Constraint–based Diagnosis for Intelligent Language Tutoring Systems

Wolfgang Menzel & Ingo Schröder
Projekt DAWAI · Fachbereich Informatik · Universität Hamburg
Vogt–Kölln–Straße 30 · 22527 Hamburg · Germany
menzel|ingo.schroeder@informatik.uni-hamburg.de

Abstract

If the student of a foreign language is expected to benefit from the interactive nature of computer–based tutoring systems, solutions are required which combine the ability to accept free form input with the production of helpful feedback on the quality of the utterances received. A solution is presented which provides language learning systems with the desired diagnostic capabilities for a wide range of syntactic, semantic, and domain–related phenomena. It is based on a procedure for structural disambiguation in a multi–level representation using graded constraints.

Zusammenfassung

Die sinnvolle Nutzung der interaktiven Möglichkeiten von Sprachlehrsystemen erfordert Analyseverfahren mit der Fähigkeit, auch für frei formulierte Schülerlösungen eine präzise Bewertung der sprachlichen Qualität zu ermitteln und diese in aussagekräftige Lernhinweise für den Schüler umzusetzen. Es wird ein Ansatz vorgestellt, der die gewünschten Diagnosefähigkeiten für einen großen Bereich an syntaktischen, semantischen und domämenspezifischen Phänomen zur Verfügung stellt. Ausgangspunkt ist ein Verfahren zur strukturellen Disambiguierung einer Mehrebenenrepräsentation auf der Basis von bewerteten Constraints.

1 Introduction

Two requirements have to be fulfilled by any stimulating language learning environment:

1. It should encourage the creative use of language in communicatively relevant settings.
2. It should provide the student with adequate feedback regarding the quality of her utterance, covering both grammaticality and communicative appropriateness.

So far no technical solution exists which satisfies both requirements at the same time. If, on the one hand, the system design has put emphasis on high quality feedback the student will almost certainly have to restrict her choice to a few predefined items for which specifically tailored responses can be provided. Unfortunately, such restricted exercises bear only little resemblance to a situation of natural person–to–person communication and any creative use of language is severely hampered. On the other hand, a broad coverage in free form input is always achieved at the expense of sacrificing the diagnostic capabilities of the system.

The difficulty of combining these two desiderata in a single solution obviously results from a basic characteristic of natural language which comes along with a vast amount of local ambiguity if special aspects (e. g. its syntax, semantics, pragmatics) are treated in isolation. Under such conditions the parsing of perfectly composed utterances becomes a serious problem let alone the possibility to accept different kinds of deviations as they are regularly produced by beginning students. Here, the system has to solve two tasks at the same time:

1. Robust parsing: Try to obtain a structural interpretation of the student's utterance even if it possibly contains unexpected or unacceptable constructions.
2. Fault diagnosis: Try to identify the particular kind of problem in terms of explanation possibilities and strategies for remedy.

Although being fundamentally different, these two tasks are highly interrelated and depend strongly on each other: Whereas diagnosis is possible only with respect to a presumed underlying structure of the erroneous utterance (e. g. 'If this constituent is meant to be the subject of the sentence, it is of wrong case'), parsing can be performed efficiently only if strong hypotheses about the particular kind of errors are available (e. g. 'If the case of this noun was nominative instead of dative, it could be the subject of the sentence'). Thus, preliminary parsing hypotheses are always needed prior to the diagnosis, and at the same time diagnostic results are a prerequisite of the parsing procedure. Moreover, while robust parsing requires an at least partial ignoring of certain regularities of the language system (otherwise no interpretation can be found for a deviant utterance), diagnosis needs to check whether the same well–formedness conditions hold (otherwise not a single error can be detected).

In this respect the diagnosis of natural language utterances differs remarkably from other diagnostic tasks, where the structure of the system under diagnosis is expected to be known in advance (Davis 1994, Struss 1992). The natural language diagnosis requires a structural identification to become part of the diagnosis proper. Because of the mutual dependency, both tasks will have to be carried out in a highly integrated computational framework which allows to check structural hypotheses and well–formedness conditions simultaneously.

Robust behavior for natural language parsing systems is usually attempted by means of over–generating rule systems which contain *error rules* for extra–grammatical phenomena. For language learning purposes this would imply to anticipate and explicitly specify every erroneous construction which could possibly be produced by a student (Yazdani 1986). An alternative approach uses *constraint retraction* techniques where certain well–formedness conditions are temporarily ignored if otherwise no consistent structural description can be generated (Uszkoreit 1991, Erbach 1993). Thus, weaker instances of grammar rules are derived from the normal ones whenever this seems necessary.

Applications to the area of foreign language learning usually require a combination of both techniques. Schwind (1995), for instance, uses a constraint retraction approach for the class of agreement errors and error rules for structural faults (e. g. missing constituents, inappropriate linear orderings etc.). While error rules lend themselves easily to the creation of small scale demonstration systems, it seems, however, infeasible to exhaustively anticipate every potential error configuration and to describe it by means of corresponding error rules. Constraint retraction techniques, on the other hand, require a rather strong structural backbone to rely upon. Therefore, their application is usually restricted to selected types of regularities and severely limited exercises (Menzel 1988, Menzel 1990).

Both error rules and constraint retractions provide a good starting point for the derivation of error diagnoses. As long as singleton errors are considered, simple error explanation schemes can be defined and used to produce the desired feedback for the student.

Unfortunately, both techniques lead to tremendous efficiency problems since they neutralize valuable information which even in the error free case is urgently needed to constrain the search space. This problem becomes a particularly serious one, because neither approach uses graded ratings for (partial) structural hypotheses and, therefore, does without an important means to guide the search for an appropriate solution.

Particularly, the application of empirically obtained gradings in probabilistic grammars has turned out to be a major factor for introducing a considerably higher degree of robustness in the parsing of natural language (cf. Briscoe 1994). However, probabilistic grammars have to be trained on huge corpora of natural language examples, taken e. g. from running newspaper texts. If a grammar for diagnostic purposes is required it will need to be trained on similar collections of typical learner utterances. This approach does not seem particularly promising since it can hardly be imagined how a corpus could be collected, which is statistically representative not only with respect to relevant language structures but moreover to possible error situations *and* exercise types. Notice that one and the same utterance can be acceptable in one context but quite inappropriate in another. Therefore, the probabilistic approach would require corpora which are properly annotated according to the different error categories and exercise contexts, because only then it might become possible to induce the relevant information on the distinction between the acceptable and the unacceptable case from the given data.

