A Diagnostic Tool for German Syntax [*]

John Nerbonne, Klaus Netter
Abdel Kader Diagne, Judith Klein[†] and
Ludwig Dickmann[‡]

[†] Deutsches Forschungszentrum für Künstliche Intelligenz, GmbH
Stuhlsatzenhausweg 3, D-6600 Saarbrücken 11, FRG
phone: (+49 681) 302-5300
e-mail: nerbonne@dfki.uni-sb.de

and

[‡]Institut für Computerlinguistik, Universität des Saarlandes
Im Stadtwald, D-6600 Saarbrücken 11, FRG

Abstract

In this paper we describe an ongoing effort to construct a catalogue of syntactic data exemplifying the major syntactic patterns of German. The purpose of the corpus is to support the diagnosis of errors in the syntactic components of natural language processing (NLP) systems. Secondary aims are the evaluation of NLP syntax components and support of theoretical and empirical work on German syntax.

The data consist of artificially and systematically constructed expressions, including also negative (ungrammatical) examples. The data are organized into a relational database and annotated with some basic information about the phenomena illustrated and the internal structure of the sample sentences. The organization of the data supports selected systematic testing of specific areas of syntax, but also serves the purpose of a linguistic database.

The paper first gives some general motivation for the necessity of syntactic precision in some areas of NLP and discusses the potential contribution of a syntactic database to the field of component evaluation. The second part of the paper describes the set up and control methods applied in the construction of the sentence suite and annotations to the examples. We illustrate the approach with examples from verbal government and sentential coordination. This section also contains a description of the abstract data model, the design of the database and the query language used to access the data. The final sections compare our work to existing approaches and sketch some future extensions.

We invite other research groups to participate in our effort, so that the diagnostics tool can eventually become public domain. Several groups have already accepted this invitation, and progress is being made.

[*]This work was supported by a research grant, ITW 9002 0, from the German Bundesministerium für Forschung und Technologie to the DFKI project DISCO and by IBM Germany through the project LILOG-SB conducted at the University of Saarbrücken. These projects are led by Hans Uszkoreit. We are grateful to Andrew White, Trinity College, Dublin, and Francesco Infantino and Claudio Piorenza, both of Universiteta di Bari, for support programming in the context of student projects.

1 Introduction

This paper describes an ongoing effort to construct a catalogue of syntactic data which is intended eventually to exemplify the major syntactic patterns of the German language. Our purpose in developing the catalogue and related facilities is to obtain an empirical basis for diagnosing errors in natural language processing systems analyzing German syntax, but the catalogue may also be of interest to theoretical syntacticians and to researchers in speech and related areas. The data collection differs from most related enterprises in two respects: (i) the material consists of systematically and artificially constructed sentences rather than naturally occurring text, and (ii) the material is annotated with information about the syntactic phenomena illustrated which goes beyond tagging parts of speech. The catalogue currently treats verbal government, sentential coordination, fixed (semi-idiomatic) verbal structures (*Funktionsverbgefüge*), and relative clauses.

The data consists of linguistic expressions (mostly short sentences designed to exemplify one syntactic phenomenon) together with annotations describing selected syntactic properties of the expression. The annotations of the linguistic material serve (i) to identify and label construction types in order to allow selected systematic testing of specific areas of syntax, e.g., sentential coordination; and (ii) to provide a linguistic knowledge base supporting the research and development of natural language processing (NLP) systems. Besides this information, the annotations contain information about the precise structure of the sentence such as the position of the finite verb (e.g., as 'fifth word') and the positions of other phrases. The purpose of including the latter sort is that it represents very concrete data against which to test accuracy.

In order to probe the accuracy of NLP systems, especially the detection of unwanted overgeneration, the test material includes not only genuine sentences, but also some syntactically ill-formed strings.

The syntactic material, together with its annotations is being organized into a relational database in order to ease access, maintain consistency, and allow variable logical views of the data. The database system is in the public domain and is (mostly) independently supported.

We are publishing this work—both the test material and the database of annotations in order to interest potential contributing partners; later we shall share the corpus and tools with the general research and development community.

2 Goals of a Diagnostic Tool

Our goal in collecting and annotating syntactic material is to develop a diagnostic tool for natural language processing systems, but we believe the material may be of interest to other researchers in natural language, particularly syntactic theoreticians. Finally, although this is not an evaluation tool by itself, our work points to possiblities for evaluating systems of syntactic analysis by allowing the systematic verification of claims about, and investigation of, the coverage and precision of systems.

2.1 Natural Language Processing

There is general consensus, both in theoretical computational linguistics and in practical, industrially sponsored research in natural language processing, that systems for syntactic analysis (parsing, classification and recognition) are possible and valuable. The applications of syntactic analysis currently under investigation include grammar and style checking; machine translation; natural language unterstanding (particularly interfaces to databases, expert systems, and other software systems); information retrieval; automatic indexing; speech synthesis; and speech recognition. The potential impact of syntactic analysis technology is technically and financially profound.

But if we are to realize the full benefits of syntactic analysis, then we must ensure that correct analyses are provided. The development of a diagnostic tool serves just this purpose—pointing out where analyses are correct, and where incorrect. There are, of course, other measures of quality which apply to natural language software, e.g., general software standards. Systems which perform syntactic analysis are naturally subject to the same general standards of software quality that are imposed throughout the software engineering field, e.g., efficiency, modularity, modifiability, compatibility, and ease of installation and maintenance. Special-purpose systems may be subject to further standards; e.g., interface software is generally required to have clear and intuitive boundaries (transparency). Compared to such general software standards, correctness of syntactic analysis is an orthogonal criterion, though for many applications, an overriding one. Attending exclusively to general software standards means risking incorrectness—whether this be incorrectness of matrix multiplication in a linear algebra package or misanalyses in a natural language parser. The ultimate costs of such misanalysis depend, of course, on the particular application, but these costs may easily outweigh the benefits of the system deployed.

The importance of precision in syntactic analysis is occasionally disputed. It is pointed out, for example, that humans make speech errors (and typos), and that natural language understanding systems will have to be sufficiently robust to deal with these. Here, it is claimed, less precise systems may even have an advantage over more exact, and hence "brittle" competitors. What is correct about this point is that systems should be able to deal with ill-formed input. What is questionable is the suggestion that one deal with it by relaxing syntactic or other constraints *generally* (although it might be quite reasonable to use constraint relaxation where no exact analysis may be found—as a processing strategy).

