Syntactic Annotation of Non-Canonical Linguistic Structures

Hagen Hirschmann and Seanna Doolittle and Anke Lüdeling
Institut für deutsche Sprache und Linguistik
Humboldt-Universität zu Berlin
hagen_h@yahoo.com and seaka@web.de and anke.luedeling@rz.hu-berlin.de

1 Introduction

This paper deals with the syntactic annotation of corpora that contain both 'canonical' and 'non-canonical' sentences.

Consider Examples (1) and (2) from the German learner corpus Falko which will be introduced below. (1) represents a syntactically correct (although perhaps not very enlightening) utterance to which it is easy to assign a syntactic structure. The utterance in (2), on the other hand, would be considered incorrect (and probably be interpreted as a word order error) – it is much more difficult to assign a syntactic structure to it. The question is: how can (1) and (2) be annotated in a uniform way that shows that there is a difference and makes clear exactly where that difference lies?

(1) *Vieles kann man nur mit einem Wort sagen .*
much can one only with one word say
(Much can be said with only one word.)

(2) *Er tatsächlich war sehr wohlhabend gewesen .*
He really was very wealthy been
(He really had been very rich.)

We will not speak about 'grammatical' or 'ungrammatical' utterances here, but rather about 'canonical' and 'non-canonical' utterances. 'Non-canonical' in this paper refers to structures that cannot be described or generated by a given linguistic framework – canonicity can only be defined with respect to that framework. A structure may be non-canonical because it is ungrammatical, or it may be non-canonical because the given framework is not able to analyse it. For annotation purposes the reason for non-canonicity does not matter but for the interpretation of the non-canonical structures, it does. Most non-canonical structures in a learner corpus can be interpreted as errors (Section 2) whereas many non-canonical structures in a corpus of spoken language or computer-mediated communication may be considered interesting features of those varieties.

Many existing syntactically annotated corpora (or treebanks) consist of written language, very often from taken newspapers.[1] While annotation frameworks differ with respect to the underlying theory and the formalism (see Nivre, to appear, for an overview), they make the common assumption that the sentences in the corpora are 'correct' or 'grammatical'.

Language varieties that contain non-canonical as well as canonical sentences such as learner language, spoken language, dialects, the language produced in many computer-mediated communication (CMC) situations, and so forth cannot be directly annotated with the

[1] There are, of course, some treebanks for spoken language such as the CHRISTINE corpus (Sampson 1995, 2003) or TüBa-D/S (Tübinger Baumbank des Deutschen/Spontansprache, http://www.sfs.uni-tuebingen.de/de_tuebads.shtml, Stegmann, Telljohann & Hinrichs 2000) and the parsed Switchboard corpus (http://www.cis.upenn.edu/~treebank/home.html). In recent years parsing of spoken language has become more important (witness the SParseval competition, Roark et al. 2006). We will come back to spoken language in Section 3.

same annotation schemes that are used for 'canonical' treebanks. There are three possible reactions to this:

(a) Write a different grammar or change the grammar that deals with the non-canonical variety at hand.
(b) Ignore the non-canonical utterances by either not annotating the non-canonical structures at all or choosing an inappropriate structure.
(c) Mark the non-canonical sentences as errors and deal with them in a different way.

Solution (a) is the solution that some treebanks for spoken language (such as TüBa-D/S, Stegmann, Telljohann and Hinrichs, 2000), as well as some dialect corpora or historical corpora have taken. It is certainly adequate for many research questions. However, a variety-specific annotation scheme makes a comparison between a canonical and a non-canonical treebank difficult. Solution (b) makes it impossible to do structured searches for non-canonical utterances (see Section 1.1). Solution (c) which is a common solution for learner language, on the other hand, neglects the canonical sentences (see Section 1.2).

Our goal is to develop a syntactic annotation scheme that is able to distinguish between canonical and non-canonical structures and to give adequate descriptions to both.

We want to exemplify this using the Falko corpus which is a learner corpus that consists of texts from advanced learners of German (Lüdeling *et al*., 2005, Siemen, Lüdeling & Müller, 2006).[2] The corpus is stored in a multi-layer model. For our purposes we annotate the corpus with a simple topological structure which is explained in Section 2.2. Section 2.3 then deals with the problem of applying the topological structure model to the learner data. We propose that it is necessary to formulate a 'target hypothesis' against which the non-canonical utterances can be annotated. In the remainder of this section, we first want to show how other treebank schemes deal with non-canonical data (Section 1.1) and then discuss error tagging (Section 1.2). In our examples we focus on German corpora but the problems we describe are not language specific.

1.1 Non-canonical syntactic structures in German corpora

As stated above, many treebanks contain written language data of a fairly standardized variety (often newspaper data). As a result only canonical sentences are expected in these corpora. Therefore the annotation schemes often do not anticipate the problem of non-canonical utterances.

However, even newspapers contain utterances (sentences, phrases, word forms *etc.*) that cannot be regarded as canonical. These utterances are not necessarily ungrammatical, but sometimes they are not well-formed with regard to the syntactic annotation scheme of the corpus. The basic problem is that whenever a non-canonical utterance occurs, the annotation scheme does not provide adequate means of describing it.