Like most contemporary approaches to robust parsing probabilistic grammars suffer from a biased focus on the isolated treatment of syntactic phenomena. This syntax–oriented approach not only causes severe difficulties with respect to local ambiguity and efficiency, in addition, it puts tremendous limitations on the quality of diagnostic results since it reduces diagnosis to a context insensitive similarity comparison. For example, a purely syntax–based diagnosis will certainly find two equally likely explanations for the number disagreement in the example (1) where the noun can be corrected to singular, or alternatively the verb can be changed to plural.

(1) * The cars drives fast.

Given a suitable context (e. g. where only one car is under consideration) this diagnostic uncertainty immediately disappears. In certain cases contextual information might even put a much stronger perspective on an utterance which eventually can override syntactic evidence. In such cases a convincing diagnosis can only be obtained if the diagnostic component takes into consideration as much contextual information as possible. Such a representation of context conditions should include knowledge about the domain of discourse, about the discourse situation (who is speaking to whom, where, and when) as well as about previous discourse contributions.

This contextual embedding then provides an anchor point for the diagnosis, and error explanations can be found which are well motivated in the given situation. A quite similar strategy can also be observed with human teachers, who never consider an erroneous utterance in isolation, but try to collect evidence from very different sources to infer the most likely intention behind the student's contribution. These assumptions about the underlying intention are not only used to find a plausible diagnosis but, furthermore, serve as a reference for possible repair proposals: 'If you want to express this and that, better try it the following way....'

On the other hand, it should be noticed that contextual information never provides an absolute point of reference. In any case it is based on nothing but *assumptions* on likely student behavior (e. g. she will answer a given question properly) and nobody is able to prevent a student from producing strange responses. Under these circumstances any assumption about plausible behavior is doomed to fail and might become subject of diagnostic efforts itself.

To avoid a system break down under such circumstances every piece of model information has to be defeasible, and partial representations for the different levels of language have to be loosely coupled. The robust behavior and diagnostic abilities of the system are based on the assumption that combined deviations on different levels will be encountered only in rather exceptional cases. Under usual conditions a bidirectional information flow among representational levels will facilitate mutual support which allows to overcome, for instance, syntactic difficulties by means of semantic or domain–specific knowledge and vice versa.

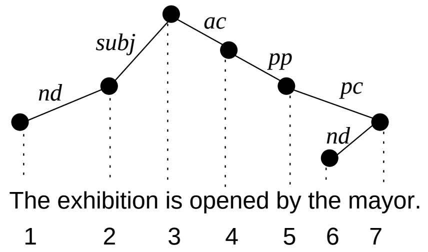
Whereas with the advent of multimedia–based tutoring systems a rich body of possibilities for a close–to–reality *presentation* of exercise contexts are available, there is an obvious lack of appropriate means for the *representation* of these knowledge components in a way which facilitates their purposeful exploitation in procedures for robust parsing and error diagnosis. This paper presents a proposal for an integrated approach to robust parsing and error diagnosis combining

- a multi–level representation which allows to bring together syntactic, semantic, pragmatic, and domain–specific knowledge in a uniform way,
- the use of graded, hence defeasible, constraints on all levels, and
- a common arbitration mechanism which allows to weigh evidence from very different sources against one another.

The approach is based on a procedure for structural disambiguation which eliminates elementary structural descriptions from an initially complete, but highly underspecified representation of all structural interpretations for a given utterance. Thus, it especially facilitates the comparison of alternative interpretations and the arbitration of possibly contradicting evidence. Section 2 gives a short introduction to the underlying eliminative parsing mechanism which is based on constraint satisfaction techniques. This basic procedure will be extended to accommodate graded constraints and is applied to a multi–level representation. Section 3 and 4 contain the corresponding details and analyze the consequences for the robustness of the resulting parsing procedure. Afterwards, Section 5 gives a number of examples for the diagnostic capabilities achieved so far. Finally, we summarize the approach in Section 6.

2 Eliminative Parsing

Parsing by means of constraint satisfaction has first been described by Maruyama (1990*b*). It was developed for the application in an interactive machine translation system (Maruyama 1990*a*). Later the idea has been extended to the processing of word hypothesis lattices instead of linear strings (Harper, Jamieson, Zoltowski & Helzerman 1992, Harper, Jamieson, Mitchell, Ying, Potisuk, Srinivasan, Chen, Zoltowski, McPheters, Pellom & Helzerman 1994, Harper & Helzerman 1994) and has been implemented on a massively parallel hardware architecture (Helzerman & Harper 1992).



(a)

$$\begin{aligned} v_1 &= \langle nd, 2\rangle \\ v_2 &= \langle subj, 3\rangle \\ v_3 &= \langle nil, 0\rangle \\ v_4 &= \langle ac, 3\rangle \end{aligned} \qquad \begin{aligned} v_5 &= \langle pp, 4\rangle \\ v_6 &= \langle nd, 7\rangle \\ v_7 &= \langle pc, 5\rangle \end{aligned}$$

(b)

Figure 1: *(a) Syntactic dependency tree for an example utterance: For each word form a unique subordination and a label, which characterizes the kind of subordination, are to be found. (b) Labellings for a set of constraint variables: Each variable corresponds to a word form and takes a pairing consisting of a label and an index (corresponding to the superordinated word form) as a value. A value of* $\langle nil, 0\rangle$ *indicates the root of the tree.*

Parsing by constraint satisfaction aims at producing a dependency tree (cf. Figure 1a), where each word form of an utterance is unambiguously subordinated to another with a unique label describing the kind of dependency relation between the two candidates. Admissible dependency relations are specified using constraints (cf. Figure 2).

{X} : ArticleAgree : Article : X↓cat=ART → X.label=nd ∧ X↑cat=NOUN ∧
X↓pos<X↑pos ∧ X↓case=X↑case ∧ X↓num=X↑num ∧ X↓gender=X↑gender

'Articles modify nouns to the right and agree with them with respect to case, number, and gender.'