The problem with general constraint relaxation is that it inevitably involves not only providing analyses for ill-formed input (as intended), but also providing additional incorrect analyses for well-formed input—"spurious ambiguity". To see this, consider agreement, probably a good candidate for a less important "detail" of syntax which might safely be ignored. For example, it might be argued that sentence (1) below ought to be regarded as syntactically acceptable, since it's clear enough what's intended (and since the error is even common enough as a result of incomplete editing):

(1) ** Liste alle Sekretärinnen, die einen PC benutzt*
List all secretaries who uses a PC. [*sic*]

Syntactically tolerant systems would accept this sentence, but they would then have no way of distinguishing correct and incorrect parses of sentences such as (2), which are distinguished only by agreement:

(2) *Liste jede Sekretärin in Finanzabteilungen, die einen PC benutzt.*
List every secretary in finance departments who uses a PC.

The relative clause *die einen PC benutzt* can of course only be understood as modifying *jede Sekretärin* (the only NP with which it agrees), but a system which ignored agreement information would have no way of eliminating the parse in which the relative clause is construed as modifying *Finanzabteilungen.* Sag [20] demonstrates that this problem of SPURIOUS AMBIGUITY is potentially ubiquitous in NLU systems which try to understanding natural language without linguistic expertise.

Furthermore, even if we accepted the argument that some applications may ignore syntactic accuracy, we are still faced with the applications at the other end of the spectrum of syntactic sensitivity, i.e., applications where syntactic accuracy is essential. Applications of this sort are found where the microstructure of the text plays an important role, e.g., grammar or style checking, and generally the entire area of NL generation: clearly, nobody wants a system which over-generates in synthesis. Similarily it is hard to find any advantage for underconstrained systems in applications such as speech understanding, where the whole point of using syntactic information is to reduce the number of hypotheses—a goal served only by maximally constrained systems.

We therefore believe that syntactic precision is indispensable for some applications and valuable even in applications in which ill-formed input may be expected.

The diagnostic tool assesses correctness of syntactic analysis—it supports the recognition of bugs in the linguistic analysis. This in turn provides both a means of assessing the effects of proposed changes in syntactic analysis as well as a means of tracking progress in system coverage over time. Neither of these derivative tasks is realistically feasible without the aid of an automated tool. Humans may spot individual errors when attending propitiously, but we are poor at systematic checks and comparisons, especially in large systems created by groups over relatively long periods of time.

2.2 Linguistic Research

This is an appropriate point at which to acknowledge our own debt to descriptive and theoretical linguistics, from which our primary data—the German sentences themselves—have been gathered. We expect to reciprocate, i.e., we expect that descriptive linguistics and even linguistic theory may benefit from the data collection effort we have undertaken. These benefits may take different forms: first, we have begun gathering the data in a single place; second, we are organizing it into a database in a fairly general way, i.e., with relatively little theoretical prejudice, so that variable perspectives on the data are enabled; third, in addition to relatively crude data analysis routinely provided in linguistic data collections—which seldom extends beyond marking ill-formedness/well-formedness, we have provided further fundamental data annotations. Fourth, and most intriguingly, the time may not be distant when linguistic hypotheses may be tested directly on the computer. Many contemporary computational systems for natural language syntax are based on ideas of current interest in theoretical linguistics as well, and there is interest in general machinery for implementing syntactic analysis for wide varieties of linguistic theories. At the point where syntacticians can test their hypotheses easily by running grammars, the use of diagnostic tools will be of immediate interest in linguistic research as well.

In sketching these potential benefits of the general data collection and analysis effort we have begun, it should be clear that we do not intend to speak only to linguists emploring "corpus-based" methodologies: our information includes facts about the ill-formedness of strings as well as rudimentary data analysis. This will become clearer below.

2.3 Toward Evaluation

The catalogue of syntactic material we have collated is intended for deployment in diagnosis—the recognition and characterization of problems in syntactic analysis. This is a task different from general system evaluation, which in most cases will judge the performance of a system relative to the achievement of a goal which is set by an application. Even if we limit evaluation to the performance of the syntactic component of a system, there are still some differences which have to be kept in mind.

There are two very central issues which arise in applying test suites to the problem of evaluation: the choice of phenomena, and the relative importance of the phenomena chosen. The choice of phenomena is critical because we have attempted to choose syntactic annotations whose importance is broadly recognized. There is really no alternative if the diagnostic tool is to be generally useful. But in choosing such areas we automatically exclude areas which are theoretically controversial—with the inevitable consequence that these areas become immune to effective testing. We chose to annotate the position of NPs and PPs as relatively uncontroversial examples, but we could not have chosen adverbial attachment (VP vs. S) since there would have been too little agreement about the putative constituents. But this means that our suite does not test for adverbial attachment. Now, obviously controversy does not imply insignificance (if anything the opposite), which leaves us in the uncomfortable position of being unable to test important areas.

Of course this is not peculiar to the use of test suites as opposed to naturally occurring data, as Black et al. [3] demonstrate. We have only two remarks to make. First, this is a genuine consequence of trying to provide a tool whose generality justifies its development. The generality allows the tool to be employed in combination with various syntactic theories and allows it to be used to compare theories. More sophisticated data annotations would simply not be useful here (although it would do no harm to include them among others). Second, it is worthwhile recalling the experience of speech community, which has long since accepted the use of test suites which systematically restrict evaluation to relatively objective criteria (text correctness, ignoring stress, duration and intensity). In spite of the focus on a subset of the issues, the overall use of the test suites is felt to have contributed to progress in speech recognition.

Turning to the problem of the relative importance of various phenomena, the contrast between diagnosis and evaluation can be appreciated if one considers the case of applying our diagnostic tool to two different systems. In virtually every case, the result we obtain will show that neither system is perfect (nor perfectly incorrect), and that neither one analyzes exactly a subset of the constructions of the other. Suppose, for the sake of illustration, that one system is superior in treating long-distance (multi-clausal) dependencies, while the other is better at simple clause structure, but that the performance of the two systems is otherwise the same. The diagnosis is complete, but the evaluation still needs to determine the relative importance of the areas in which coverage diverged.[1] If matters were always as simple as in this illustration, we might appeal to a consensus of informed opinion, which would in this case certainly regard the treatment of simple clause structure as more important than that of long-distance dependencies—and would therefore evaluate the systems accordingly. But matters need not and normally are not so simple at all. There simply is not a consensus of informed opinion about the relative importance of various areas of grammatical coverage.

An example of crucial information that is lacking from our catalogue of syntactic material is information about relative frequency of occurrence. If this information could be obtained and added to the database, then it should be possible to develop an evaluation system of sorts from our diagnosis system.[2] It would be much more difficult to obtain information about the relative significance of constructions apart from frequency since this probably depends on the application and may be only tangentially related to syntax.