Owing to these facts the annotator can

(a) try to find the best-fitting description for the utterance. This means that certain elements may not be tagged appropriately, as we see in Figure 1 where an equation is assigned a sentence structure.
(b) skip the annotation of the structure or do only a partial parse. That means the problematic structures are (often) syntactically isolated from the rest of the sentence. This is illustrated by Figure 2 where one of the constituents is not connected. Those structures are not integrated into the syntactic structure and in most cases they cannot

[2] The corpus is available at http://www2.hu-berlin.de/korpling/projekte/falko/.

directly be searched for in the corpus. The information that an utterance cannot be annotated based on the underlying model is given only implicitly.

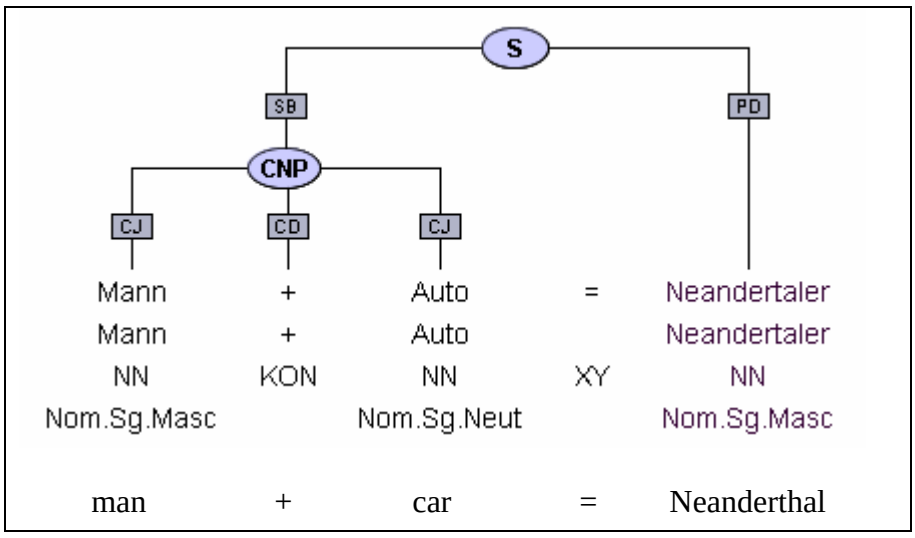

Figure 1: The equation is annotated with a sentence node label (S): *Mann + Auto* (man + car) is given the function of a subject (see the SB edge label) and *Neandertaler* is labelled as a predicate (PD).
(Tiger-corpus, release 2005, http://www.ims.uni-stuttgart.de/projekte/TIGER/)

