Errors, Intentions, and Explanations: Feedback Generation for Language Tutoring Systems

Wolfgang Menzel
Fachbereich Informatik, Universität Hamburg
Vogt-Kölln-Straße 30
D-22527 Hamburg
menzel@informatik.uni-hamburg.de

Abstract

This paper tries to adopt intentions as an additional dynamic source of knowledge, which can be used to bias the feedback generation components of a tutoring system for foreign language learning. By doing so, it departs from the common similarity-based approach, which derives the plausibility of an error diagnosis from the effort needed to transform the student's solution into the next best interpretation which can be established by the system. Given the limited knowledge of a tutoring system, this popular approach might lead to misguiding or even confusing explanations, especially in cases where alternative explanations would have been possible, but the system was not able to identify them properly or even needed to resort to arbitrary selections. The paper investigates two different options for integrating dynamically changing contextual expectations into the diagnostic analysis of erroneous utterances.

1 Introduction

Devising advanced language learning systems, which can be expected to provide helpful feedback without putting inacceptably strong limitations on the creative language production of their users, is a balancing act. On the one hand, the system designer is forced to restrict the complexity of a tutoring system's model as much as possible to arrive at a practical solution at all. On the other hand, the desire to produce good explanations requires to include a large body of nowledge about language and the non-linguistic context of language use into the model.

This fundamental bias shows up on every level of language description and can even be observed for rather simple problems like the correction of spelling errors. Starting with an analysis of this application domain in Section 2, we try to identify basic requirements for the design of a trustworthy diagnostic procedure and apply them to other areas subsequently. By considering morpho-syntactic and syntactic errors in Section 3 and Section 4, we successively relax the limitations of our input language. However, the less constrained our input language becomes, the more support from extra-linguistic sources will be needed, to counter the potential explosion of the space for potential error descriptions. An architecture for sentence parsing is presented, which not only is error sensitive, but also able to integrate the available knowledge into the decision on the most plausible interpretation for a student's utterance.

As useful as domain knowledge might be for error selection, it still falls short of the ideal point of reference for any explanatory effort: the intention of the learner, which every human teacher routinely makes use of. Here, the evidence can either be inferred from past behaviour of the student, domain knowledge, or the discourse context. If necessary, it can even be directly elicited from the student by asking

> "What did you want to say?"

where the student has both linguistic and non-linguistic means available to convey her message. We will discuss the necessity for similar solutions and possibilities to include them into a language learning environment in Section 5.

2 Error diagnosis

Ambiguity is pervasive in natural language. It is a phenomenon which can be observed on virtually all levels of linguistic description. Ambiguity, however, seems to hinge upon a limited scope of attention: it tends to disappear if language is put in context, be it a linguistic or extra-linguistic one. Deviating language, as it is a typical phenomenon in a language learning environment, obviously has the opposite effect. As it forces the hearer/reader to take errors into account, an additional source of uncertainty, hence ambiguity, is introduced into the language understanding process.

This phenomenon, perhaps, can best be illustrated with simple examples from the domain of spelling correction. Here, each correction proposal produced by a spell checker can be interpreted as a kind of error explanation. If we (a) restrict our diagnosis to only single-character substitutions, insertions, and deletions, (b) consider them to appear with equal probability, and (c) limit the scope of attention of our model to merely a list of potential word forms (i.e. excluding all syntactic, semantic, or pragmatic factors from consideration), the sentence

(1) **It is very kold in here.*

can be corrected by replacing the unknown word *kold* with at least the following alternatives

kold $\rightarrow$ {*bold, cold, gold, fold, hold, mold, sold, told, old*}

If the available knowledge is now extended to also include syntactic regularities, all but the adjectives are removed from this list:

kold $\rightarrow$ {*bold, cold, gold, old*}.

Considering also semantic fit would exclude additional candidates, leaving perhaps:

kold $\rightarrow$ {*cold, old*}

which e.g. by means of domain knowledge can be fairly safe disambiguated to a single proposal

kold $\rightarrow$ {*cold*}.

Of course, this solution could have been obtained more easily by directly considering domain knowledge right from the beginning.