3 The Diagnostic Facility

We include here a brief description of the diagnostic facility; more detailed documentation, especially for the various areas of coverage of the syntactic catalogue, is likewise available. (Cf. Diagne [6], Klein and Dickmann [13])

3.1 Sentence Suite

As noted in the introduction, our material consists of sentences we have carefully constructed to illustrate syntactic phenomena; we have not attempted to collect examples from naturally occurring text. Several considerations weighed in favor of using the the artificially constructed data:

[1] Strictly speaking, this is not necessary; we could evaluate all such cases as equally significant, but (i) the results of such "evaluation" would be too coarse to be of much use; and (ii) this simply goes against good sense. Some areas of grammatical coverage simply are more important than others. See the example in text, where simple clause structure is certainly more important than long-distance (multi-clausal) dependency—at least for most applications in German.

[2] But it is not clear that this is the best way to go about developing an evaluation system. For example, we are not making any effort to keep some of the material secret, as speech evaluation systems routinely do in order to prevent a bias toward test material.

- since the aims are error detection, support of system development, and evaluation of systematic coverage, we need optimal control over the test data. Clearly, it is easier to construct data than to collect it naturally when we have to examine (i) a systematic range of phenomena or (ii) very specific combinations of phenomena.

- we wished to include negative (ill-formedness) data in order to test more precisely (cf. discussion in Section 2.1 on "spurious ambiguity" and also on the needs of generation). Negative data is not available naturally. (This is not to say that natural corpora contain no errors, only that they are not marked as such.)

- we wished to keep the diagnostic facility small in vocabulary. This is desirable if we are to diagnose errors in a range of systems. The vocabulary used in the diagnostic tool must either (i) be found in the system already, or (ii) be added to it easily. But then the vocabulary must be limited.

- we wished to exploit existing collections of data in descriptive and theoretical linguistics. These are virtually all constructed examples, not naturally occurring text.

- data construction in linguistics is analogous to the control in experimental fields—it allows the testing of maximally precise hypotheses.

We have no objection to including naturally occurring data in the catalogue, subject to the restrictions above (especially constraining the size of the facility).

In this connection we should again note that the considerable benefits which diagnostic and evaluation efforts have brought to speech research have occurred almost entirely using data which are to some extent constructed—e.g., which are read in laboratory conditions, or which have been collected for use with a specific application. Cf. Pallett et al. [17] for a contemporary use of a constructed corpus in speech research.

The vocabulary for the test suite has been taken from the domain of personnel management wherever possible. We chose this domain because it is popular in natural language processing, both as a textbook example and as an industrial test case. The domain of personnel management would also be useful in case we are to diagnose errors in semantics as well as syntax (which we are not attempting to do at present, but which is an interesting prospect for the future). It presents a reasonably constrained and accessible semantic domain. Where no suitable vocabulary from the domain of personnel management presented itself, we have extended the vocabulary in *ad hoc* ways.

The suite of test sentences is being collated by various contributors, each specializing in a single area of coverage, e.g., verb government, coordination, or NP constructions. Because of the range of syntactic material which is eventually to be included, it is difficult to draw precise guidelines about the sentences. Still, several factors have been borne in mind while constructing the syntactic examples.

- lexicon size (cf. above)

- adherence to the following standards: (somewhat) formal, conversational High German; i.e., we have avoided colloquialisms, literary peculiarties, and regional dialects.
- selected testing of negative examples. We have tried to keep the catalogue small, but not at the cost of using great ingenuity to create minimal sets of testing data, nor at the cost of introducing very unnatural examples into the test catalogue. We have not rigorously purged superfluous examples.
- minimization of irrelevant ambiguity (bearing in mind that it cannot be fully eliminated).
- attention to analytical problems. We have attempted to catalogue not only the constructions, but also the problems known to be difficult in their analysis.

We do not deceive ourselves about our chances for success with respect to the last point: our catalogue is doubtlessly incomplete in many respects, but most sorely in this one. We invite comment and contribution everywhere, but most especially in further cataloguing the known analytical problems in German syntax.

In stressing our intention to catalogue analytical problems as well as the basic range of syntactic construction types, we do not wish to suggest that we are gathering a collection of "cute examples". We have gathered some cute examples, but these are relatively few in the general catalogue. Our primary goal remains a coverage of phenomena which is as comprehensive as feasible, even if this involves the rather tedious compilation of theoretically relatively well-explored and scientifically "uninteresting" constructions, such as the full paradigms illustrating determiner-adjective-noun agreement in German or the different types of verbal subcategorization. From our experience, it is above all the absence of systematic and comprehensive test-beds which hampers system development, rather than the lack of ingenious examples (which frustrate all systems in some way or other). Our goal is thus not primarily to show what systems cannot do, but to support the extension of what they can do.

3.2 Syntactic Annotations

In choosing which annotations about the sentences might be sensible, we have been guided by two considerations. First, the catalogue will be much more useful if examples from *selected* areas can be provided on demand. For example, it is useful to be able to ask for examples of coordination involving deletion of the subject in the first conjunct—as opposed to simply coordination (an area of coverage). This means that we need to provide annotations about which area of coverage a given sentence (or ill-formed string) is intended to illustrate. With regard to these annotations, we have merely attempted to use standard (traditional) linguistic terminology.

Second, we can exploit some annotations to check further on precision of analysis. This is the purpose of annotations such as:

- well-formed vs. ill-formed
- position of finite matrix verb
- position of NPs
- position of PPs

So, in a sentence such as (3), the following database values are encoded:

(3) *Der Manager bittet um den Vertrag.*
the student asks for the contract

Category	Position
finite matrix verb	3
position of NPs	1-2, 5-6
position of PPs	4-6

In selecting these properties as worthy of annotation, we were motivated primarily by a wish to focus on properties about which there would be little theoretical dispute, which would be relatively easy to test, and which would still provide a reasonable reflection of a system's accuracy. Let us note finally that in annotating sentences with their syntactic properties we do not suppose that there is only one view about syntactic structure in general, but only that some subset of the properties annotated will be recognized by all or most researchers. (Cf. Black et al. [3] for application in English syntax.)

3.3 Example 1: Verbal Government

One of the phenomena which the data collection already covers is the area of verbal government, i.e., verbal subcategorization frames. The aim was to compile a comprehensive list of *combinations of obligatory complements* of verbs, forming the basis of different sentence patterns in German. We ignore both adjuncts and optional complements in restricting ourselves to obligatory complements, which can be tested by an operationalizable criterion, a specific sort of right extraposition:

(4) *Er hat gegessen, und zwar Bohnen.*
he has eaten, namely beans.