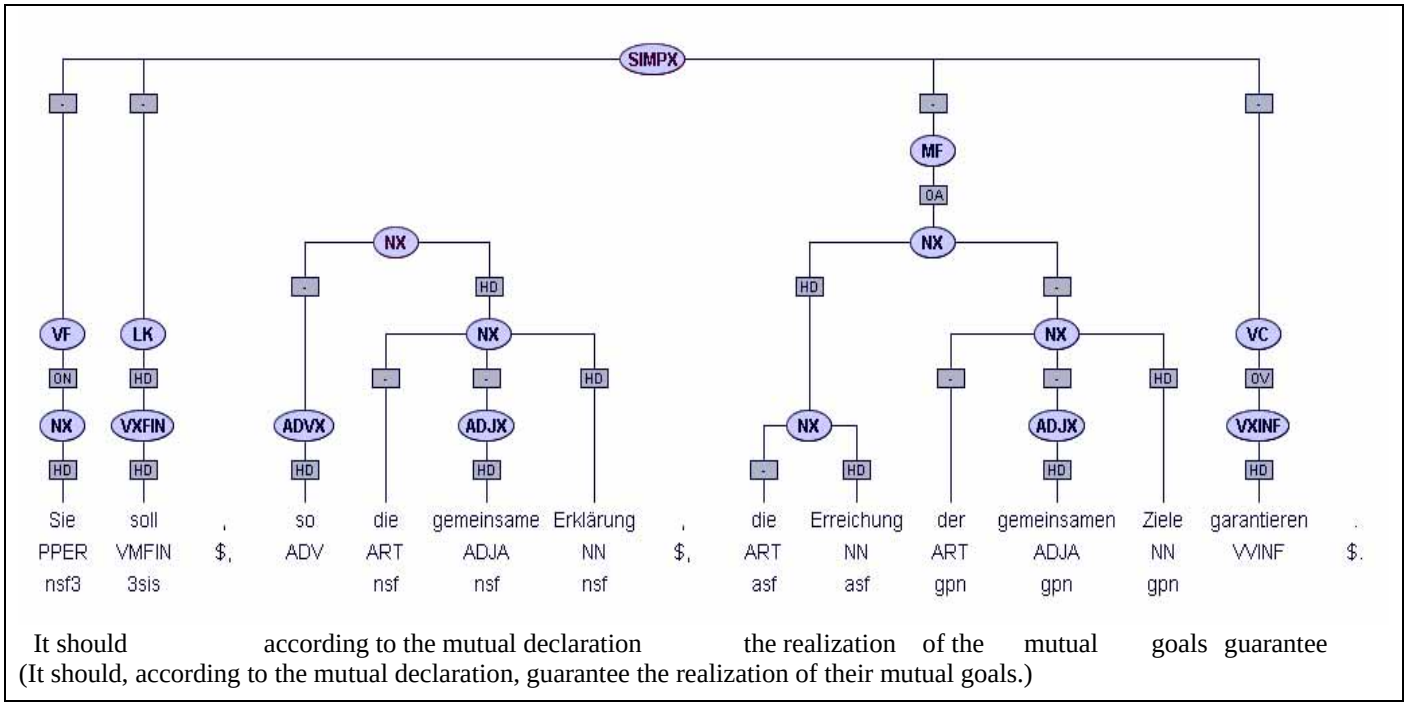

Figure 2: The parenthesis, "*so die gemeinsame Erklärung*" ("according to the mutual declaration"), is not integrated into the sentence because it cannot be assigned to a topological field.[3]
(TüBa-D/Z corpus www.sfs.uni-tuebingen.de/resources/sty.ps)

[3] Topological fields are introduced in Section 2.1. Note that in this case the parse is entirely correct. The problem is that it cannot be formally distinguished from cases where elements are left unintegrated which are not correct.

A third option would be to mark the structure as not describable with reference to the underlying annotation scheme. We could not find a German treebank where this is done.

While the utterances above contain grammatical (or acceptable) structures that cannot be described by the annotation schemes, Figures 3 and 4 contain ungrammatical utterances. As we said above, ungrammaticality is just one of the reasons for non-canonicity. In both cases the noun has the wrong inflection (*Schülern* instead of *Schüler* (pupils) and *Haushaltsjahrs* instead of *Haushaltsjahr* (financial year)) Using the annotation scheme for canonical structures, the annotator has two choices: annotate a grammatical structure and ignore the wrong case (as shown in Figure 3), or annotate the inappropriate case (as shown in Figure 4):

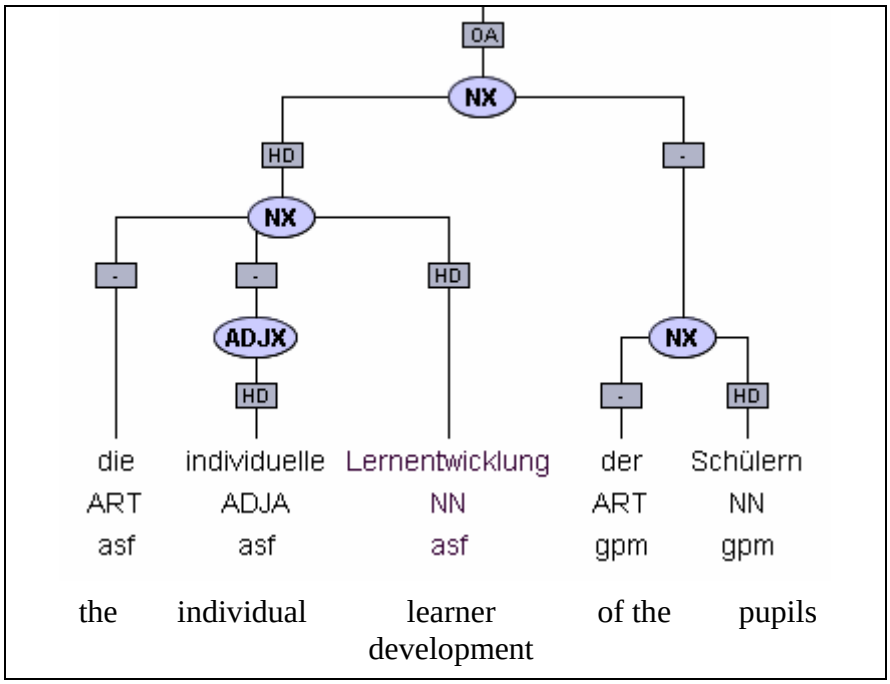

Figure 3: The noun phrase *der Schüler* (of the pupils) is a genitive attribute, whereas *Schülern* is dative case. The morphological tag "gpm" (genitive plural masculine) marks the noun as genitive, corresponding to the syntactic function of the noun phrase, but ignoring the morphological form.
(TüBa-D/Z corpus, www.sfs.uni-tuebingen.de/resources/sty.ps)