Unfortunately, any restriction of the available knowledge in a system component requires us to even consider a (locally) acceptable input as being potentially wrong, since it might be so indeed, if later other knowledge sources are taken into account. This for instance happens, if the sentence had been

(2) **It is very told in here.*

which consists solely of perfect word forms but is inacceptable from a syntactic point of view.

Note that the picture painted so far is a highly simplified one, since the spurious ambiguity introduced by a spelling error grows considerably, if more complex types of errors are also modelled. Thus, the replacement of two characters introduces additional correction proposals like:

kold $\rightarrow$ {*bald, bolt, cord, colt, golf, kind, ...*}.

Usually, a simpler correction proposal is considered more plausible compared to a more complex one. As we will see, however, even this heuristics can be misleading.

Obviously, correction proposals as discussed so far are rather poor explanations of a spelling error, because they hardly give us any indication *why* the error might have appeared at all. Explanations in the stronger sense of the word would require to also distinguish underlying causes of an error, like the substitution of a character by a neighboring one on the keyboard, a phonetic similarity which might have led to confusion, or a misapplication of a word-formation rule. These different *error perspectives* provide us with better indications of how such an error can be avoided in the future. At the same time, however, they introduce yet another source of ambiguity into the diagnosis problem, while providing additional selection criteria for choosing the most plausible explanation. In example (2), for instance, a phonetic confusion seems most plausible, while an explanation based on keybord layout seems rather unlikely.

In some cases, one and the same error happens to be differently complex if it is considered from a different perspective. This is important, since less complex error descriptions are usually preferred for error explanation. Thus, in a sentence like

(3) **It was there fault.*

the most plausible explanation is a phonetic confusion of the word forms *their* and *there*, since both can be mapped to the same pronunciation. Therefore, no error assumption is required at a purely phonetic level. A keyboard based explanation, on the other hand, would require to assume two character substitutions instead, while still much cheaper corrections (like *there* $\rightarrow$ *here*) are available. Despite their apparent graphical similarity to the original form they should be disfavoured because they do not remove the syntactic inconsistency.

A number of lessons can be learned from these examples:

- The necessity to consider erroneous input increases (local) ambiguity. Enumerating all possible correction possibilities is neither feasible nor desirable.
- An error can be explained under quite different perspectives, where the particular perspective might influence the decision on the most plausible explanation, and plausibility usually is a gradual notion.
- Considering least effort corrections only, does not always yield acceptable error explanations.
- If possible, an error diagnosis should be aware of alternative explanation/correction possibilities, since these can turn out to be more plausible from another perspective.
- Considering an error might even be necessary if the input seems (locally) acceptable.

Taking all these requirements into account, a huge number of error hypotheses can be derived even for relatively simple problems. Therefore, most practical tutoring systems implicitly make simplifying assumptions. If, however, a more general solution is aspired to a need for strong constraints will arise, which can be used to narrow down the space of possible alternatives. There are two major sources of such constraints: (a) artificial limitations of the sublanguage under consideration or (b) (dynamic) domain knowledge derived from the current state of affairs, the latter approach being of more general interest. To make domain knowledge available to the feedback generation process two alternative architectures can be pursued:

1. An overgenerating diagnosis component produces an as large as possible set of potential error explanations, from which the most plausible one is selected, or
2. the domain knowledge is directly integrated into the diagnosis procedure to guide it towards the most plausible explanation.

While overgeneration can only be applied to highly restricted diagnostic problems, the direct integration of domain knowledge requires a sophisticated infrastructure for information fusion. Both approaches and their limitations will be discussed in the following sections.

3 Morphosyntax

Morphosyntactic regularities in natural language are typically described by means of features and relations such as agreement or government. Agreement violations e.g. can then be interpreted as a lack of a particular rule in the current grammatical knowledge of the student or as a sign of this rule accidentally being ignored. There is, however, a different interpretation available, assuming that the student is fully aware of the corresponding rule but erred with respect to the feature assignment in the dictionary. Similar to the situation with spelling errors, we have now a different perspective to describe an error like the one in

(4) **We expect many child to come.*

On the one hand, there is a *grammar rule* not obeyed by the student, i.e.