Figure 2: *A simple constraint: It consists of a variable declaration, a name, a class and a formula of propositional logic which encodes grammatical knowledge.*

Basically, a constraint consists of a logical formula which is parameterized by variables (in our example X) which can be bound to an edge in the dependency tree. It is associated with a name (e. g. ArticleAgree) and a class (e. g. Article) for identification and modularization purposes respectively. Selector functions are provided which facilitate access to the label of an edge (e. g. X.label) and to lexical properties of the dominating node (e. g. X↑case) and the dominated one (e. g. X↓case). Being universally quantified, a typical constraint takes the form of an implication with the premise describing the conditions for its application. Accordingly, the constraint of Figure 2 reads as follows: Each article (X↓cat=ART) modifies a noun (X↑cat=NOUN) to the right (X↓pos<X↑pos) as a noun modifier (X.label=nd) and agrees with its dominating form in regard to case, number, and gender (X↓case=X↑case . . .). In order to restrict the complexity of the constraint satisfaction problem (CSP) only unary and

binary constraints are used. Hence, no more than two variables are allowed to appear in a constraint, and it is not possible to express conditions for a structural configuration of more than two dependency edges. This, certainly, is a rather strong restriction. It puts severe limitations on the possibility to model grammatical and extra–grammatical knowledge, which will be discussed in Section 5.2.

Given the above specification of a parsing problem, the word forms of an utterance can be interpreted as the variables of a CSP, which are to receive a unique value assignment as a pair consisting of a label and a dominating word form. Figure 1b shows such an assignment which exactly corresponds to the structure in Figure 1a.

Initially the constraint net contains all possible structural interpretations in a highly dense representation. This initial state corresponds to a structural description of maximum ambiguity. Each variable's domain is bound to the complete set of subordination possibilities. The constraint satisfaction procedure successively discards local assignments if they are not licensed by the set of constraints. Eventually, an almost disambiguated structure with mostly unique value assignments might become available, from which a single structural description can easily be extracted.

Unfortunately, in the presence of an inconsistent CSP the procedure described so far is not able to find a solution. It, therefore, still lacks the desired robust behavior which would enable it to determine a structural description for erroneous utterances, too. Although the basic algorithm can easily be modified to let the last value assignment survive under any circumstances, this only introduces a rudimentary notion of robustness which is highly sensitive to arbitrary variations, e. g. in the sequence of constraint applications.

3 Robust Parsing with Graded Constraints

For deviant input sentences it is usually not possible to find a structural interpretation which satisfies all constraints simultaneously. In terms of CSP the problem is *over–constrained*. For such a problem one can try to find a solution which at least partly fulfills the requirements. *Partial constraint satisfaction problems* (PCSP; Tsang 1993, Freuder & Wallace 1992, Wallace & Freuder 1995) can be divided into *minimal violation problems* (MVP), where you want to find a labelling such that the minimum constraints are violated, and the *maximal utility problem* (MUP), the solution of which assigns values to a maximum subset of the variables with no constraints violated. Therefore, the partial constraint satisfaction problem can be seen as a generalization of the traditional CSP. In the context of constraint parsing the minimal violation interpretation seems more appropriate since a solution of the parsing process (as opposed to scheduling tasks for instance) should be a structure that covers the whole sentence, not just parts of it. The MVP approach introduces robustness into constraint parsing because now a solution for arbitrary input, possibly with some constraints violated, can be found.

This kind of robustness, however, is not quite satisfactory because all the constraints are treated as being of equal importance which, in general, is not the case. Therefore, every constraint c receives a weight $w(c)$ chosen from the interval $[0, 1]$ to denote how serious one considers a violation of the constraint (cf. Figure 3). Furthermore, complex constraints like the one in Figure 2 are broken down into smaller ones in order to facilitate an as fine grained distinction as possible among different kinds of constraint violations.

{X} : SubjInit : Subj : 0.0 :
  X.label=subj → X↓cat=NOUN ∧ X↑cat=FINVERB
'A subject is a noun and it modifies a finite verb.'

{X} : SubjNumber : Subj : 0.1 :
  X.label=subj → X↓num=X↑num
'The subject agrees with the verb with respect to number.'

{X} : SubjOrder : Subj : 0.9 :
  X.label=subj → X↓pos<X↑pos
'The subject is placed in front of the verb.'

{X, Y} : SubjUnique : Subj : 0.0 :
  X.label=subj ∧ X↑id=Y↑id → Y.label≠subj
'The subject is unique for a given verb.'

Figure 3: *Very restrictive constraint grammar fragment for subject treatment in German: Graded constraints are additionally annotated with a score.*

Now different types of conditions can easily be expressed with constraints:

- Hard constraints with score $w(c) = 0.0$ (e. g. constraint SubjUnique) exclude totally unacceptable structures from consideration. This kind of constraint also initializes the space of admissible solutions (e. g. constraint SubjInit; Menzel 1994).
- Typical well–formedness conditions like agreement or word order are specified by means of weaker constraints with score $0.0 < w(c) \ll 1.0$, e. g. constraint SubjNumber.
- Weak constraints with score $0.0 \ll w(c) < 1.0$ can be used for conditions that are merely preferences rather than error conditions, e. g. constraint SubjOrder prefers subject topicalization to object topicalization in German, but does not enforce it (and does not even put a strong penalty on it). Uncertain information, e. g. derived from prosodic clues or fuzzy domain–specific knowledge, can also be incorporated by weak constraints. Uncertain and preference–based information makes sure that, similar to a human listener, *only a single* structure that fits the given conditions best will be produced. As long as there is any kind of preference, be it grammatical or not, no enumeration of possible solutions will be generated.[1]
- Constraints with score $w(c) = 1.0$ are totally ineffective due to the multiplicative combination.

The solution of such a partial constraint satisfaction problem with scores[2] is the dependency structure of the utterance that violates the fewest and weakest constraints. In order to formalize this intuitive notion, the notation of constraint weights is extended to scores for dependency structures. The scores of all constraints c violated by the structure under consideration s are multiplied and a minimum selection is carried out to find the solution s' of the PCSP:

[1]If, for some reason, more than one possible interpretation of an utterance is desired the constraint parsing approach can easily be modified to return all the structures whose ratings do not differ too much from the best rating.