(5) **Er hat verzehrt, und zwar Bohnen.*
he has consumed, namely beans

(6) **Er hat das Buch gelegt, und zwar auf den Tisch.*
he has put the book, namely on the table

(7) *Er hat Maria geküßt, und zwar auf die Wange.*
he has kissed Mary, namely on the cheek

We attempted to find instances of all possible combinations of nominal, prepositional, sentential, but also adjectival complements.[3] Clearly, we could not immediately cover the entire field in full depth, so that we decided to adopt a breadth-first strategy; e.g., we ignored the more finegrained distinctions to be made in the area of infinitival complementation or expletive complements. The description in these areas will be elaborated at later stages.

The result of the collection is a list of about 70 combinations which are exemplified in about 220 sample sentences (440 sentences including the negative examples).

The sentences illustrate:

- combinations of nominal, prepositional and adjectival complements, viz.,
  - nominal complements only:

    (8) *Der Manager gibt dem Studenten den Computer.*
    the manager gives the student the computer

  - nominal and prepositional complements with semantically empty (9) or non-empty prepositions (10):

    (9) *Der Vorschlag bringt den Studenten auf den Lösungsweg.*
    the suggestion takes the student to the solution

    (10) *Der Manager vermutet den Studenten in dem Saal.*
    the manager assumes the student in the hall

  - nominal and adjectival (or predicative) complements

    (11) *Der Manager wird krank.*
    the manager becomes ill

- nominal complements combined with finite (subordinate) clauses, introduced by the complementizers *daß* (12), *ob* (13) or some wh-element (14):

  (12) *Daß der Student kommt, stimmt.*
  that the student comes, is-correct

  (13) *Dem Manager entfällt, ob der Student kommt.*
  it escapes the manager, whether the student comes

  (14) *Der Manager fragt, wer kommt.*
  the manager asks who comes

- nominal complements in combination with infinitival complements, illustrating bare infinitives (15) and *zu*-infinitives (16):

  (15) *Der Manager hört den Studenten kommen.*
  the manager hears the student come

[3] At the basis of our list were collections to be found in the literature, such as Colliander [4], Engel [8], Engelen [9], Goetze [10], Helbig [12] and Weisgerber [24]. We are also grateful to Stefanie Schachtl, Siemens Munich, who provided us with some of her material.

(16) *Der Manager behauptet, den Studenten zu kennen.*
the manager claims to know the student

- examples involving some of the combination above in connection with expletive or correlative prepositional pronouns or expletive *es*:

(17) *Der Vorschlag dient dazu, den Plan zu erklären.*
the proposal serves (to-it) to explain the plan

(18) *Der Manager achtet darauf, ob der Student kommt.*
the manager checks (on-it) whether the student comes

(19) *Der Manager hält es für notwendig, den Vertrag zu kündigen.*
the manager considers it (for) necessary to cancel the contract

Since we are interested only in verbal government here, we tried to keep as many other parameters as possible carefully under control: as already mentioned, the noun phrases in the sample sentences are built from a limited vocabulary. Almost all noun phrase and prepositional complements have a definite determiner. In the case of prepositional phrases the fusion of preposition and determiner (*in dem* → *im*) is avoided. Since German has relatively free word order, the different complements have to be identified by their case marking in most cases—as a consequence, morphological ambiguities of case (e.g., between feminine or neuter nominative and accusative) were excluded. The matrix and subordinate clauses all have only one verbal head (i.e., they do not have any temporal auxiliary), and the verb's morphological form is the third person, singular, present, indicative active form. The sentences do not contain any additional irrelevant modifiers, adjuncts or particles. The word order of the sample sentences is meant to illustrate the "un-marked" order, although this should not play an important role, since the complements are uniquely case marked, as mentioned.

Every combination of complements is illustrated by at least one example. In addition, each government type is paired with a set of ill-formed sentences, which illustrate three types of errors relevant for verbal government:

- an obligatory complement is missing;
- there is one complement too many;
- one of the complements has the wrong form.

In describing the complement structure of the sentences we chose a vocabulary which is of course not theory neutral, but which at least can be expected to meet common agreement. We tried to avoid theory-specific notions such as *subject* or *direct object*, and identified the complements on the basis of morphological case marking, prepositions, complementizers and/or the morphology of the verb. Obviously, this vocabulary cannot exhaustively characterize the properties of individual complements. For example, with those few verbs which subcategorize for two accusative NPs it is quite unlikely that both NPs behave in the same way with respect to passivization. Similarily, a nominative complement ("subject") may have different propertives depending on the verb being un-accusative or un-ergative. However, we think that distinctions of this kind should be dealt with seperately in data sets on e.g., passivization, ergativity, etc.

3.4 Example 2: Coordination

A second set of test sentences covers data illustrating coordination phenomena such as gapping, right-node-raising and 'left-' and 'right-deletion'. The data are limited to sentential coordination involving NP complements only. Phenomena which so far have not been considered include coordination of adjuncts and nouns, of prepositions and prepositional phrases and 'asymmetrical coordination', i.e. non-clausal coordination of unlike categories.

The basis of the sentence suite were two complex sentences consisting of two simple clauses connected by *und*. These clauses each contain a main verb and three NP complements (with unambiguous nominative, dative and accusative case). The finite main verb occurs in final (20) or second (21) position:

(20) *Ich glaube, daß der Professor der Sekretärin den Blumenstrauß schenkte, und der Student dem Kommilitonen den Roman verkaufte.*
(I believe that the professor gave a bunch of flowers to the secretary and [that] the student sold the novel to his colleague.)

(21) *Der Professor schenkte der Sekretärin den Blumenstrauß und der Student verkaufte dem Kommilitonen den Roman.*

These two sentences were submitted to systematic variations. On the one hand all possible deletions of constituents in both sentences are exemplified. We ommitted all patently impossible combinations, e.g., sentences where the same grammatical function (e.g., the nominative subject noun phrase) was deleted in both conjuncts. Second, the order of NPs in the second conjunct was systematically permuted. We constrained the permutations to one of the conjuncts in order to avoid a combinatory explosion.

The combination of the variations together with the restrictions led to a total number of 530 grammatical and ungrammatical sample sentences. Due to the systematic deletion and order permutations we not only automatically cover all possible cases of specific forms of coordination, such as gapping, right node raising etc., but also some interesting variations of coordination where two internally dissimilar clauses or partial clauses are coordinated.

The ungrammatical sentences cover interesting generalizations (such as the prohibition of finite verb gapping in the first conjunct in verb-second clauses in contrast to verb-final clauses), and they also allow one to test a more extreme 'misbehaviour' of systems. Sentence (22) for example contains four constituents, namely NPs in nominative, dative and accusative form plus a finite main verb, which together would make up a complete clause but are clearly insufficient for a coordinate clause.

(22) **Ich glaube, daß einen Blumenstrauß schenkte, und der Student dem Kommilitonen.*

A simplistic system which indiscriminately reconstructed in each conjunct constituents which are to be found in other conjuncts might easily accept this sentence as well-formed.