> The noun *child* and its determiner *many* have to agree in number.

on the other, there are two alternative hypotheses for describing which *factual information* from the dictionary has not been properly taken into consideration

> The determiner *many* is plural but should be singular.
> The noun *child* is singular but should be plural.

Diagnoses of this kind can be obtained by successively *retracting* an increasing number of *constraints*, used to model the underlying agreement requirements an lexical feature assignments of a particular linguistic construction. Note that the example above only serves to illustrate the most simple case of feature agreement. More difficult diagnostic problems result, e.g. from the occurrence of morphosyntactically ambiguous word forms like *fish* in

(5) **These fish smells.*

which can even lead to alternative rule error descriptions. More complex agreement networks are typically found in languages with a richer inflectional morphology (Menzel, 1988; Menzel, 1990; Menzel, 1992). Figure 1 shows the agreement requirements within a German prepositional phrase. The network consists of nine agreement and fourteen value assignment constraints (only the former ones are shown in the picture).

Most of the existing proposals for diagnosing agreement requirements simply ignore some types of possible error interpretations, by either restricting themselves to only a single perspective or by resorting to an a priori selection schema. While (Schwind, 1995) mainly focusses on rule errors, (Heift, 1998) is biased towards fact error diagnoses, which are directly coded into HPSG grammar constraints.

Both options, however, are quite unsatisfactory, because different perspectives as well as explanation alternatives might result in error descriptions which can be explained with different effort. Consider for example the following sentence

(6) **Many child play in the backyard.*

Ignoring the unlikely conjunctive reading of *play*, a rule error description based on two constraint violations is obtained

> The noun *child* has to agree with its determiner *many* and the governing noun *play* in number.

Again, there are two fact error descriptions, however, of different complexity. While the first one is based on two constraint violations

> The determiner *many* and the finite verb *play* are plural but should be singular to agree with the noun *child*.

the second one requires to assume only a single violation

> The noun *child* is singular but should be plural to agree with the finite verb *play* and the determiner *many*.

Usually, algorithms are used which implicitly apply a simplifying heuristics to obtain a unique error description in such a case. Such heuristics could be:

- Dominance: Always correct at the syntactically dependent word form, or
- Economy: Always choose the description with a minimum number of constraint violations.

Obviously these heuristics are partly contradictory and the outcome crucially depends on which one is taking preference. While for example (6) the dominance criterion gives no clear guidance, economy considerations coincide with the intuition. Note, however that this result has been obtained by chance, since the intuitive plausible correction at the noun is favoured because of the number feature at the determiner being asymmetric for semantic reasons: Assuming the student to choose *many* although she wanted to speak about a single child must be considered highly implausible.

Of course, this kind of reasoning fails in the complementary case

(7) **Many child plays in the backyard.*

where a rule error description with a single constraint violation

> The determiner *many* has to agree with the noun *child* in number.

corresponds to two differently complex fact error descriptions:

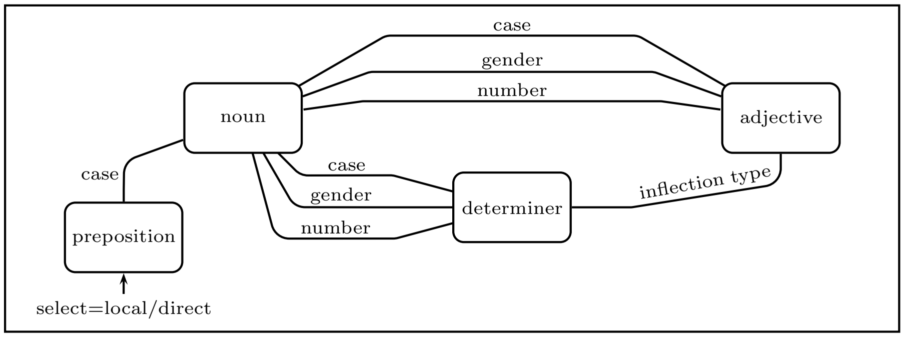

Figure 1: A constraint based-model for the agreement within a German prepositional phrase

The noun *child* and the finite verb *plays* are both singular but should be plural to agree with the determiner (2 errors).