[2]Sometimes this kind of CSP is also called *stochastic CSP* or *constraint satisfaction optimization problem*.

$$s' = \arg\min_s \prod_c w(c,s)$$

Since a satisfied constraint should not decrease the score of a structure it holds that:

$$w(c,s) = \begin{cases} w(c) & : \text{ if structure } s \text{ violates constraint } c \\ 1.0 & : \text{ else} \end{cases}$$

The use of scores contributes directly to an improved robustness because it is now possible to rank constraint violations according to their impact on the acceptability of a solution.

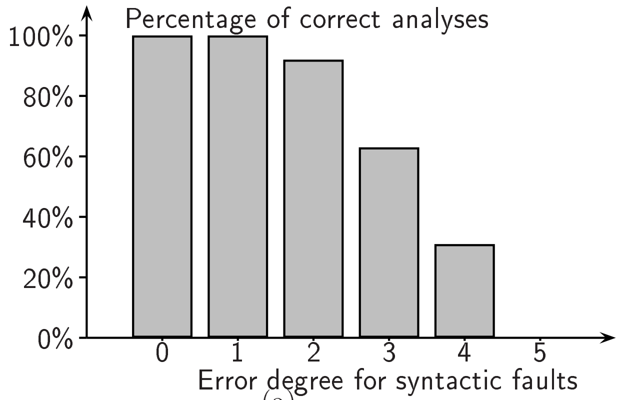
For evaluation purposes a prototypical diagnosis component for German as a foreign language has been developed. Although the prototype is limited yet, it has shown to be sophisticated enough to be immediately applied within a teaching unit. So far the grammar contains nearly 160 constraints and covers the following syntactic phenomena: active (future, present, perfect, past, and past perfect) and passive (present and past) voice of the verb, verbal and nominal genitive attributes, nominal groups including articles, adjectives, and nouns (declination classes, definiteness, and adverbial adjective modifiers), prepositional phrases, and simple subordinated clauses. Modal verbs, negations, relative clauses, and coordinations have not been dealt with yet. In order to study the robustness properties of this grammar the German sentence 'Der Mann besichtigt den Marktplatz. (The man visits the marketplace.)' has been systematically distorted by introducing different kinds of syntactic errors, and a global error measure has been defined to describe the degree of disorder for the resulting variations of the example utterance:

Case agreement The case of the subject as well as the object has been varied to take nominative, genitive, and accusative case respectively. While a shift from nominative or accusative to genitive counts as a single error, mixing up nominative and accusative counts as a double fault, because it is more difficult then to get the structural interpretation right.

Number agreement Analogous to the case parameter the subject and the object have been assigned singular and plural word forms. Note that, although there is no number agreement between finite verb and object in German, the chance of interchanging subject and object increases if the desired object agrees with the verb. Therefore, the analysis becomes increasingly more difficult when one abolishes the agreement of the subject and establishes it for the object.

Word order While German has a relatively free word order, there is nevertheless a slight preference of placing the subject in front of the object. It should be noted that the marked word order does not count as an error, but a preferred word order nevertheless helps to find the correct analysis.

The resulting 72 variations, some examples of which illustrate the kind of errors and the error measure e in Figure 4b, have then been analyzed using the above mentioned grammar. Figure 4a shows the accumulated results (for this simple example). Utterances that contain



(a)

(2) 'Den Marktplatz besichtigen die Männer.'
→ $e = 0$, no error, but marked word order

(3) 'Die Männer besichtigen der Marktplatz.'
→ $e = 2$, nominative case instead of accusative

(4) 'Den Marktplatz besichtigen den Mann.'
→ $e = 4$, combined errors

(5) 'Die Marktplätze besichtigen den Mann.'
→ $e = 7$, combined errors

(b)

Figure 4: *Percentage of correct analyses depending on the number of syntactic errors for syntactic grammar using graded constraints: As long as only up to two errors are made the correct analysis can usually be found; naturally, more and combined errors make the analysis fail.*

only a few rather simple errors are analyzed correctly; only in cases of combined constraint violations the analysis starts to break down.

The use of graded constraints for the parsing process guarantees that the least dispreferred structure is selected as the solution. The number of violated constraints is minimized as opposed to the traditional approach where the application of error rules is minimized.

4 Multi–Level Parsing

Human language understanding processes are extremely robust, because they exploit all kinds of information necessary to disambiguate an utterance and identify its meaning. Not only grammatical knowledge (or knowledge about language in general), but also contextual, domain–specific, and even common sense knowledge contribute to the overall task.

To mimic a similar behavior a multi–level parsing is adopted. Different description levels for a natural language utterance are established in parallel, and partial descriptions are mapped onto each other by means of graded constraints, thus providing a loose coupling among description levels (Menzel 1995). Evidence for a structure on one level leads to preferred structures on other levels without creating a fatal dependency: Mutual reinforcement takes place as long as supporting cross–level evidence is available, while its absence leads to autonomous decisions, and even contradicting information will not result in a failure of the overall analysis. The approach shows some resemblance to results from psycholinguistic research, which, on the one hand, support the autonomy of different description levels during human language understanding (Forster 1979) and, on the other hand, corroborate the mutual influence of these levels (Marslen-Wilson & Tyler 1987).

Parsing by constraint satisfaction as described so far must be modified in order to allow the use of multiple description levels. Instead of one constraint variable for each word form of the utterance one has to provide a constraint variable for each pairing of a word form and a description level. Constraints can be divided into intralevel and interlevel constraints now, depending on whether they pose restrictions on subordination edges on one level or on different levels. The solution does not form one single dependency tree, but a whole set of trees. Figure 5 shows two such dependency trees for the levels of syntax and semantics respectively.[3] Constraints for the semantic level result from lexical properties of the word forms and contextual information of the exercise.

