistent annotation of noun phrases and we can evaluate the agreement relation more directly.

4 Evaluation

We evaluate the data-driven dependency parser described in Bohnet (2010), a state-of-the-art second-order maximum spanning tree parser performing second on German in the CoNLL 2009 Shared Task. The parser uses a rich feature model (Bohnet, 2009) and is fully lexicalized. We used statistical tools[4] to automatically lemmatize, part-of-speech tag, and morphologically annotate the training section (36k sentences) of the data by using ten-fold cross annotation. We parsed the whole corpus creating three different models: one using the gold morphology, one using predicted morphology, and one using no explicit morphology. Morphological information was represented as in the CoNLL 2009 Shared Task.

In the following two experiments we test, whether the parser does need explicit morphological information to correctly recover noun phrases and their grammatical function. Since the parser is fully lexicalized, we expect the parser to learn at least some morphological information even when it is not explicitly given.

4.1 Agreement

In order to evaluate how well the parser learns the agreement between a noun and its dependents, we measure the number of times, a parser correctly establishes all links labelled by *NK* between a noun (NN) and its dependent determiners (ART), adjectives (ADJA), and attributive pronouns (PDAT, PIAT, PPOSAT). The total number of these edges is 115,136, the total number of words involved is 206,026. The accuracy of the morphological tagger on gender, number, and case values of these words is 92.79%. Table 1 shows the results in terms of precision, recall, and f-score.

[4] `http://code.google.com/p/mate-tools/`

All three models perform very well and close to each other. Even the model without any explicit morphology achieves an f-score of over 99%. We conclude that lexicalization and configurational information seem to suffice to a large extent for a second-order parser when recovering the internal structure of noun phrases. However, all the differences in f-score between the three models turn out to be statistically significant,[5] so there seems to be a small number of cases where morphology can help in disambiguation. Noun phrases like in Example 1 illustrate these cases where, in principle, arbitrarily many phrases can appear between the determiner and the head noun.

	prec	rec	f1
gold-morph	99.34	99.78	99.56
pred-morph	98.83	99.66	99.24
no-morph	98.77	99.62	99.19

Table 1: Evaluation of NK-edges between ART, ADJA, PIAT, PDAT, PPOSAT, and NN, which represent the agreement relation inside a noun phrase.

4.2 Case – Function Mapping

The second phenomenon we evaluate is the ability of the parser to learn the case – function mapping of German. If the parser is able to learn it, we expect it to only make errors that are related to either the inherent syncretism of the morphological system or to mapping ambiguities between a case value and the functions that it signals.

We evaluate nouns, proper nouns, adjectives, and substituting pronouns (marked for case) for f-score on the functions related to case.[6]

Table 2 shows a clear ranking of the three models:[7] the model using gold morphology outperforms the one using predicted morphology which itself outperforms the third model that uses no explicit morphology. The good performance of the gold morphology model can to a big extent be explained by the fact that in the gold morphology even those ambiguities are resolved that are inherent to the case system of German (see discussion above) and would normally need syntactic or semantic information to be resolved. The biggest difference between the model without morphology and the one using predicted morphology appears for *DA* and *OG*. These two functions are indicated by dative and genitive case respectively. For the

[5] measured on sentence level with a sign test, $\alpha = 0.001$

[6] SB – subject, PD – predicate, OA – accusative object, DA – dative obj., OG – genitive obj., AG – genitive adjunct

[7] All differences are statistically significant except for *PD*, and between pred-m and no-m for *OG*, test see Footnote 5

other functions, the difference in performance is not as high.

	gold-m	pred-m	no-m
SB	95.85	90.85	89.36
PD	76.51	75.40	74.73
OA	94.63	85.22	83.04
DA	88.55	71.59	62.79
OG	58.40	40.98	36.54
AG	96.36	94.05	91.94
total	94.73	88.80	86.82

Table 2: Evaluation of grammatical function assignment to case marked elements (nouns, proper nouns, adjectives, and pronouns) in terms of f-score.

One ambiguity the parser has to deal with is that the same case value can be mapped to two or more different functions. This happens with nominative case, which can be mapped to either *SB* or *PD*, and with genitive case, which can be mapped either to *OG* or *AG*. We expect this ambiguity to pose problems to all models, especially for the one that uses gold morphology. Table 3 shows the fraction of the recall errors for one function where it has been confused with the other possible one. Here we get an interesting picture: while *PD* and *OG* most of the time get confused with their counterpart, the effect is less strong the other way around. Knowing, that *PD* and *OG* are much less frequent than their counterparts may explain the results and gives us a first hint that the mapping learnt by the parser is probably skewed by frequency effects.

SB – PD	23.77	PD – SB	71.55
OG – AG	51.95	AG – OG	2.11

Table 3: Confusion of ambiguous case mappings (in percent) for the model using gold morphology.

The inherent ambiguity of the case system is resolved in the model using gold morphology. We would however expect the model using predicted morphology to additionally have problems to tell apart *SB/PD* (nominative) from *OA* (accusative), and *DA* (dative) from *OG/AG* (genitive).