3.5 Database

The material is organized in a relational database, such that queries can ask either for a description or classification of a sentence or for sentences matching combinations of descriptive parameters.

3.5.1 Abstract Data Model

The purpose of the database is to maintain and allow access to data concerning syntactic material, which may cover several areas. Our goal in developing the database was to (i) provide a concise organization of syntactic data, (ii) ease access to syntactic information, (iii) maintain consistency, (iv) allow variable logical views of data, (v) allow an efficient extension of the syntactic material to treat further areas.

In this section we will present the conceptual schema of the database, without giving details on the implementation. We rather present the database from a conceptual point of view. Figure 1 represents the Entity Relationship (ER) schema diagram of the database. Its current content (syntactic material) treats—enumerating in the order of their development—verbal government, coordination and fixed verbal structures (Funktionsverbgefüge or FVG). i.e., semi-idiomatic constructions with a semantically almost void verbal head and some more or less fixed nominal, prepositional or adjectival complement.

According to the **Entity-Relationship** (ER) terminology, we can identify entity types (CATEGORY, SENTENCE, VERBAL-GOVERNMENT, COORDINATION, FVS), relationship types (S-CATEGORY, S-VERBAL-GOVERNMENT, S-COORDINATION, S-FVS) and attributes for each relation. From a mathematical point of view both entity and relationship types are relations.

The major relation in the database is the entity type SENTENCE. A tuple of SENTENCE contains (i) an identifier **s-id** for tuple which is unique within the relation (primary key), (ii) a sentence that exemplifies the given properties **s-example** (according to the underlying area of application), (iii) the sentence length **s-length** (can be computed by some triggers), (iv) a specification of the wellformedness of the sentence example **wf**, (v) an error-code for ill-formed sentences **error-code**, and (vi) additional comment **s-comment**.

The CATEGORY relation may contain any category specification, e.g., ‘np’. Its participation in the S-CATEGORY relation is partial. A tuple of S-CATEGORY specifies the position and lexical form of a given category in a given sentence. A category may occur in several sentences and a sentence may contain several categories (m:n relationship).

A new application area is recorded in the database by defining one or more new entity types whose attributes and values depend on the underlying syntactic phenomenon.

The specification of a new entity type is primarily based on linguistic properties. In accordance with this specification new sentences (tuples) that exemplify the underlying syntactic phenomenon will then be inserted into the SENTENCE relation. Finally, a new relationship type with the participants SENTENCE and the new entity type must be defined to relate each sentence to the syntactical phenomenon it describes.

The following example shows database entries for sentence (23).

(23) *Der Manager hindert den Studenten daran, den Plan zu erklären.*

SENTENCE

s-id	**s-example**	**s-length**	**wf**	**error-code**	**s-comment**
1210	(23)	10	1	1	-

CATEGORY

cat-desc	**comment**
cor	correlate
inf-comp	zu-infinitive
f-verb	finite matrix verb
np	nominal phrase

S-CATEGORY

s-id	**cat-desc**	**pos-from**	**pos-to**	**substring**
1210	cor	6	6	daran
1210	f-verb	5	5	hindert
1210	np	1	2	Der Manager
1210	np	4	5	den Studenten
1210	np	7	8	den Plan
1210	inf-comp	7	10	den Plan zu erklären.

Splitting the position attribute into **pos-from** and **pos-to** makes the trigger-based generation of the corresponding substrings possible and facilitates a concistency check (e.g., **pos-from** must be a positive integer number equal or less than **pos-to**, which is also a natural number; **pos-to** must be equal or greater than **pos-from** and equal or less than the sentence length **s-length**.). The query language allows the use of the attribute **cat-position** to get the position of a category in a sentence, i.e., a combination of **pos-from** and **pos-to**.

The following example illustrates the database entry for sentence (24), representing an instance from the domain of coordination:

(24) ** Der Professor schenkte den Blumenstrauß und dem Kommilitonen der Student den Roman.*

COORDINATION

coord-id	**coord-desc**	**s-cat**	**n1**	**v1**	**d1**	**a1**
3081	N1_V1_A1_D2_N2_A2	V-second	0	0	2	0

n2	**v2**	**d2**	**a2**	**n_d**	**n_a**	**d_a**	**coord-comment**
1	2	1	0	0	1	1	null

where **coord-id** encodes the identifier for the example; **coord-desc** gives a description in the form of a linear encoding of the overtly realized constituents; **s-cat** encodes the sentence type (V-first or V-second); **n1, v1, ... n2** encode presence (0), absence (2) or permutation (1) of the nominative NP, the verb, etc. in the first or second conjunct; **n_d 0/1** etc. encodes whether the nominative NP precedes the dative NP in the second conjunct etc.; **coord-comment** finally gives some informal information, mainly on what kind of phenomenon is exemplified by the sentence.

3.5.2 Database System

The database is administered in the programming language **awk**. Some of the reasons which speak in favor of **awk** are:

- **awk** is in the public domain running under UNIX and should run in other environments; in particular, it runs on MS-DOS.
- Its ability to handle strings of characters as conveniently as most languages handle numbers makes it for our purposes more suitable than standard relational database systems; i.e., more powerful data validation, increasing availability of information with a minimal number of relations and attributes.

Compared to standard databases **awk** has a restricted area of application and does not provide fast access methods to information, but it is a good language for developing a simple relational database where character strings are involved. Additional resources and tools like a report generator and query languages were easily implemented. The database includes a reduced **sql**-like query language. We use the database entries of the examples given above to ask the following queries:

1. retrieve all sentences including a correlate, an accusative np and a *zu*–infinitive.

```
query:  retrieve s-id, s-example
        where match(comp-desc, cor) and match(comp-desc, acc)
              and match(comp-desc, zu-inf)

result: 1210  Der Manager hindert den Studenten daran, den Plan zu erklaeren.
        1211  Der Vorschlag bringt den Studenten darauf, den Vortrag zu ueberarbeiten.
```

2. retrieve the position and the lexical form of all NPs of sentence 1210.

```
query:  retrieve cat-desc, position, substring
        where s-id = 1210 and cat-desc = “np”

result: np  1  2  der Manager
        np  4  5  den Studenten
        np  7  8  den Plan
```

3. retrieve all sentences featuring correct gapping (i.e., deletion of the main verb and one or more NPs) in the second conjunct

   ```
   query: retrieve s-id s-example coord-desc
          where wf=1 and v2=2 and (n2=2 or d2=2 or a2=2)

   result: 3025   der professor schenkte der sekretaerin den blumenstrauss
                  und dem studenten den roman.
                  N1_V1_D1_A1_D2_A2
   (This result is obtained together with 17 other sentences)
   ```

The query language has been developed under SunOS using the utilities **lex** and **yacc**. **lex** is a lexical analyzer generator designed for processing of character input streams. **yacc**, a LALR(1) parser generator, is an ancronym for Yet Another Compiler Compiler. It provides a general tool for describing an input language to a computer programm.