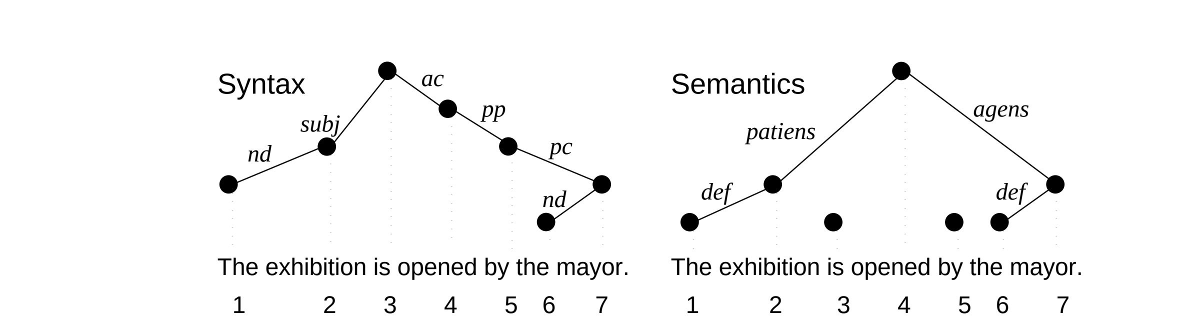


Figure 5: *Collection of dependency trees: Each tree represents a description level.*

The simplest way to present the contextual embedding to the student is to provide her with a textual description of a simple situation like the following and to ask her to answer corresponding questions. Alternative presentation modes might include spoken or visual information.

> Der Mann besichtigt den Marktplatz. Dort trifft er Anne. Aber sie ignoriert ihn. Verärgert geht der Mann in das alte Rathaus. Dort wird eine Ausstellung vom Bürgermeister eröffnet.

> The man visits the marketplace. He meets Anne there. But she ignores him. Being annoyed, the man enters the old town hall. There, an exhibition is opened by the mayor.

As long as this description is simple enough a suitable representation of its propositional content can be derived automatically. It later is fed into the parsing system again to constrain the semantic level when analyzing a student's response. Thus, expectations are geared towards sensible and relevant contributions without confining the student to particular syntactic constructions.

Using again the German sentence 'Der Mann besichtigt den Marktplatz.' it can be illustrated how the representation of different description levels helps to increase the robustness of the system. The 72 variations (cf. Section 3) are analyzed in nine different contexts where the sortal restrictions of the verb as well as the domain knowledge either support, do not influence, or contradict the desired solution.

Therefore, in addition to the syntactic parameters from Section 3 two more dimensions are introduced:

Sortal restrictions This criteria determines whether the semantic classes of the desired arguments match the sortal restrictions of the verb. A neutral value means that no sortal restrictions are checked, e. g. due to missing information.

[3] The visualization as trees is especially helpful for the grammar designer. The semantic level makes clear that it is not always beneficial to structure the solutions strictly as trees (cf. Section 5.2).

Domain knowledge This parameter determines whether the desired utterance is considered true or false in the domain. Given that the context of the exercise supports the desired interpretation, the unwanted readings get penalized, while a contradiction with the context leads to a degradation of the desired interpretation. If the domain knowledge is neutral regarding the interpretation under consideration no structural configuration gets negative support.[4]

Table 1 shows for all 648 more or less deviant variants of the example utterance whether the parsing process manages to find the desired solution.[5]

The rows and columns are ordered by the number of errors they contain, so that you find the seriousness of deviations increasing when proceeding from left to right and from top to bottom. In other words, the top left hand corner of Table 1 contains the results for utterances with no or few errors, the bottom right hand corner those with combined errors. A dark background coloring ($\boxed{++}$) indicates those sentences which could be analyzed as desired with unmarked as well as marked word order, while a light coloring ($\boxed{+-}$) denotes success for the marked case and failure for the unmarked one. White as the background color ($\boxed{--}$) finally signals that the original structure of the utterance could not be found in either case.

This kind of coloring gives a visual impression of the system behavior with respect to robustness. Using the available information on a complementary level as an anchor point even utterances with a remarkable number of errors can be analyzed correctly. The analysis fails to find the desired interpretation only in cases of highly complex distortions. While nearly half of the results for sentences with an error measure of three were wrong when only the syntactic level was represented, almost all utterances with an error measure up to five are interpreted correctly when enough semantic and domain–specific support is available. Of course, contradicting semantic and/or domain–specific expectations lead to a decrease in syntactic robustness. This was to be expected because of the symmetry of representation levels and constraints. The use of different knowledge levels leads to synergy effects, since none of the representation levels alone could achieve a similar degree of robustness.

Figure 6, which is an extension of Figure 4, shows the percentage of correctly analyzed utterances depending solely on the error measure for a supporting, neutral, and violating context respectively. If one does not consider the source of the errors, the above tendency becomes even clearer: The use of semantic and domain–specific knowledge greatly enhances the syntactical robustness in the supporting case. Naturally, the robustness is reduced if the additional information contradicts the intended interpretation.

It should be noted that, although we have stressed the robustness against syntactic deviations to enable the comparison of the multi–level representation with the syntax–only case, the robust behavior is symmetrical with respect to the different levels. Thus, positive information on the syntactic level also helps to find a semantic interpretation (which resembles more traditional serial architectures where semantic processing is based on the output of the syntactic component).

[4]The simplest way to incorporate domain knowledge into the constraint system is to encode the propositional content of the context directly as constraints.

[5]It is not possible to provide unique test sentences for every parameter combination since in the German language word forms often coincide, e. g. the nominative and accusative case of nouns.