The determiner *many* is plural but should be singular to agree with the noun (1 error).

Here, both heuristics prefer the second fact error description, thus being in obvious conflict with our intuition. Unfortunately, there seems to be no heuristics available which can produce the most plausible explanation for all such cases, without giving the diagnosis procedure access to additional knowledge, like the semantic preference as discussed above. If this cannot be achieved easily, *all* interpretations will need to be computed for a later selection.

Note that it is not even sufficient to compute alternative interpretations with a single perspective in mind, since the results do not always correspond to a description from another perspective in a one-to-one manner. If, for instance, correction proposals (i.e. fact errors) are directly derived from rule error descriptions, the sentence

(8) **Many children plays in the backyard.*

produces a rule error description with a single constraint violation

The subject noun phrase *many children* has to agree with the finite verb *plays* in number.

Hence the dominance heuristics would prefer a correction at the depending noun, which however results in a contingent error as discussed e.g. in (Heift, 1998): Subsequently the determiner has to be corrected as well. Even worse, the more obvious correction at the finite verb is no longer available. The problem can only be avoided, if fact errors are diagnosed independently, thereby producing both correction proposals

The noun *children* and its determiner *many* are both plural but should be singular to agree with the finite verb (2 errors).

The verb *plays* is singular but should be plural to agree with the subject noun phrase (1 error).

More serious problems arise, if the diagnosis problem is artificially restricted to local domains. Usually this happens if the error diagnosis is integrated into the rule application of a classical parser based on phrase structure rules, where local decisions are enforced although the constraint net is only partially visible within the scope of a rule. This almost certainly results in misleading diagnoses in some cases like in the famous German example taken from (Schwind, 1995)

**Der Götter zürnen.* (The gods are angry.)

where morpho-syntactic ambiguity leads to a situation in which *der Götter* forms a perfect noun phrase. Since this local correctness assigns genitive case to the noun phrase, a constraint violation one level up will occur (the subject has to be nominative). The equally complex, but much more plausible error description which assumes a missing number agreement between determiner and noun is not visible to the diagnosis, because it requires to acknowledge a case agreement crossing a phrase boundary. As a consequence, errors are percolated up the syntax tree as far as possible, which clearly contradicts the dominance heuristics. Other examples for misleading explanations, generated by diagnosis procedures not general enough, can also be found in (Holland, 1994)).

Both contingent errors and implausible diagnostic results in fact turn out to be artefacts of implicitly biasing the procedure towards a particular type of error decriptions. They can be completely avoided by using a more general approach which produces a rich enough variety of error descriptions. These description can then be subjected to a *post hoc* selection of the most appropriate error explanation, which takes into account domain knowledge or, even better, hypotheses about the communicative intent of the student. If, for instance, in the domain under consideration only a reference to a single child makes sense, any proposal to correct *child* to *children* would not be very helpful. Again, this kind of inference strongly relies on the assumption that the student indeed is referring to the scene on display. Especially in case of doubt or if the current state of affairs in the domain does not allow to uniquely