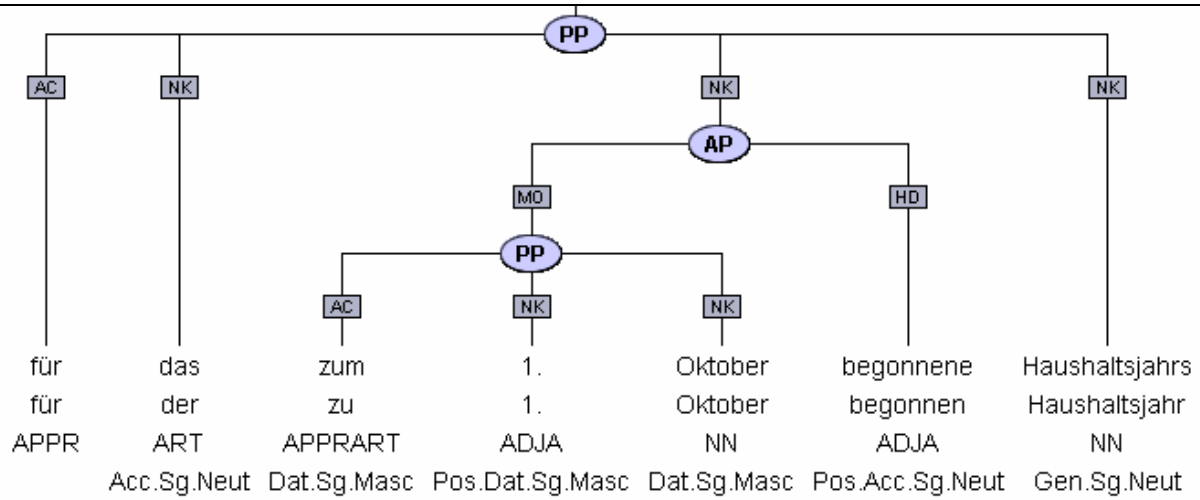

Figure 4: The preposition *für* (for) calls for the accusative case. *Haushaltsjahr*s* (financial year), however, is genitive. This is expressed by the morphological tag "Gen.Sg.Neut" (genitive singular neuter). (It is not possible to have different “NK”-elements (noun kernel elements) with different cases in one phrase).
(Tiger-corpus, release 2005, http://www.ims.uni-stuttgart.de/projekte/TIGER/)

No matter what the annotator decides to do, the annotation scheme is violated. In Figure 3 the morphological annotation is incorrect and in Figure 4 the syntactic annotation is inconsistent. Furthermore, in both cases the ungrammatical structure cannot easily be found in the corpus, although structures like these could be of special interest.

Figures 1-4 show that even in corpora which are expected to exclusively contain canonical material, non-canonical structures can be found. Some structures are grammatical but still do not fit into the scheme, while others are ungrammatical. In addition, it is usually not possible to specifically search for non-canonical structures in these corpora.

1.2 Error annotation

A different approach for interpreting corpus data is taken in the annotation of learner data. Research in this area does not focus on canonical structures but rather on errors (*i.e.* non-canonical structures) because they provide insights into acquisition strategies and hypotheses of the learner. Therefore learner corpora are often error tagged (see Granger 2002 for an overview).

In existing learner corpora, error analysis is usually based on a pre-defined error tagset (the granularity and scope of error tagsets differ significantly). The tags are assigned to the erroneous words (or sequences of words).

(3), taken from Weinberger (2002), shows a word order error. The complex error tag is inserted before the wrong element (or sequence). <GrVrWoMa> is the tag for a grammatical error affecting the verb and its word order in the main clause (Gr = grammar; Vr = verb; Wo = word order; Ma= main clause).

(3) **Zum Beispiel sie* **<GrVrWoMa>*sind*** *ein bißchen rebellisch …*
(for example they **are** a little rebellious)

In error-tagged corpora a systematic search for different types of errors is possible. However, error-tagged corpora usually do not contain parses for canonical utterances.

1.3 Combining syntactic analysis and deviation analysis

We showed that in 'canonical' treebanks it is not possible to adequately search for the non-canonical structures and that in error-tagged corpora one cannot usually do a search within the canonical structures. Since there are many corpora that contain both canonical and non-canonical structures we argue that an annotation scheme should combine the advantages of both annotation schemes in the same corpus: This can only be achieved by a corpus architecture which contains different independent levels of annotation – as we will show later, three annotation levels are needed to annotate both the canonical and the non-canonical syntactic-topological structures in a corpus. The first level being annotation of all canonical structures, the second level of analysis is the formulation of a target hypothesis and the third level is error tagging based on the target hypothesis, so that it can be seen, what exactly makes the sentence not describable.

The advantages of separating these three levels of annotation are: first, the ability to compare the canonical structures in the corpus with canonical structures in other corpora (other varieties, languages, dialects *etc.*) and second, the option to make qualitative and quantitative analyses of the deviation from the underlying model. To do this the deviations first have to be categorized as non-canonical structures (they simply can not be described with the underlying model represented by the annotation scheme). Depending on the model and the reason for not fitting into that model, deviations can be categorized differently. In a learner corpus they will mostly be classified as errors, in a spoken language corpus they could be analysed as properties of a spoken register.

We demonstrate our scheme using the learner corpus Falko annotated on the basis of the topological field model.

2. A case study: Annotation of word order in German

We have chosen the annotation of word order as one aspect of syntax or as one component of syntactic annotation in order to illustrate a multi-layer syntax annotation of canonical and non-canonical utterances in the same corpus. The model is simple and easy to implement and can be annotated in a linear fashion but can, in essence, only describe verb placement errors (it has nothing to say about the order of components in the middle field or word order inside components). The general argument, however, carries over to trees or graphs.

2.1 Modelling the linear sentence structure of German

There are two important factors that a model depicting German word order must cope with. First of all unlike English, as a rule[4], German word order in main clauses (SVO) and subordinate clauses (SOV) differs.[5]

In general, German is considered to be a language with (fairly) flexible word order. But the finite verb has a fix position in the sentence. Its position is used to describe the three classes of German sentences – namely: (4) verb second (e.g. main clause), (5) verb first (e.g. yes-no questions) and (6) verb last sentences (e.g. subordinate clauses).

[4] which, of course, was made to be broken.
[5] For a general overview, see for example Comrie 1981, Chapter 4.