Table 4 shows the top three functions that the model using predicted morphology confused a function with. For *SB* and *OA* we get the expected picture: without the oracle disambiguation of case, the parser makes the expected errors. For *DA* on the other hand, the parser seems to have problems to recognize the dative as such and so confuses it with *SB* and *OA*, both functions that cannot be marked by dative case. We used a finite state morphology (Schiller, 1994) to annotate every case-bearing word (nouns, determiners, adjectives, pronouns, proper nouns, determined by the automatically assigned part-of-speech tag) with every possible gender, number, and case value that this word form might have. We then disambiguated this annotation further by taking intra-noun phrase agreement into account and found out that 19.63% of the errors could have been fully disambiguated to dative case. This shows that the parser does not learn the mapping between dative case and the label *DA* well enough. A likely reason for that is the lower frequency of *DA*, which occurs 8 times less than *OA*. For *OG*, Table 4 shows a frequent confusion with *AG* and *DA*, which is predicted by the case syncretism in the system.

	1.		2.		3.	
SB	OA	45.86	NK	11.57	PD	9.76
OA	SB	56.82	NK	10.36	CJ	5.17
DA	SB	28.98	OA	20.30	AG	11.88
OG	AG	33.03	DA	21.10	OA	13.76

Table 4: Top three functions that a function has been confused with (in percent) by the model using predicted morphology.

5 Discussion

Recovering the internal structure of a German noun phrase does not seem to pose a big problem for the parser. For most cases, lexicalization and configurational information seem to suffice, although a small portion of the noun phrases can be better disambiguated when explicit feature representations are given.

Good accuracy on the noun phrase's internal structure should then provide a good basis for determining its grammatical function in a broader sentential context because a second-order parser has all the information even though it is distributed on different parts of the phrase (function sharing). However, our second experiment indicates problems for the parser that exceed those caused by inherent ambiguities. A clear sign is the *DA* function, which should only appear with dative case but is frequently confused with other functions that cannot be marked by dative. The low frequency of *DA* might explain the confusion with e.g. *SB* and *OA*, which occur much more often.

Our next steps will include determining an upper bound on gold morphology that is not disambiguated for its inherent syncretism and investigating verbal frames, which may contribute independent information to function selection.

References

Barry J. Blake. 2001. *Case*. Cambridge University Press, Cambridge, New York, 2nd edition.

Bernd Bohnet. 2009. Efficient Parsing of Syntactic and Semantic Dependency Structures. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, volume 2007, Boulder, Colorado. Association for Computational Linguistics.

Bernd Bohnet. 2010. Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 89–97, Beijing, China. Association for Computational Linguistics.

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The TIGER treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, pages 24–41.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 149–164, Morristown, NJ, USA. Association for Computational Linguistics.

Amit Dubey and Frank Keller. 2003. Probabilistic parsing for German using sister-head dependencies. In *Proceedings of ACL 2003*, pages 96–103, Morristown, NJ, USA. Association for Computational Linguistics.

Amit Dubey. 2005. What to do when lexicalization fails: parsing German with suffix analysis and smoothing. In *Proceedings of ACL 2005*, pages 314 – 321, Ann Arbor, Michigan. Association for Computational Linguistics.

Peter Eisenberg. 2006. *Grundriss der deutschen Grammatik: Der Satz.* J.B. Metzler, Stuttgart, 3 edition.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Stepánek, Pavel Stranák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2009. The CoNLL-2009 shared task: Syntactic and Semantic dependencies in multiple languages. In *Proceedings of the 13th CoNLL Shared Task*, pages 1–18, Boulder, Colorado.

Sandra Kübler, Erhard W. Hinrichs, and Wolfgang Maier. 2006. Is it really that difficult to parse German? In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing - EMNLP '06*, page 111, Morristown, NJ, USA. Association for Computational Linguistics.

Sandra Kübler. 2008. The PaGe 2008 shared task on parsing German. In *Proceedings of the Workshop on Parsing German*, pages 55–63, Morristown, NJ, USA. Association for Computational Linguistics.

Anna N. Rafferty and Christopher D. Manning. 2008. Parsing three German treebanks: Lexicalized and unlexicalized baselines. In *Proceedings of the Workshop on Parsing German*, pages 40–46. Association for Computational Linguistics.

Michael Schiehlen. 2004. Annotation strategies for probabilistic parsing in German. In *Proceedings of the 20th international conference on Computational Linguistics*, pages 390–397. Association for Computational Linguistics.

Anne Schiller. 1994. Dmor - user's guide. Technical report, University of Stuttgart.

Reut Tsarfaty, Djamé Seddah, Yoav Goldberg, Sandra Kübler, Marie Candito, Jennifer Foster, Yannick Versley, Ines Rehbein, and Lamia Tounsi. 2010. Statistical parsing of morphologically rich languages (SPMRL): what, how and whither. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 1–12. Association for Computational Linguistics.