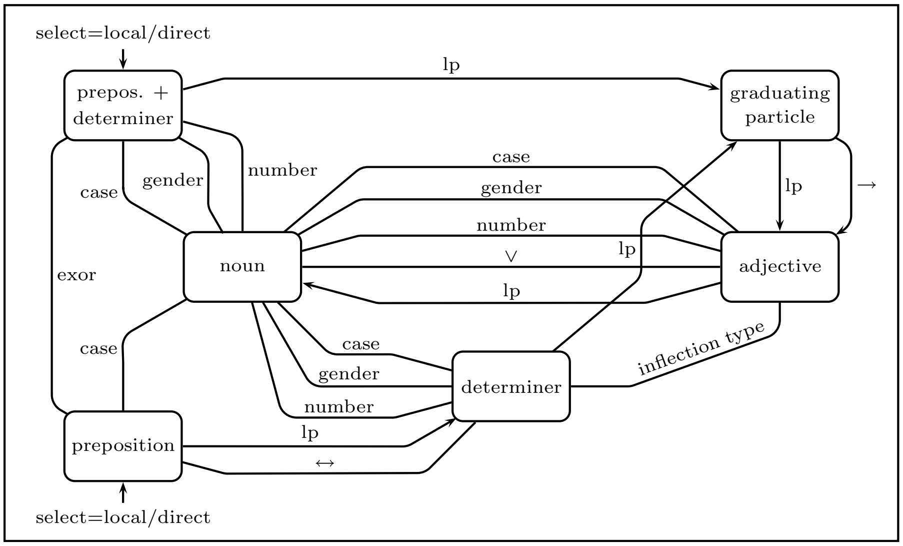

Figure 2: A constraint based-model for a German PP including optionality and linear ordering

determine a proper reference, it might be more appropriate to ask back whether the student wanted to talk about one particular child or a number of them.

Of course, any exhaustive diagnosis which also supports multiple perspectives is not only rather expensive to compute, but also produces a potentially very large set of different error explanations. If we now attempt to extend the approach to the diagnosis of less constrained utterances occurring in natural discourse we run into serious efficiency problems. Due to the recursive nature of language the set of error explanations can even become infinite in the extreme case. Thus, the overgenerating approach seems to be restricted to artificially constrained exercises in limited domains of grammatical constructions. Although it can be extended to simple structural relationships like linear precedence and dominance (c.f. Figure 2), it is no longer feasible if realistic communication scenarios are considered. Such tasks will require a diagnostic approach which directly integrates hypothesis selection into hypothesis generation. We will propose a corresponding architecture in the next section.

4 Syntax

An integration of hypothesis generation and selection needs to fulfill three major requirements:

- It supports constraint retraction to model constraint violations and their non-local consequences.
- It facilitates the integration of (possibly dynamic) influences from the context.
- It is based on a mechanism for plausibility-based reasoning which enables it to arbitrate between possibly contradicting pieces of evidence.

The formalism of weighted constraint dependency grammar (WCDG) has been devised with these three requirements in mind (Schröder et al., 2000; Foth et al., to appear). All conditions on grammatical well-formedness are expressed by means of weighted, i.e. retractable constraints, and different procedures for constraint solving are available, which can determine the optimal interpretation even for a possibly highly distorted utterance.

Due to the use of weighted constraints, WCDG can demonstrate an excellent potential for the fusion of uncertain evidence from quite different information sources (Daum et al., 2003). It is therefore in a position to successfully counter the huge amount of spurious ambiguity introduced by erroneous language in addition to what must be considered inherent structural ambiguity of natural language. Such problems are typically caused by the necessity to consider multiple attachment possibilities as in

(8) *the apple in front of the plate with the green dots*

or by the uncertainty of assigning syntactic functions in free word order languages, like e.g. in German

(9) *Schlaf braucht das Kind.* (lit. Sleep needs the child).

Although there is a clear preference for the subject preceding the object in German, it is misleading in this case, since it can be overridden in a topicalized construction.

Different kinds of language regularities are modelled by constraints with differently strong weights. Thus, the violation of e.g. an agreement requirement is heavily punished, but not strictly enforced, while linear precedence relations like the one between subject and object in a German clause receive a considerably smaller penalty. Mere preferences, which e.g. favour short attachments over longer ones, are ranked even lower.

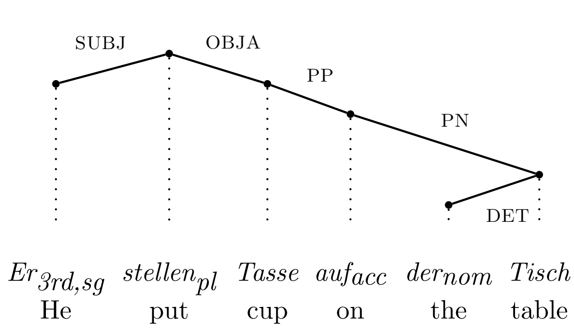

Figure 3: A dependency tree for an ungrammatical German sentence

It is one of the attractive properties of WCDG that constraint violations can easily be interpreted as error explanations. Thus, WCDG parsing is error sensitive by definition without further provision. Since preferential knowledge is used extensively throughout the grammar, even a completely perfect utterance will usually cause some constraints to be violated. Constraints are therefore grouped into serious and less serious ones with only the former ones being explained to the student. Figure 3 shows the dependency analysis of a distorted German sentence[1], which raises three serious constraint violations, namely

- the number agreement between the subject *Er* and the finite verb *stellen* has not been established,
- there is no case agreement within the prepositional phrase *auf der Tisch*, and
- the direct object *Tasse* lacks a determiner.