(4) *Das Kind* ***isst*** *Erbsen .*
fin. verb
The child eats peas
(The child is eating peas)

(5) ***Isst*** *das Kind Erbsen ?*
fin. Verb
Eats the child peas ?
(Does the child eat peas?)

(6) *..., dass das Kind Erbsen* ***isst.***
fin. verb
..., that the child peas eats.
(..., that the child eats peas.)

The second factor that has to be dealt with is split constituents, the German verbal group being the most predominant example. The verb complex (finite verb and other verbal arguments like infinite verb and verb particles) does not necessarily form a linear unit in the sentence and hence a type of verbal bracket (*Satzklammer*) is created as illustrated in (7).

(7) *Das Kind* ***hat*** *Erbsen* ***gegessen.***
fin. verb **inf. Verb**
The child has peas eaten
(The child has eaten peas)

The topological field model (Drach 1937, Höhle 1986) has proven quite useful in describing these features. In this model, the two possible positions of the verbal components namely the left bracket and the right bracket form the cornerstones or boundaries for the division of the sentences into fields. In verb second sentences, for example, which are in most cases declarative sentences (statements), up to three fields can be formed. Figure 5 illustrates this for example (7). The initial field is located left of the finite verb in the left bracket. The middle field can be found directly on the right of the left bracket and the final after the right bracket on the right side. In our example this field is empty.

initial field	**left bracket**	middle field	**right bracket**	final field
Das Kind (the child)	hat (has)	Erbsen (peas)	gegessen (eaten)	[empty]

Fig. 5 . topological field diagramm for a main clause

Although there are some restrictions concerning what kind of and how many constituents may occupy these fields, there is still a high degree of positional flexibility.

The topological field model is a widely used descriptive model for German word order and numerous phrase-based generative analyses of German build on it (Grewendorf, Hamm and Sternefeld 1987). These are good reasons for its use as a model for annotation that can be reproduced by different annotators and meets with the annotation standard of consensual analyses.

2.2 Description of Falko's syntactic field annotation:

This section shows how the topological field model is used to annotate the Falko corpus[6]. The multi-layer architecture of the corpus (Lüdeling *et al.* 2005) enables us to assign more than one tag to a token or token group, making it possible to segment the text into token groups which can be labelled at multiple levels.[7] Consider Figure 6 where (1) is presented in a multi-layer table.

The [word] level is the electronic reproduction of the learner's text. It constitutes the tokenized corpus. The following two rows represent a simplified version of our topological field annotation. The utterance is identified and marked with an "x" at the [utterance] level. In the next level [top. fields] the topological fields are tagged.

[word]	Vieles Much	kann can	man man	nur only	mit with	einem a	Wort word	sagen say	.
[utterance]	x								
[top.-fields]	initial field	left bracket	middle field					right bracket	

Figure 6: Example for a topological field annotation of a canonical utterance in a multi level corpus architecture

The elements left of the finite verb are tagged as the initial field of the main clause. As a rule, only one constituent can occupy the initial field but further elements can be located in front of the left verbal bracket and there are many different approaches for naming and classifying these elements (*cf.* Hoberg 1997 and Pasch 2003).

The field immediately following the finite verb is the middle field. This field can consist of more than one constituent and there is a fair amount of flexibility in the word order.

In our example, the verbal complex has two elements and, as mentioned above, the infinitive verb form in main clauses (verb second structures) is defined as the right sentence bracket.

As can be seen in this example not all fields must be occupied. A final field has not been annotated – which in the literature is often seen as a field for extraposition of longer sentence elements, for example subordinate clauses.

Using this method, it is possible not only to search for sentences and fields, but since each annotation layer implemented in our corpus is aligned with the other layers it is possible to search for elements/structures in specific syntactic-topological contexts.

For example, by taking the part of speech-level into consideration, it would also be possible to research further features at the sentence and field levels. Not only can complexity be measured by the sentence length or the number and types of subordination, but also by the complexity and contents of the topological fields.

2.3 Annotating non-canonical word order structures

After the brief introduction of how canonical topological field structures of German can be annotated we show how non-canonical structures are annotated. Figure 7 exemplifies the problem. This utterance does not correspond with the German topological field model, because there are two constituents in the initial field: *Er* (subject) and *tatsächlich* (adverbial)

[6]Large parts of the Falko corpus are annotated according to a (slightly more complex) scheme.

[7]The annotation tool we use for Falko is EXMARaLDA (Schmidt 2004).

which means (in accordance to the topological field model) the infinitive verb is not in its obligatory verb second position.

[word]	Er *he*	tatsächlich *really*	war *was*	sehr *very*	wohlhabend *wealthy*	gewesen *been*	.
[utterance]	*x*						
[top. field annotation of **utterance**]	f_ = non-canonical (annotation not possible)						

Figure 7: Example for a non-canonical utterance with a topological field annotation scheme

This problem holds true for every non-canonical structure – when, for whatever reason, it cannot be explained by the (grammatical) model on which the annotation scheme is based. All that can be done at the annotation level which describes canonical structures is tag the structure that does not fit as non-canonical.

2.3.1 Target hypothesis and error annotation

Analysing an error (a non-canonical utterance) always involves saying something about its deviation from the corresponding "correct" (or canonical) structure. If this relationship is not taken into consideration, nothing can be said about the error – not even that it is non-canonical.