		Error	0	1	1	2	2	2	3	3	4
		Domain	true	neut.	true	false	neut.	true	false	neut.	false
		Sorts	corr.	corr.	neut.	corr.	neut.	viol.	neut.	viol.	viol.
0	Nom, Acc	c, v	++	++	++	++	++	++	++	++	+ −
1	Nom, Acc	c, c	++	++	++	++	++	++	++	++	+ −
1	Nom, Gen	c, v	++	++	++	++	++	++	++	++	+ −
1	Gen, Acc	c, v	++	++	++	++	++	+ −	+ −	− −	− −
2	Nom, Acc	v, v	++	++	++	++	++	++	++	++	+ −
2	Nom, Gen	c, c	++	++	++	++	++	++	++	− −	− −
2	Nom, Nom	c, v	++	++	++	++	++	++	++	++	+ −
2	Gen, Acc	c, c	++	++	++	++	++	+ −	+ −	− −	− −
2	Gen, Gen	c, v	++	++	++	++	++	+ −	+ −	− −	− −
2	Acc, Acc	c, v	++	++	++	++	+ −	− −	− −	− −	− −
3	Nom, Acc	v, c	++	++	++	++	− −	− −	− −	− −	− −
3	Nom, Gen	v, v	++	++	++	++	++	++	++	− −	− −
3	Nom, Nom	c, c	++	++	++	++	+ −	− −	− −	− −	− −
3	Gen, Acc	v, v	++	++	++	++	++	+ −	+ −	− −	− −
3	Gen, Gen	c, c	++	++	++	++	+ −	− −	− −	− −	− −
3	Gen, Nom	c, v	++	++	++	++	++	+ −	+ −	− −	− −
3	Acc, Acc	c, c	++	++	++	++	+ −	− −	− −	− −	− −
3	Acc, Gen	c, v	++	++	++	++	+ −	− −	− −	− −	− −
4	Nom, Gen	v, c	++	++	++	++	+ −	− −	− −	− −	− −
4	Nom, Nom	v, v	++	++	++	++	+ −	− −	− −	− −	− −
4	Gen, Acc	v, c	++	+ −	− −	− −	− −	− −	− −	− −	− −
4	Gen, Gen	v, v	++	++	++	++	+ −	− −	− −	− −	− −
4	Gen, Nom	c, c	++	++	− −	− −	− −	− −	− −	− −	− −
4	Acc, Acc	v, v	++	++	++	++	+ −	− −	− −	− −	− −
4	Acc, Gen	c, c	++	++	+ −	+ −	− −	− −	− −	− −	− −
4	Acc, Nom	c, v	++	++	++	++	+ −	− −	− −	− −	− −
5	Nom, Nom	v, c	++	++	++	++	− −	− −	− −	− −	− −
5	Gen, Gen	v, c	++	+ −	− −	− −	− −	− −	− −	− −	− −
5	Gen, Nom	v, v	++	++	− −	− −	− −	− −	− −	− −	− −
5	Acc, Acc	v, c	+ −	− −	− −	− −	− −	− −	− −	− −	− −
5	Acc, Gen	v, v	++	++	+ −	+ −	− −	− −	− −	− −	− −
5	Acc, Nom	c, c	+ −	− −	− −	− −	− −	− −	− −	− −	− −
6	Gen, Nom	v, c	++	+ −	− −	− −	− −	− −	− −	− −	− −
6	Acc, Gen	v, c	+ −	− −	− −	− −	− −	− −	− −	− −	− −
6	Acc, Nom	v, v	+ −	− −	− −	− −	− −	− −	− −	− −	− −
7	Acc, Nom	v, c	+ −	− −	− −	− −	− −	− −	− −	− −	− −
Error	Case	Number									

Table 1: *Parsing results for a systematically distorted sentence: In the table from left to right and top to bottom the number and seriousness of errors increase. Case and number agreement (‘c’ means correct, ‘v’ violated) is given for the subject and object respectively.*

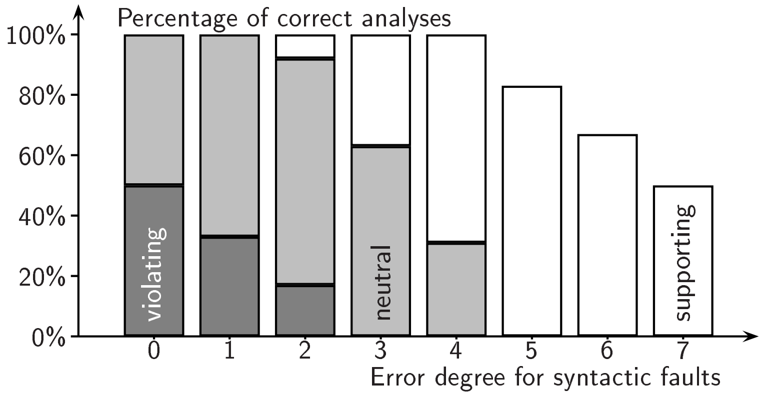


Figure 6: *Percentage of correct analyses depending on the number of syntactic errors and, additionally, parameterized by the kind of semantic and domain–specific support: The better the semantic and domain–specific support, the more errors can be compensated for. Neutral support means no support as in Figure 4.*

For all applications that use some kind of score or probability a major concern is the acquisition of those numbers. Although some research dealing with the automatic extraction of constraints from tree corpora has been carried out (Schröder 1996), the grammar (including the constraint weights) for the diagnosis has been developed manually by means of plausibility considerations only. It turned out that the main results — robustness against a remarkable number of errors and support from complementary levels — remain surprisingly stable when subjected to small modifications of the constraint weights.

5 Diagnosis

Based on the constraint parsing procedures introduced and extended in the Sections 2 to 4 a simple language learning system has been implemented. It is rather limited in its breadth since only a few teaching units have been designed. In these units the students are asked to answer some questions about a given situation or describe it in their own words. Thus, the students have to produce free form sentences and (optionally) understand given language input.

Since mistakes in students' language use lead to constraint violations it is quite a simple task to identify the errors in the parsing result. An appropriate interpretation component has to map the set of constraint violations to a set of possible explanations. It is not a trivial one–to–one mapping, since some weak constraints should probably not lead to an explanation, while others have to be grouped into clusters and reported to the student as one consistent compound diagnosis. Nevertheless it has been found that the design of the interpretation process is fairly straightforward.

5.1 Sample results

In the following we give some examples for the quality of errors that can be recognized by the diagnosis component.

Partial parsing If it is not possible to find a complete main clause that can be structured as a tree with the finite verb as the root, a partial analysis is performed.

'Das alte Rathaus.' ⟶ Missing subordination of the noun 'Rathaus'
'Der Mann... Klaus schläft.' ⟶ Missing subordination of the noun 'Mann'

Agreement Verbs and their arguments, articles and nouns, adjectives and nouns, prepositions and nouns etc. agree with respect to gender, case, number, person etc.