Since in WCDG the diagnosis process is completely integrated into the normal course of sentence parsing, considerably more complex constructions as the one discussed in Section 3 can now be handled. Among the phenomena properly treated are linear ordering regularities and the omission of obligatory material (e.g. obligatory complements). Inserting superfluous material usually causes the parse tree to fall apart into fragments. Since actual insertions are really rare in student utterances, they can be neglected for practical purposes (Reuer, 2003). Instead, whenever the parser returns tree fragments for an utterance, this might indicate a category confusion, which can be identified using simple pattern-based transformation rules (Menzel and Schröder, 1998) (c.f. Figure 4).

For reasons of efficiency the expressive power of constraints has been limited to consider only two word-to-word dependencies at a time. This has the unfortunate consequence that e.g. agreement requirements that can be composed into rather long

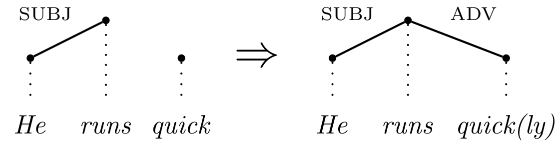

Figure 4: Example of a pattern-based transformation rule for the diagnosis of improper category selections

chains of conditions can only be diagnosed approximately. Moreover, since the error-sensitive parser always tries to obtain the single best interpretation of an utterance, only a single error perspective (namely the rule based one) is supported and alternative interpretations are not available. Therefore we coupled a WCDG based diagnosis, which can determine the optimal syntactic structure of a sentence but produces imperfect agreement diagnoses, with a thorough simulation of agreement errors as described in Section 3, which requires the structure be known in advance but can produce the desired multi-perspective results for agreement errors. Thus, the respective advantages of both procedures are combined for mutual benefit (Menzel and Schröder, 1998; Menzel and Schröder, 1999).

5 Intentions in error diagnosis

To effectively constrain the huge hypothesis space of a general diagnostic procedure requires both a strong support from external knowledge and powerful mechanisms to integrate this knowledge into the diagnostic engine. Therefore the traditional approach to produce semantic interpretations from syntactic structures, has to be extended to also accept semantic expectations when building a syntax tree. Such a bidirectional coupling has first been introduced in the framework of head-driven phrase structure grammar (HPSG). HPSG, however, suffers from being based on classical two-valued logic. Therefore, uncertain and preferential knowledge can not be handled properly and the utility of the approach for error diagnosis is limited.

WCDG on the other hand, carries out a plausibility-guided search taking into account all the evidence available and therefore is able to integrate additional knowledge sources in a soft manner. This soft integration makes sure that less plausible interpretations are not simply discarded and therefore lost irrevocably, but instead are disfavoured to a certain degree. They have to compete with alternative interpretations and can even survive as long as no better solution becomes available.

Simular to HPSG, additional description levels can be established parallel to the main syntactic one. The mapping between these structures, however, is not established by means of (crisp) coreference relationships, but makes use of defeasible constraints again, hence providing a fail-soft mechanism which is

[1] An online demo of the parser (with a grammar optimized for broad coverage of German, not for error diagnosis) can be found on the web under
`nats-www.informatik.uni-hamburg.de/parse/Papa/ParserDemo`