3.5.3 Auxiliary Materials

The database of syntactic material is accompanied by a few auxiliary development tools. First, in order to support further development of the catalogue and database, it is possible to obtain a list of words used (so that we minimize vocabulary size); we will also provide a list of differentiating concepts (so that categorization names may be accessed easily). Second, documentation is available on each of the areas of syntactic coverage included. This is to cover (minimally) the delimitation of the area of coverage, the scheme of categorization, and the sources used to compile the catalogue.

Third, a small amount of auxiliary code may be supplied to support development of interfaces to parsers. This need not do more than dispatch sentences to the parser, and check for the correctness of results. DiTo currently supports this only by writing sentences to a specified file.

4 Comparison to Other Work

Batori and Volk [2] suggest that test suites be seen as a type of corpus particularly well suited for the SYSTEMATIC testing of computational linguistics work. We find this view congenial, and in particular see no contradiction in using both test suites and corpora of naturally occurring data.

Ours appears to be the first attempt to construct a general diagnostic facility for German syntax, even if virtually every natural language processing group working on German has a small suite of sentences used for internal monitoring and debugging. (Cf. Volk and Ridder [22] for an interesting example.)

There have been several related efforts concerned with English syntax. Guida and Mauri [11] report on attempts to evaluate *system* performance for natural language processing systems (*n.b.*, not merely syntax) in which they attempt to finesse the issue of correctness (which we argue to be central) by measuring user satisfaction. We have attempted to address the issue of syntactic correctness head-on.

Hewlett-Packard Laboratories compiled a test suite of approximately $1,500$ sentences which it distributed at the *Public Forum on Evaluating Natural Language Systems* at the 1987 Meeting of the Assocation for Computational Linguistics, cf. Flickinger [7]. That effort differed from the present one in that it tried to evaluate semantics and pragmatics, as well as syntax, and in that it consisted essentially of sentences without annotated properties. The sentences were not organized into a database.

Read et al. [19] advocate a "sourcebook" approach, in which fewer examples are submitted to much closer scrutiny. The closer scrutiny does not seem subject to automation, at least at present. Furthermore, their emphasis is on evaluating *systems* for natural language understanding, and the primary focus seems to be on domain modeling, conceptual analysis and inferential capabilities, not syntax. It is similar to the HP approach (and to ours) in employing primarily constructed examples, rather than naturally occurring ones.

The Natural Language group at Bolt, Beranek, and Newman Systems and Technologies Corporation circulated a corpus of approximately $3,200$ sample database queries formulated in English at the 1989 DARPA Workshop on Evaluating Natural Language, cf. Palmer and Finin [18]. The emphasis here, too, was on system (natural language understanding) performance, rather than specializations, but most of their examples seem to come from actual trial use of a natural language interface program, which gives their work added value.

The University of Pennsylvania's "Treebank" project (similar to a project of the same name at the University of Lancaster sponsored by IBM) has begun an effort to annotate naturally occurring text and speech, and to organize the annotations into a "Treebank". The annotations are phonetic, syntactic, semantic and pragmatic, and the intended scope is monumental. Since they wish to gather representative and varied data, they hope to collect and annotate approximately 10^8 words. It was the consensus of the DARPA panel (discussed in Palmer and Finin [18]) that the Treebank proposal be pursued as the most promising for syntactic evaluation. Our proposal differs only in using constructed data.

The *Data Collection Initiative* of the *Association for Computational Linguistics* (ACL / DCI) is a loosely organized confederation of efforts concerned with the classification and annotation of various sorts of texts (cf. Liberman [15]). Our work will be made available to this group.

Finally, Black et al. [3] propose an evaluation scheme which mixes some elements of corpus-based and suite-based approaches. They take example sentences from the Brown Corpus (of English prose) which are hand-annotated with simple constituent structure markings much like those we use. They then input the sentences to various parsers and compare the labeled bracketings obtained with those from the hand annotations, showing how one can develop a measure of deviation for less than perfect results. They do not address the problems of relative importance of constructions or that of competing readings.

5 Current State, Future Plans

5.1 Emerging Issues in Suite-Based Diagnostics

Perhaps the most difficult issue in suite-based diagnostics and evaluation is that of accommodating unintended readings. Try as one will, it seems impossible to construct large numbers of sentences which do not include at least some arguably ambiguous examples. But then sentences appear in the test suite with a single marking of constituent structure when more than one is possible, and others appear with the '*' of unacceptability, when they would be acceptable under an unlikely interpretation. An example may clarify the difficulty. One grammar we developed assigned (as a first analysis) to the sentence below the bracketing it appears with (the bracketing for the intended reading is as one would expect in English, which is shown):

(25) Schmidt ((ist (in dem)) Haus)
Schmidt (is (in (the house)))

But the unintended bracketing comes about when the parser allows that the determiner *dem* appear anaphorically (*In dem [Auto] fahren wir!* 'In that [car] we drive'), and further allows that the copula appear with a bare $\bar{\mathbf{N}}$, which is common enough in German (*Sie ist Ingineurin* 'She is [an] engineer'). The copula plus $\bar{\mathbf{N}}$ construction is limited to a class of nouns (potentially) denoting classes of humans (professions, nationalities, etc.), but this is arguably semantic information. In other words, the parser understood the sentences as syntactically analogous to the following:

(26) Arbeitet Schmidt bei dem Projekt?
Schmidt ((ist (bei dem)) Manager)
Schmidt is with that manager

We do not think that mere cleverness in constructing examples can totally eliminate this kind of ambiguity, even if we can try to minimize it. But the ambiguity infects the use of suites (and annotated corpora) in at least two ways: first, the possibility of ambiguity implies that it can only be reasonable to ask of a well-formed example whether a parse with the given syntactic annotations is provided, but not whether NO other is; and second and more seriously, it calls into question the marking of sentences as unacceptable (with '*'),[4] since it means that the only import of '*' we can vouch for is that the string is unacceptable without special context. We conclude from this circumstance that test suites should not be employed—either for diagnostics or evaluation—without submitting the results to an analysis of errors.

A second difficulty we have alluded to above (Sec. 2.3), and that is the step from diagnosis to evaluation, since this inevitably involves some relative weighting of the data. We have no brief in this matter, but the example of the speech community is again instructive, where they have allowed frequency to determine relative weighting.