'Die schöne Mann schläft.' ⟶ Missing gender agreement for the article 'die'
'Der Mann besichtigt dem Marktplatz.' ⟶ Missing case agreement for the article 'dem'

Word order In the German language verbs are placed at the second position in main clauses, and the rest of the verb arguments has a canonical ordering, too.

'Die Stadt besichtigt der Mann.' ⟶ Object topicalization
'Der Mann die Stadt besichtigt.' ⟶ Wrong word order, verb not in second position

It should be noted that the first sentence is syntactically absolutely correct. But in the absence of other (non–syntactic) reasons for the object topicalization it reads a little bit strange.

Auxiliary selection Verbs determine which auxiliary ('haben' or 'sein') is used for their perfect form.

'Der Mann ist die Stadt besichtigt.' ⟶ Wrong auxiliary 'ist'
'Der Mann hat in die Stadt gegangen.' ⟶ Wrong auxiliary 'hat'

Case frames Verbs have case frames that must be filled by verb arguments.

'Schläft.' ⟶ Missing first argument
'Der Mann besichtigt.' ⟶ Missing second argument

Sortal restrictions Verbs pose certain restrictions on the semantic classes of their arguments like animacy.

'Die Stadt schläft.' ⟶ Violation of sortal restrictions for the first argument
'Der Mann sieht die Idee.' ⟶ Violation of sortal restrictions for the second argument

The violation of sortal restrictions often indicates a metaphorical use of the verb. Both example sentences do have a plausible interpretation under certain assumptions. Whether such a metaphorical use should be allowed can be controlled by the corresponding constraints.

Contextual restrictions A representation of the embedding context makes it possible to diagnose not only syntactic and semantic, i. e. language inherent, mistakes, but also errors regarding the propositional content or pragmatic aspects of the utterance. Thus, comprehension problems of the student while reading the introductory text can be identified.

'Anne besichtigt die Stadt.' ⟶ Propositional content not supported by the context
'Anne wird von dem Mann ignoriert.' ⟶ Propositional content not supported by the context

5.2 Results and Remaining Difficulties

The prototype is capable of performing a diagnosis of (relatively simple) natural language sentences. By using graded constraints on all levels of processing the analysis shows a universal robustness against a wide range of ungrammaticality and different violations of context induced expectations. Error diagnoses can be easily extracted from the parsing results for deviant input and immediately transformed into error explanations. Partial parsing is used as a fall–back in case a single structure for the utterance cannot be found.

There is, however, a number of difficulties and problems which need to be discussed in more detail. Although the restriction to at most binary constraints does not entail a limitation of the theoretical expressiveness of the formalism,[6] it definitely has some practical consequences. To be able to treat particular linguistic phenomena by means of binary constraints, sometimes the grammar writer has to adopt rather artificial constructs. Complex verbal groups, like modal verb constructions, for instance, normally need more than two subordination edges to be constrained simultaneously. In the worst case, transitivity chains of arbitrary length may exist. Only at the expense of introducing additional linguistically unusual as well as computationally expensive labels and/or levels these constructs can be described by binary constraints. This problem is even more urgent for interlevel–constraints that have to relate information on different levels to each other. It is, however, possible to approximate some ternary constraints by a set of binary ones. A possible solution to the problem above could be to postpone some of the more difficult constraint checks until the structure has settled. Then even complete transitivity checks can be performed efficiently (Menzel 1992).

While dependency trees are well suited for syntactic descriptions, they pose some problems on other levels. For instance, it is not always possible and often difficult to express domain information as subordination structures. Since dependency trees use only word forms as nodes, no distinction between word form and reference object can be made. Their identification, however, is viable only in applications which require a limited degree of variation in the context. This seems to be appropriate for a wide range of language learning situations where the designer of an exercise can control the context to a large extent. For more ambitious applications a more general solution is required.

Constraint parsing does not employ knowledge about the native language of the student, although mistakes resulting from a transfer of regularities from the mother tongue to the foreign language are quite common. These kinds of errors are easily identified by error rules, but require special treatment in the case of constraint parsing. Therefore, an integration of special error rules into the constraint parsing procedures may be desirable.

Finally, our prototype does not check the appropriateness of the student's utterance in regard to the task in question. As long as the answer does not violate any of the syntactic, semantic, and domain–specific expectations no corrections will be generated even if the utterance totally misses the topic. To overcome this deficiency the system has to be modified to use dynamic constraints, i. e. constraints specific to a particular task must be added when applicable (cf. Weischedel, Voge & James 1978). These constraints will be violated if the answer does not contain a minimum amount of relevant information.

[6] For instance, Nudel (1983) showed that every CSP with constraints of arbitrary arity can be transformed into a binary CSP at the expense of dramatically increasing the number of possible domain values. Since we model natural language, it may also be interesting that constraint grammars with binary constraints are strictly more expressive than context free grammars (Maruyama 1990*a*).

A general problem not restricted to our proposal, but for all free form diagnosis components concerns the certainty of diagnosis results. No diagnostic system for free form input (not even a human teacher) works absolutely correct, because the relevant information is too complex. So, while very simple exercise types like completion tests, where the number of valid answers is small, can guarantee the correctness of their diagnosis, this is not true for free form exercises. It, therefore, seems appropriate to inform the user about this uncertainty and to recommend additional advice from a human teacher.

6 Conclusions

We proposed multi–level parsing with graded constraints as a new technical solution for diagnosis of free form input in intelligent language tutoring systems. The system finds the most appropriate interpretation of a possibly faulty utterance and identifies the well–formedness conditions violated by the student. The key features of the approach are

- scoring of all partial and complete analyses and
- use of all kind of information, be it syntactic, semantic, domain–specific, contextual, or what else seems appropriate.

The system of constraints constitutes a model of appropriate language use. Both the structural interpretation and the diagnostic results for deviant input are derived from this model of correctness. This characteristic clearly justifies the classification of the approach as model–based, although constraint diagnosis superficially seems quite different from other model–based diagnosis approaches (Struss 1992). While other approaches can and do assume the structure to be static, both the behavior *and* the structural interpretation have to be described and restricted by constraints in constraint parsing systems.

A prototypical implementation has shown its applicability to language learning exercises of at least modest degree of sophistication. It holds high promise for the development of complete language tutoring solutions which successfully combine a close–to–reality interaction with the ability to provide the necessary feedback for improvement.