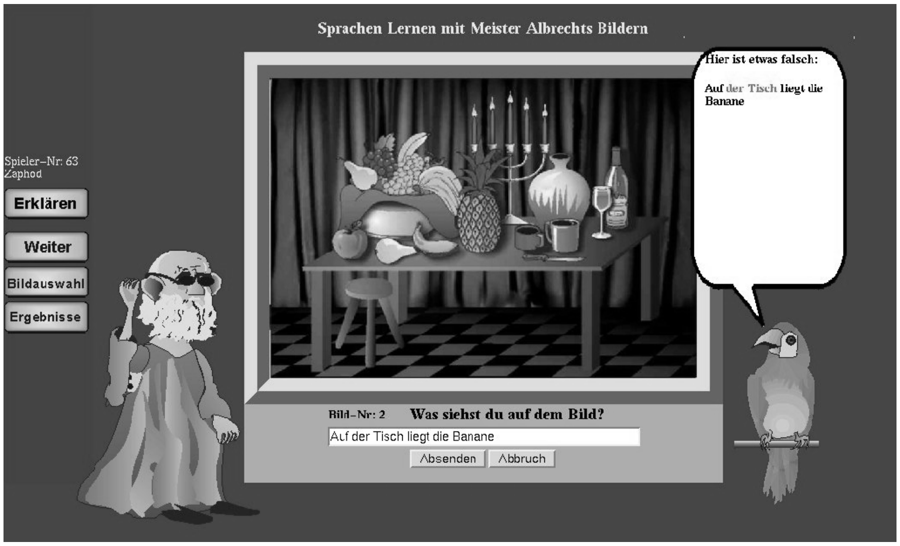

Figure 5: A static micro-world used in a (web-based) language-driven computer game.

important in cases of local conflicts between the different representations. This approach makes it possible to integrate a simple semantic representation into the error sensitive syntactic parser, which by providing access to a dynamically changing knowledge base can be biased towards a plausible interpretation.

Having a bidirectional interface to semantic structures available, which makes the error-sensitive parser susceptible to external influences, we still have to identify promising sources and the most effective way to gain access to the desired biasing information. The most simple approach makes use of a small scene or situation presented to the student by means of language, pictures or a combination thereof. Figure 5 shows such a scene, which has been used in a web-based computer game, where the student has to describe the picture to a blind artist who tries to remember, if he ever painted such a still. The parrot comments on possible language errors.

The propositional content of the scene can now be fed directly into the error sensitive parser. This however is equivalent to making the assumption that the student is really talking about the scene on display. Therefore, an effective guidance of the available domain knowledge can only be expected if

- the student is cooperative, i.e. complies to the task assigned by the system, and
- the number of options to communicate something about the micro-world is suffiently small given the current state of affairs.

Note that rather strong restrictions on the set of possible options are highly desirable from the systems point of view, since a strong external influence greatly increases the chances to arrive at a plausible error diagnosis. On the other hand, too strong limitations might increase the risk of the student trying to escape from them. Therefore, only rather simple domains can be used, because otherwise there will be too many things to talk about and the utility of contextual contributions is considerably diminished.

A more attractive alternative would therefore be the presentation of a dynamic scene as used e.g. in the system *PromisD* (Reuer, 2003), where the student is asked to describe a car accident to a policeman. Dynamically evolving scenarios provide quite reliable possibilities for focussing the attention of the student on the changing aspects of the scene. However, they require to take temporal aspects into consideration, which usually do not easily map to natural language expressions. Even if these problems can be solved satisfactorily, the basic assumption of the knowledge-based approach remains the same: the student is expected to only talk about what has been considered important by the system designer.

Whenever we want to give the student a higher degree of freedom for formulating her contributions, we need to search for an alternative source of constraining evidence. This gap could be filled by hypotheses about the *intentions* of a student. In contrast to general knowledge about the discourse domain, speaker/writer intentions have the major advantage that they are not only precisely focussed on the diagnostic problem at hand, but can also be used as an excellent background for error explanations:

> "If you wanted to say this, it should have been that."

To gain access to this additional knowledge source, the system can ask the student to further elaborate on what her intention was. As in a human-to-human setting linguistic and graphical means are available for communicating the intended propositional content back to the system. This, however, would require a truly interactive microworld, where in addition to providing content and training exercises the system also accepts natural language commands or direct manipulation gestures (Hamburger, 1995). Besides quite traditional interaction modes like

- *quizmaster*: The system asks something and the student answers, and
- *movecaster*: The system shows something and the student describes it,

others which require a more active involvement of the student could also be realized:

- *oracle*: The student asks and the system answers,
- *servant*: The student gives a command and the system carries it out, or
- *interpreter*: The student tells a story and the system displays it.