To be able to measure this deviation, the corresponding canonical structure has to be formulated. Often different readings and consequently different ways of annotating a non-canonical sentence are possible (see Corder 1981 and Lüdeling, to appear, for a discussion). So first, we have to predefine what the corresponding canonical structure of the non-canonical sentence is. We call this assumption target hypothesis. It determines the annotation of the non-canonical structures and provides the link between the learner sentence and the "error annotation".

In regard to topological aspects (as well as to other grammatical aspects), the target hypothesis gives an implication of where certain elements cannot be placed in accordance with the underlying model.

The target hypothesis has to refer precisely to the non-canonical structures in the learner text. In order to make the target hypothesises as reliable as possible, we align it as close as possible to the learner text– word by word.

As can be seen in Figure 8, the canonical structures are duplicated in the target hypothesis level. In this case the tokens are matched.

[word]	Vieles	kann	man	nur	mit	einem	Wort	sagen
[target hypothesis]	Vieles	kann	man	nur	mit	einem	Wort	sagen

Figure 8: Example for the annotation layer "target hypothesis", tagging a canonical utterance of a learner in the Falko corpus (http://korpling.german.hu-berlin.de/falko/, subcorpus "Falko-Zusammenfassungen 1.0")

Divergences of the learner text to the target hypothesis directly indicate non-canonical structures.

If an utterance is non-canonical, there are three different possibilities of how a token (word) in the target hypothesis can deviate from the surface of the learner text:

1. A token is deleted.
2. A token is inserted.
3. A token is substituted.

Sentences with non-canonical topological structures that are "corrected" will mostly contain the options 1. and 2., because words or phrases are reordered, which means they are deleted at their original position and inserted at another:

[word]	Er		tatsächlich	war	sehr	wohlhabend	gewesen	.
[target hypothesis]	Er (he)	war (was)	tatsächlich (really)		sehr (very)	wohlhabend (wealthy)	gewesen (been)	.
description of deviation		token inserted		token deleted				

Figure 9: Example for the annotation layer "target hypothesis", tagging a non-canonical utterance of a learner in the Falko corpus (http://korpling.german.hu-berlin.de/falko/, subcorpus "Falko-Georgetown")

In order to make this learner utterance canonical, the verb must be placed directly after the first constituent *Er* in a verb second position.

As might be evident from this example, there are different possibilities for alignment but it is standard to define such learner structures as verb placement errors. So, in order to illustrate this, the verb (and not for example *tatsächlich*) is deleted at its original non-canonical position and it is inserted at its canonical position.

2.3.2 Interpretation of the deviation

Figure 10 illustrates how word order errors can be described based on the topological field annotation of the target hypothesis. By using the field annotation of the target hypothesis as a template that is placed over both structures, a possible way of describing the error would be to say that the finite verb is erroneously positioned in the targeted middle field (deletion) but it should be located in the left verbal bracket (insertion).

[word]	Er		tatsächlich	**war**	sehr	wohlhabend	gewesen	.
[target hypothesis]	Er	**war**	tatsächlich		sehr	wohlhabend	gewesen	.
description of deviation		token inserted		token deleted				
top. field annotation of **target hypothesis**	initial field	left bracket	middle field				right bracket	

Figure 10: Topological field annotation of a target hypothesis, aligned to a non-canonical utterance

3. Non-canonical structures in other contexts

Learner corpora might be an obvious example of texts that contain both canonical and non-canonical structures. But many other varieties are similar, although, the specific 'deviations' of course might differ. In these varieties, the non-canonical structures are not 'errors' but interesting and characteristic properties.

In this section, we briefly show that our annotation scheme might be very helpful in annotating these other varieties as well. We will use examples from spontaneous spoken language and computer-mediated communication.

3.1 Spontaneous spoken language

Spoken language syntactically differs from written language in many ways (for a thorough discussion of features of spoken German see Schwitalla 2006). These differences are sometimes qualitative (there are structures that occur only in written registers and structures that occur only in spoken registers) and sometimes quantitative (some structures occur markedly more often in one of the registers than in the other). As stated above, treebanks for spoken language often develop their own annotation schemes (the most specific one is probably the CHRISTINE scheme[8]) and this might well be necessary to cover phenomena such as hesitations, self-corrections and the like. Schemes like the TüBa-D/S or CHRISTINE typically mark elements that are syntactically unconnected as such and do not attach everything to a single top node. Again, this might be the most appropriate way of annotating spoken language. There are two problems with this, however: First, unconnected elements like hesitations, interjections *etc.* that are very typical of spoken registers cannot be formally distinguished from unconnected elements like the parenthesis in Figure 2 which is very typical of written registers. And second, it is difficult to systematically describe the differences between written and spoken registers in a precise way if the structures cannot be mapped onto each other.

One of the structures that is always listed as typical for spoken language is the ellipsis (Schwitalla 2006) which is illustrated in Figure 12 which stems from a dialogue between a mother and her daughter[9]. The mother complains that her daughter always uses the parents' bathroom and takes the parents' towels *etc.* From *deine* to *fehlen* the utterance in (8) can be described by the regular field model but the utterance that immediately follows in (9) does not fit into the model because there is no finite verb and because of this no bracketing structure can be assigned. The annotation of (8) is unproblematic, as shown in Figure 11. (9), on the other hand, can only be annotated after a target hypothesis is formulated, as shown in Figure 12.