[4] And indeed, the experience of at least some who have used the HP test suite has been that sentences are too liberally marked as '*'. This was noted by Daniel D.K. Sleator and Davy Temperley, who employed the suite at Carnegie Mellon. Cf. [21].

The evaluation data for speech are collected in a way that guarantees representativeness, and no further effort is made to determine the relative importance of phenomena.

5.2 Eventual Range of Syntax Catalogue

As mentioned, we regard our work only as a starting point which has to be complemented by contributions from other groups and individual experts. As to extensions of the database, one of the strongest criterion on including a suite of data covering a phenomenon will be whether the data can be described in a uniform and uncontroversial manner. This will a priori limit the possible range mainly to phenomena for which we have morpho-syntactic clues in the widest sense. This does not mean of course that the level of morpho-syntax will have to be considered the appropriate level of *explanation* for a phenomenon.

Consider the example of control in non-finite structures. Whether this phenomenon is to be treated on the level of semantics, on the level of syntax or a mixture of the two, is a question which is very much open to discussion. However, whatever the theory and the explanation of that phenomenon is, it will have to account for the grammaticality and ungrammaticality of the morpho-syntactic agreement relations between some constituent of the matrix clause and the reflexive pronoun in the infinitive:

(27) *Ich* behaupte, *mich*/**dich*/**sich* zu freuen
(I claim to enjoy myself/*yourself/*himself)

(28) Er forderte *mich* auf, *mich*/**dich*/**sich* zu freuen
(He asked me to enjoy myself/*yourself/*himself)

Thus, the following list, which we intend to be suggestive rather than definitive, can only give a rough idea of what aspects of German syntax could be considered important to be covered by a system. The list should be read under the above-mentioned restriction that not all aspects of these phenomena, but only the surface-oriented aspects should and could be treated in a syntactically oriented test suite:

Syntax of the simple clause, including verbal government and *genera verbi* (passive, etc.), negation, word order, and adverbial modification. Verb phrase complementation including argument sharing or inheritance (*auf Hans ist er stolz*), clause union, extraposition, modal and auxiliary verbs. Verbal complex, fixed verbal structures (*Funktionsverbgefüge*), separable prefix verbs, idioms and special constructions.

Noun phrase syntax, including determiner and numeral (and measure) system, relative clauses of various sorts (including preposed participial phrases), pre- and postnominal adjectival modification, noun phrase coordination, and plurals. Pronominal system and anaphora.

Prepositons and postpositions, cliticization, particles (e.g., *als, ja, je, denn*).

Questions, including long-distance (multi-clause) dependence. Imperative and subjunctive moods. Adjectival and nominal government, modification, and specification. Equative, comparative, and superlative constructions. Coordination and ellipsis.

5.3 Collaborations

We have contacted research groups in NLP and machine translation in the interest of exchanging specialized (and annotated) data sets in exchange for the rest of the database. Several groups are now involved in active cooperation: Institut für angewandte Informationswissenschaft (IAI), Saarbrücken, Institut für Computerlinguistik at the University of Koblenz (ICL-Koblenz), and the Gesellschaft für Mathematik und Datenverarbeitung (GMD), Darmstadt.

Brigitte Krenn [14] at the IAI has compiled data on the structure of semi-idiomatic verbal constructions (*Funktionsverbgefüge*); Martin Volk [23] at Koblenz has developed a test suite of relative clauses, and Renate Henschel and Elke Teich at the GMD have proposed a data collection and annotation effort on the syntax of modal and auxiliary verbs.

A diagnostic and evaluation tool of this sort ought to be commonly developed, used and maintained. Diagnosis improves in quality and general acceptance as further groups become involved, which in turn enables an increase in the quality and comparability of systems—the more so as the tool itself improves from common use. We invite further interested groups to contact us about collaborations.

References

[1] Alfred V. Aho, Brian W. Kernighan and Peter J. Weinberger: *The awk programming language.* Wokingham et al., Addison Wesley, 1988

[2] István Batori and Martin Volk: Das Verhältnis von natürlichsprachlichen Korpora zu systematischen Sammlungen konstruierter Texte. Workshop presentation, *Repräsentatives Korpus der deutschen Gegenwartssprache*, 15-16.Oct.1992. to appear in a report of the *Institut für Kommunikationsforschung und Phonetik*, Bonn.

[3] E. Black et al.: A Procedure for Quantitatively Comparing the Syntactic Coverage of English Grammars. *Proceedings of the February 1991 Speech and Natural Language Workshop.* Morgan Kaufmann: San Mateo, California, 1991

[4] Peter Colliander: *Das Korrelat und die obligatorische Extraposition.* Kopenhagener Beiträge zur Germanistischen Linguistik. Sonderband 2. Kopenhagen, 1983.

[5] P. Chen: The entity-relationship model. Toward a unified view of data. *ACM Transactions on Database Systems.* No. 1, 1976.

[6] Abdel Kader Diagne: DiTo – *A Diagnostic Tool for German Syntax. Data Base and User's Manual.* DFKI Technical Document D-92-05, DFKI, Saarbrücken 1992.

[7] Daniel Flickinger, John Nerbonne, Ivan Sag, and Thomas Wasow: Towards evaluation of natural language processing systems. Technical report, Hewlett-Packard Laboratories, 1987.

[8] Ulrich Engel: Die deutschen Satzbaupläne. In *Wirkendes Wort 20*, pages 361–392, 1970.

[9] Bernhard Engelen: *Untersuchungen zu Satzbauplan und Wortfeld in der geschriebenen deutschen Sprache der Gegenwart.* Reihe I 3.3 Verblisten. München, 1975.

[10] Lutz Götze: *Valenzstrukturen deutscher Verben und Adjektive.* München, 1979.

[11] Giovanni Guida and Giancarlo Mauri: Evaluation of natural language processing systems: Issues and approaches. *Proceedings of the IEEE*, 74(7):1026–1035, 1986.

[12] Gerhard Helbig: *Wörterbuch zur Valenz und Distribution deutscher Verben.* Leipzig, 5th ed., 1980.

[13] Judith Klein and Ludwig Dickmann: *Daten–Dokumentation: Verbrektion und Koordination.* DFKI Technical Document D-92-04, DFKI, Saarbrücken 1992.

[14] Brigitte Krenn: Funktionsverbgefüge: Eine Datenbeschreibung unpub. Documentation, Institut für angewandte Informationsforschung, Saarbrücken.

[15] Mark Liberman: Text on Tap: the ACL/DCI. In *Proceedings of the October 1989 Speech and Natural Language Workshop.* Morgan Kaufmann: San Mateo, California, 1989, pp.173-188.