From the perspective of an intention-based diagnosis an interactive environment becomes particularly interesting, because it offers the possibility for the student to explain her communicative intent by purely non-linguistic means. If the given task in a microworld e.g. consists of preparing a breakfast table, the student can easily illustrate the propositional content of an utterance by dragging and dropping various objects to the desired position. Such a non-linguistic approach even becomes the only feasible option whenever the direct communication in the language to be learned breaks down due to insufficient proficiency of the learner. Given the sophistication of contemporary graphical user interfaces, fairly complex propositions can be communicated that way. This contrasts advantageously with a direct language-based clarification dialogue, which in cases of communication failure can only be used to communicate very simple types of information, e.g. the number of reference objects as dicussed in Section 3.

6 Conclusions

We have proposed to use communicative intentions of a language learner as an additional source of evidence when diagnosing grammatical errors she commited. Having such an information available would enable a tutoring system to produce error explanations which are grounded in what the student actually wanted to say. Intentions can either be determined by asking the student to further elaborate on certain aspects of her utterance (e.g. the number of referential objects in the domain) or to demonstrate the intended propositional content by means of a graphical user interface using direct manipulation gestures.

Two different architectures have been discussed for including the additional information into diagnosis and feedback generation, one which uses a post hoc selection from the results of an overgenerating diagnosis, and one which attempts to integrate the available propositions directly into the diagnosis procedure. While the former is limited to rather tiny types of exercises only a combination of both led to a solution in which both approaches complement each other.

To directly integrate the external evidence into the diagnosis, an architecture for information fusion has been devised which is based on weighted, hence retractable constraints. Such a grammar model can accomodate even dynamically changing propositions and use this information to guide the analysis procedure towards the desired interpretation.

References

M. Daum, K. Foth, and W. Menzel. 2003. Constraint-based integration of deep and shallow parsing techniques. In *Proc. 11th Conf. of the Europ. ACL*, pages 99–106, Budapest.

K. Foth, W. Menzel, and I. Schröder. to appear. Robust parsing with weighted constraints. *to appear in: Natural Language Engineering.*

H. Hamburger. 1995. Tutorial tools for language learning by two-medium dialogue. In V. M. Holland, J. D. Kaplan, and M. R. Sams, editors, *Intelligent Language Tutors*, pages 183–199. Lawrence Erlbaum Associates, Hillsdale, NJ.

T. Heift. 1998. *Designed Intelligence: A Language Teacher Model.* Ph.D. thesis, Simon Fraser Univ.

M. V. Holland. 1994. Lessons learned in designing intelligent CALL: Managing communication across disciplines. *Computer Assisted Language Learning*, 7(3):227–256.

W. Menzel and I. Schröder. 1998. Model-based diagnosis under structural uncertainty. In *Proc. 13th Europ. Conf. on Artificial Intelligence*, pages 284–288, Brighton.

W. Menzel and I. Schröder. 1999. Error diagnosis for language learning systems. *ReCALL*, Special issue:20–30.

W. Menzel. 1988. Error diagnosing and selection in a training system for second language learning. In *Proc. 12th Int. Conf. on Computational Linguistics*, pages 414–419, Budapest.

W. Menzel. 1990. Anticipation-free diagnosis of structural faults. In *Proc. 13th Int. Conf. on Computational Linguistics*, volume III, pages 422–424, Helsinki.

W. Menzel. 1992. *Modellbasierte Fehlerdiagnose für Sprachlehrsysteme.* Sprache und Information. Niemeyer, Tübingen.

V. Reuer. 2003. *PromisD: Ein Analyseverfahren zur antizipationsfreien Erkennung und Erklärung von grammatischen Fehlern in Sprachlehrsystemen.* Dissertation, Humboldt-Universität zu Berlin.

I. Schröder, W. Menzel, K. Foth, and M. Schulz. 2000. Modeling dependency grammar with restricted constraints. *Traitement Automatique des Langues (T.A.L.)*, 41(1):97–126.

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