Acknowledgements

The authors thank G. Evermann, K. A. Foth, M. Fürter, M. Glockemann, A. Häming, S. Hamerich, T. Kroll, A. Popa, H. Rölke, M. Schulz, T. Schöllhammer, and N. Stockfleth, members of the summer term 1997 project group "Robust processing of natural language", for their contributions to the development of the first prototype system.

This research has been partly funded by the DFG (Deutsche Forschungsgemeinschaft) under grant no. Me 1472/1–1.

References

Altmann, G. & Steedman, M. (1988), 'Interaction with context during human sentence processing', *Cognition* **30**, 191–238.

Briscoe, T. (1994), Prospects for practical parsing of unrestricted text: Robust statistical parsing techniques, *in* N. Oostdijk & P. de Haan, eds, 'Corpus–based Research into Language', Rodopi, Amsterdam.

Cooper, W. E. & Walker, E. C. T., eds (1979), *Sentence Processing: Psycholinguistic Studies Presented To Merill Garret*, Lawrence Erlbaum, Hillsdale, NJ.

Davis, R. (1994), 'Diagnostic reasoning based on structure and behavior', *Artificial Intelligence* **24**(1), 347–410.

Erbach, G. (1993), Towards a theory of degrees of grammaticality, Bericht 34, Computerlinguistik, Universität Saarbrücken.

Forster, K. I. (1979), *Levels of Processing and the Structure of the Language Processor*, in Cooper & Walker (1979), pp. 27–85.

Freuder, E. C. & Wallace, R. J. (1992), 'Partial constraint satisfaction', *Artificial Intelligence* **58**, 21–70.

Garfield, J. L., ed. (1987), *Modularity in Knowledge Representation and Natural–Language Understanding*, MIT Press, Cambridge, MA.

Harper, M. P. & Helzerman, R. A. (1994), Managing multiple knowledge sources in constraint–based parsing of spoken language, Technical Report EE 94–16, School of Electrical Engineering, Purdue University, West Lafayette, IN.

Harper, M. P., Jamieson, L. H., Mitchell, C. D., Ying, G., Potisuk, S., Srinivasan, P. N., Chen, R., Zoltowski, C. B., McPheters, L. L., Pellom, B. & Helzerman, R. A. (1994), Integrating language models with speech recognition, *in* 'Proceedings of the AAAI–94 Workshop on the Integration of Natural Language and Speech Processing', pp. 139–146.

Harper, M. P., Jamieson, L. H., Zoltowski, C. B. & Helzerman, R. A. (1992), Semantics and constraint parsing of word graphs, *in* 'Proceedings of the International Conference on Acoustics, Speech and Signal Processing', pp. 63–66.

Helzerman, R. A. & Harper, M. P. (1992), Log time parsing on the MasPar MP–1, *in* 'Proceedings of the 6th International Conference on Parallel Processing', pp. 209–217.

Marslen-Wilson, W. & Tyler, L. K. (1987), *Against Modularity*, in Garfield (1987), pp. 37–62.

Maruyama, H. (1990*a*), Constraint dependency grammar, Technical Report RT0044, IBM Research, Tokyo Research Laboratory.

Maruyama, H. (1990*b*), Structural disambiguation with constraint propagation, *in* 'Proceedings of the 28th Annual Meeting of the ACL', Pittsburgh, pp. 31–38.

Menzel, W. (1988), Error diagnosing and selection in a training system for second language learning, *in* 'Proceedings 12th International Conference on Computational Linguistics, Coling '88', Budapest, pp. 414–419.

Menzel, W. (1990), Anticipation-free diagnosing of structural faults, *in* 'Proceedings 13th International Conference on Computational Linguistics, Coling '90', Helsinki, pp. 422–424.

Menzel, W. (1992), *Modellbasierte Fehlerdiagnose in Sprachlehrsystemen*, number 24 *in* 'Sprache und Information', Niemeyer Verlag, Tübingen.

Menzel, W. (1994), Parsing of spoken language under time constraints, *in* A. Cohn, ed., 'Proceedings of the 11th European Conference on Artificial Intelligence', Amsterdam, pp. 560–564.

Menzel, W. (1995), Robust processing of natural language, *in* 'Proceedings of the 19th German Annual Conference on Artificial Intelligence', Berlin, pp. 19–34.

Nudel, B. (1983), 'Consistent–labeling problems and their algorithms: Expected complexities and theory–based heuristics', *Artificial Intelligence* **21**, 135–178.

Schröder, I. (1996), Integration statistischer Methoden in eliminative Verfahren zur Analyse von natürlicher Sprache. Diplomarbeit, Fachbereich Informatik, Universität Hamburg, http://nats-www.informatik.uni-hamburg.de/~ingo/da/.

Schwind, C. (1995), 'Error analysis and explanation in knowledge based language tutoring', *Computer Assisted Language Learning* **8**(4), 295–324.

Struss, P. (1992), Knowledge-based diagnosis: An important challenge and touchstone for AI, *in* B. Neumann, ed., 'Proceedings of the 10th European Conference on Artificail Intelligence', Vienna, Austria, pp. 863–874.

Trueswell, J. C., Tanenhaus, M. K. & Garnsey, S. M. (1994), 'Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution', *Journal of Memory and Language* **33**, 285–318.

Tsang, E. (1993), *Foundations of Constraint Satisfaction*, Academic Press, Harcort Brace and Company, London.

Uszkoreit, H. (1991), Strategies for adding control information to declarative grammars, Research Report RR–91–29, DFKI GmbH.

Wallace, R. J. & Freuder, E. C. (1995), Heuristic methods for over-constrained constraint satisfaction problems, *in* 'Proceedings of the CP 1995 Workshop on Over–Constrained Systems'.
*ftp://ftp.cs.unh.edu/pub/csp/Papers/cp95-over-rjw-ecf.ps.gz

Weischedel, R. M., Voge, W. M. & James, M. (1978), 'An artificial intelligence appraoch to language instruction', *Artificial Intelligence* **10**, 225–240.

Yazdani, M. (1986), 'Intelligent tutoring systems: An overview', *Expert Systems* **3**(3), 154–162.