[16] Judith Klein, Ludwig Dickmann, Abdel Kader Diagne, John Nerbonne, and Klaus Netter: DiTo: Ein Diagnostikwerkzeug für die syntaktische Analyse. In *Tagungsband KONVENZ 92.* Springer: Berlin, 1992, pp.380-385.

[17] Dave S. Pallett et al.: DARPA Research Management Benchmark Test Results, June 1990. In *Proceedings of the June 1990 Speech and Natural Language Workshop.* Morgan Kaufmann: San Mateo, California, 1990, pp.298-305.

[18] Martha Palmer and Tim Finin: Workshop on the Evaluation of Natural Language Processing Systems. In: *Computational Linguistics 16(3)*, 1990, pp.175-181.

[19] Walter Read, Alex Quilici, John Reeves, Michael Dyer, and Eva Baker: Evaluating natural language systems: A sourcebook approach. In *COLING '88*, pages 530–534, 1988.

[20] Ivan Sag: Linguistic Theory and Natural Language Processing. In Ewan Klein and Frank Veltman, eds. *Natural Language and Speech. Symposium Proceedings.* Springer-Verlag: Berlin, 1991.

[21] Daniel D. K. Sleator and Davy Temperley: *Parsing English with a Link Grammar.* Carnegie Mellon School of Computer Science Technical Report CMU-CS-91-196, October 1991.

[22] Martin Volk and Hanno Ridder: GTU – eine Grammatik Testumgebung mit Testsatzarchiv to appear in: *LDV-Forum 1.1992*

[23] Martin Volk: Kurzbeschreibung der Testsatzsammlung zu den Relativsätzen unpub. Documentation, Universität Koblenz.

[24] Monika Weisgerber: *Valenz und Kongruenzbeziehungen.* Frankfurt a. M., 1983.

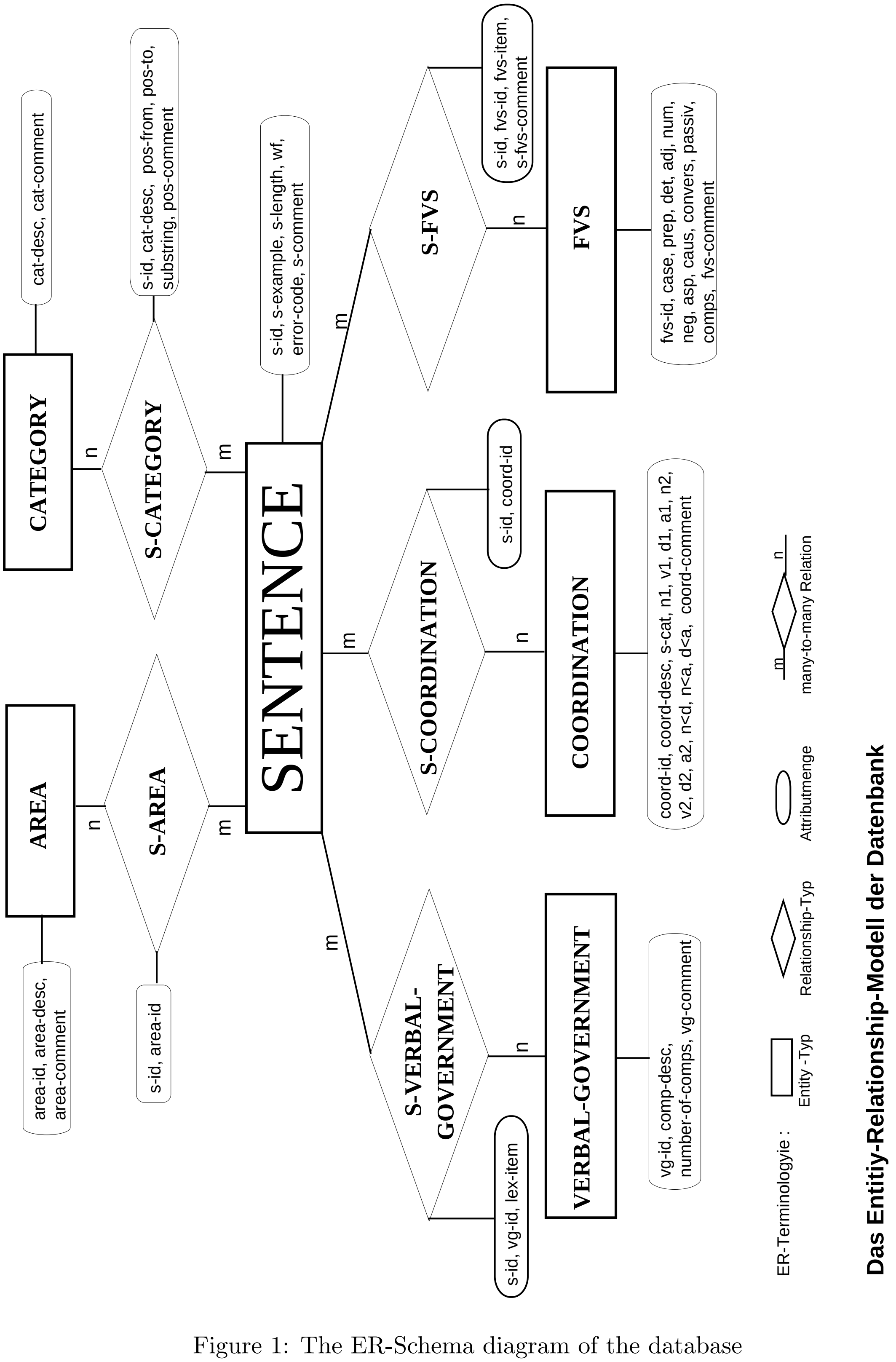
SENTENCE
s-id, s-example, s-length, wf, error-code, s-comment
AREA
area-id, area-desc, area-comment
S-AREA
s-id, area-id
CATEGORY
cat-desc, cat-comment
S-CATEGORY
s-id, cat-desc, pos-from, pos-to, substring, pos-comment
S-VERBAL-GOVERNMENT
s-id, vg-id, lex-item
VERBAL-GOVERNMENT
vg-id, comp-desc, number-of-comps, vg-comment
S-COORDINATION
s-id, coord-id
COORDINATION
coord-id, coord-desc, s-cat, n1, v1, d1, a1, n2, v2, d2, a2, n<d, n<a, d<a, coord-comment
S-FVS
s-id, fvs-id, fvs-item, s-fvs-comment
FVS
fvs-id, case, prep, det, adj, num, neg, asp, caus, convers, passiv, comps, fvs-comment
m
n
ER-Terminologyie :
Entity -Typ
Relationship-Typ
Attributmenge
many-to-many Relation
Das Entitiy-Relationship-Modell der Datenbank

Figure 1: The ER-Schema diagram of the database