(8) *deine handtücher die kannste aus=m schrank holen wenn dir welche fehlen*
(You can take towels from the closet if you need them)

[8] See Sampson (1995) and http://www.grsampson.net/ChrisDoc.html.

[9] The corpus dialogues between mothers and their daughters about controversial topics was collected in the Sonderforschungsbereich 245: "Sprechen und Sprachverstehen im sozialen Kontext" in Heidelberg and Mannheim between 1988 and 1992. More information and some of the data are available at http://www.ids-mannheim.de/ksgd/agd/korpora/ekkorpus.html. The transcription is generally in lower case.

[word]	deine	handtücher	die	kannste	aus=m	schrank	holen	wenn	dir	welche	fehlen
[target hypothesis]	deine (your)	handtücher (towels)	die (them)	kannste (can+you)	aus=m (out+of+the)	schrank (closet)	holen (take)	wenn (if)	dir (you)	welche (some)	fehlen (lack)
[top. field annotation of **utterance**]	initial field			left bracket	middle field		right bracket	final field			

Figure 11: Topological field annotation of a canonical utterance from a corpus of spoken German

(9) *aber unsre in ruh lassen okay*
(But leave ours alone, ok?)

[word]	aber	unsre			in	ruh	lassen	okay
[target hypothesis]	aber (but)	unsre (ours)	sollst (should)	du (you)	in (in)	ruh (peace)	lassen (leave)	okay (okay)
description of deviation			token inserted	token inserted				
[top. field annotation of **target hypothesis**]	(con-junction)	initial field	left bracket	middle field			right bracket	final field

Figure 12: Annotation of a (topologically) non-canonical utterance from a corpus of spoken German. This figure shows, in analogy to the method used in Figure 10, that two elements have to be inserted to conform with the underlying syntactic scheme, namely, the finite verb *sollst* (should), and the subject *du* (you). The deviation could be defined as a missing targeted left bracket and a missing element in the targeted middle field.

Most treebanks for spoken corpora might annotate (9) simply as elliptical. Then one could not show what exactly was missing. Our annotation complies with theoretical accounts of ellipsis (*cf.* Klein, 1993: 768) which state that in elliptical structures are syntactically complete but lack only phonetic material. But even if one does not share this analysis – the annotation against a target hypothesis makes it possible to search for the exact types and location of omissions.

In a corpus annotated like this, it is also possible to quantitatively compare features of spoken language to features of written language.

3.2 Computer-mediated communication

It is very often said that computer-mediated communication (CMC) is positioned somewhere between spoken registers and written registers (Beißwenger & Storrer, to appear). Many papers on CMC focus on specific features such as the use of inflectives or emoticons, others calculate quantitative differences. We are not aware of any large-scale study of the CMC's syntax (or even of syntactically annotated corpora of CMC). The following examples (10) and (11) again show a passage that is partly canonical and partly non-canonical – in

analogy to the examples (8) and (9). They stem from a forum discussion about a computer game[10]. Sentence (10) is fully canonical while the expression in (11) is again elliptical and cannot be directly assigned a field structure.

(10) *Wenn es unbedingt sein muss kann ich ja noch mal neu anfangen*
(If it is absolutely necessary I can start over again)

(11) *Ok … erst Level 10*
(Ok … first level 10)

As shown in the spoken language corpus (examples (8) and (9)), it is again possible to construct a target form for the non-canonical utterance in example (11):

[word]	Ok	erst			Level	10	
[target hypothesis]	Ok	erst (first)	muss (must)	ich (I)	Level (level)	10	machen (do)
description of deviation			token inserted	token inserted			
[top. field annotation of **target hypothesis**]	(discourse marker)	initial field	left bracket	middle field			right bracket

Figure 13: Annotation of a (topologically) non-canonical utterance from a corpus of CMC

The question of how CMC is influenced by oral registers or written registers can be answered once a CMC corpus has been annotated with the proposed scheme. It can then be compared, qualitatively and quantitatively, to other field-annotated corpora.

4. Summary

In this paper, we argued for a generalized annotation scheme for canonical and non-canonical sentences if they appear in the same corpus. We define canonicity as 'conformity with a specific annotation scheme'. We showed that many existing treebanks schemes are not prepared to deal adequately with non-canonical structures. The options open to the annotator who finds a structure that cannot be described with the scheme are either to use an inappropriate structure or to only perform a partial parse. Neither option leads to annotations that can be systematically searched when one wants to specifically look at non-canonical structures of a given type. In error-tagged corpora (for example learner corpora), on the other hand, non-canonical structures can easily be identified; but error-tagged corpora usually do not provide tagging for the canonical structures. We argue that it is important for many linguistic questions to (a) distinguish between canonical and non-canonical structures and (b) show how the non-canonical structures do not conform to the canon.

Our annotation scheme works in three steps. First, we annotate all canonical sentences within the syntactic model. In the second step, we provide a target hypothesis for all non-canonical sentences. The target hypothesis is a structure that corresponds as closely as

[10] From http://www.worldofgothic.de

possible to the original non-canonical structure and can be described by the model. Then we annotate the differences between the non-canonical structure and the target hypothesis.

The same general scheme can be used for different varieties. The interpretation of the deviations from the canonical structure is a further step that depends on the variety at hand and on the research question. In learner language, a deviation might be analysed as an error, in other varieties it might be analysed as a feature.

A corpus annotated like this provides a means for quantitative as well as qualitative research. Non-canonical structures can be compared to canonical structures in the same corpus or to other structures in different corpora.

References

Beißwenger, M. and A. Storrer (to appear) Corpora of computer-mediated communication, in A. Lüdeling and M. Kytö (eds) Corpus Linguistics. An International Handbook. Berlin: Mouton de Gruyter.

Comrie, B. (1981) Language Universals and Linguistic Typology. Oxford: Basil Blackwell.

Corder, S. P. (1981) Error Analysis and Interlanguage. Oxford: Oxford University Press.

Granger, S. (2002) A bird's-eye view of learner corpus research, in J. Hung and S. Petch-Tyson (eds) Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching, pp. 3-33. Amsterdam: John Benjamins.

Drach, E. (1937) Grundgedanken der deutschen Satzlehre. Frankfurt a. M.: Diesterweg.

Grewendorf, G., F. Hamm and W. Sternefeld (1987) Sprachliches Wissen. Eine Einführung in moderne Theorien der grammatischen Beschreibung. Frankfurt a. M.: Suhrkamp.

Hoberg, U. (1997) Die Linearstruktur des Satzes, in: G. Zifonun, L. Hoffmann and B. Strecker (eds) Grammatik der deutschen Sprache, pp. 1495-1680. Berlin: de Gruyter.

Höhle, T. (1986) Der Begriff „Mittelfeld“. Anmerkungen über Theorie der topologischen Felder, in M. Reis *et al.* (ed.) Akten des VII. Kongresses der Internationalen Vereinigung für germanistische Sprach- und Literaturwissenschaft, pp. 329-340. Tübingen: Niemeyer.

Klein, W. (1993) Ellipse, in: J. Jacobs *et al.* (eds) Syntax. An International Handbook of Contemporary Research. Vol. 1. Berlin and New York: Walter de Gruyter.

Lüdeling, A., M. Walter, E. Kroymann and P. Adolphs (2005): Multi-level error annotation in learner corpora, in: *Proceedings of Corpus Linguistics 2005*. Birmingham.
Available on-line from http://www.corpus.bham.ac.uk/PCLC/ (accessed: 28 June 2007)

Lüdeling, A. (to appear) Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora, in P. Grommes and M. Walter (eds) Fortgeschrittene Lernervarietäten. Tübingen: Niemeyer.

Nivre, J. (to appear) Treebanks, in A. Lüdeling and M. Kytö (eds) Corpus Linguistics. An International Handbook. Berlin: Mouton de Gruyter.

Pasch, R., U. Braußе, E. Breindl and H.U. Waßner (2003): Handbuch der deutschen Konnektoren. Linguistische Grundlagen der Beschreibung und syntaktische Merkmale der deutschen Satzverknüpfer (Konjunktionen, Satzadverbien und Partikeln). Berlin: Walter de Gruyter.

Roark, B., Y. Liu, M. Harper , R. Stewart, M. Lease, M. Snover. and I. Shafran (2006) SParseval: Evaluation Metrics for Parsing Speech, in *Proceedings of the International conference on Language Resources and Evaluation (LREC-2006)*, Genoa.
Available on-line from http://bllip.cs.brown.edu/publications/Matt_Lease.shtml (accessed: 28 June 2007)

Sampson, G. (1995) English for the Computer. The SUSANNE Corpus and Analytic Scheme. Oxford: Oxford University Press.

Sampson, G. (2003), Thoughts on Two Decades of Drawing Trees. In: Abeillé, A. (ed.), *Treebanks: Building and Using Parsed Corpora*, pp. 23–41, Dordrecht: Kluwer,.

Schmidt, T. (2004) Transcribing and annotating spoken language with EXMARaLDA, in: Proceedings of the LREC-Workshop on XML based richly annotated corpora, Lisbon 2004. Paris: ELRA.
Available on-line from http://www1.uni-hamburg.de/exmaralda/ (accessed: 28 June 2007)

Schwitalla, J. (2006) Gesprochenes Deutsch. Eine Einführung. Berlin: Erich Schmitt Verlag.

Siemen, P., A. Lüdeling and F. H. Müller (2006) FALKO - ein fehlerannotiertes Lernerkorpus des Deutschen, in: *Proceedings of Konvens 2006*, Konstanz.
Available on-line from http://ling.uni-konstanz.de/pages/conferences/konvens06/konvens_files/abstracts/siemenetal.pdf (accessed: 28 June 2007)

Stegmann, R., H. Telljohann and E. Hinrichs (2000) Stylebook for the German Treebank in VERBMOBIL. Technical Report 239. Verbmobil.
Available on-line from http://www.r-stegmann.de/dr-rosmary-stegmann/veroeffentlichungen-und-vortraege/ (accessed: 28 June 2007)

Weinberger, U. (2002) Error Analysis with Computer Learner Corpora. A corpus based study of errors in the written German of British University Students. MS Thesis. Lancaster University.