Resolving Ambiguity in German Adjectives

Amanda NICHOLAS and Brent MARTIN
Intelligent Computer Tutoring Group (ICTG),
University of Canterbury, Christchurch New Zealand.

Abstract. One problem in ill-defined domains is accurately identifying the source of errors. Obtaining sufficient information about the error can be difficult because doing so may interfere with the learning task.

In this paper we present the results of an experiment in the domain of German adjectives. We trialed a modified student interface that gathers more data during problem solving by requiring the student to perform a related subtask. There is evidence that the students who performed the subtask outperformed the control group on a post-test despite the extra task slowing them down, suggesting the extra effort required by the students to overcome ambiguity was worth the intervention.

Keywords. Student Modeling, Language learning, Ambiguity

Introduction

Dealing with ambiguity is a serious problem in developing Intelligent Tutoring Systems for foreign languages [1]. Natural language processing has not yet reached the point where we can process an unconstrained statement made by a student and accurately identify the source of any errors [2]. By constraining the scope of statements made by the student, it is possible to mark an answer as correct or incorrect. However, although the system can detect that the student has made an error, the source of this error may be difficult to determine. Menzel defines four sources of ambiguity: limited observability, polysemy, alternative conceptualizations of domain knowledge and structural uncertainty. In a domain with high ambiguity, feedback messages can be difficult to determine. Good feedback should refer the student to the underlying domain principle [3]. If it is not possible to determine which domain principle has been broken, correctly targeted feedback cannot be given.

One approach to avoid ambiguity is to require the student to specify the intermediate steps they carry out mentally. This approach is not popular; “such an interaction renders the exercise somehow unnatural.” [1]. Requiring the student to specify intermediate steps also raises the issue of transference [4]. When developing an ITS, the interface is generally designed to stay as close to the real world as possible, in order to ensure that the skills learnt on the computer will transfer to the real situation. By requiring the student to specify additional information the transference of skills may be weakened. This research compares two constraint-based (CBM) tutors; one that matches the real world more closely, and one that decreases ambiguity. Two Intelligent Tutoring Systems were

developed for the domain of German adjective endings, a domain where errors have a high level of ambiguity. An error in an adjective ending could be caused by a number of factors: mistaking the case, mistaking the gender, mistaking the article, or simply not knowing the correct ending in this situation. The experimental tutor required the student to specify the gender of the noun, the case of the noun, and the type of article. These three factors specify exactly the ending required. By considering these factors, the tutor could determine if it was the ending of the adjective that the student had specified incorrectly, or if they had some misconceptions about the sentence. The tutoring system could then provide targeted feedback based on the student's misconceptions. In contrast, the control only asked the student to specify the adjective ending. This matches what is required in the real world, but means that the tutoring system has to make assumptions about the error the student has made.

In the next section we further describe the problem of ambiguity in the domain of German adjectives, and constraint-based modeling (CBM) is summarized in Section 2. Sections 3 and 4 describe the experiment and present the results. Finally, we conclude in Section 5.

1. Ambiguity in German Adjectives

Adjective endings are a difficult topic for students to master. This is due to the number of endings that must be memorized, and the amount of knowledge required of the sentence to get the ending correct. Rogers studied the main areas of weakness in students with more than four years of experience learning German [5]. She states "...much anecdotal 'evidence' from teachers of German as a foreign language emphasizes morphology as a major areas of weakness (e.g. adjective endings...)". Her study showed that approximately 5% of errors made by advanced learners of German were errors in adjective endings. The number one error was in selecting gender, which could also affect the choice of adjective ending. Each error was only classified once, so if the student mistook the gender, it would not also appear as a mistaken adjective ending. The number of errors in adjective endings is therefore likely to be much higher than 5% when all reasons are considered. Further, Juozulynas studied students with two years of experience learning German and found that "The biggest problem in the students' writing seems to be syntax ... inflectional morphology with its much-feared endings takes second place. Syntax and morphology together make up 53% of the errors in the corpus." [6] Note that adjective endings are contained in inflectional morphology.

Case	Masculine	Feminine	Neuter	Plural
Nominative	-e	-e	-e	-en
Accusative	-en	-e	-e	-en
Genitive	-en	-en	-en	-en
Dative	-en	-en	-en	-en

Table 1. Adjective ending when preceded by the definite article

In German, adjectives must agree with the nouns they modify. This means that the ending of an adjective varies based on the gender and the case of the noun, and whether

the noun is preceded by the definite article, indefinite article, or no article. Table 1 lists the endings for the case where an adjective is preceded by the definite article. For example, take the sentence “Das graue Haus ist neu”. (The gray house is new). Here “Haus” is the noun, and its gender is neuter. The house is the subject of the sentence, and so it is in the nominative case. The article is “das”, and it is the direct article. The adjective is “grau”, and it takes the ending “e” because, by consulting Table 1, we see that adjectives preceding a neuter noun in the nominative case must end in “e”. If we change only the article in this sentence, so that it now read “Ein graues Haus ist neu”. (A gray house is new), the ending on the adjective changes also, from “e” to “es”. It is important to note that the endings are not unique; the ending “e” appears in a number of situations, as does “en”. This is one reason why these endings are ambiguous.

Menzel identified four major sources of ambiguity that should be considered when creating CBM tutors, particularly for foreign languages [1]. These are: a limited observability of internal variables of the problem domain; polysemy of symbols used in the problem domain (symbols with multiple meanings); alternative conceptualizations of domain knowledge; uncertainty about the intended structure of the students solution. He further suggests that because of this constraints alone are not sufficient to provide enough information to respond to students appropriately. German adjective endings suffer from three of the four defined sources of ambiguity. Limited observability and polysemy are both present in the multiple possible meanings of a single ending. For example, a student could correctly give an adjective requiring a nominative, masculine, definite article ending, the ending “e”. However, it is also possible that the student thought that the adjective required a nominative, feminine, definite article ending, for which the ending is also “e”. The student might even believe that the adjective requires a dative, masculine, definite article ending, which should be “en”, and might have given it the ending “e” incorrectly, based on their (incorrect) knowledge. Without awareness of the student’s thought processes, the tutor is unable to determine if the student has answered the question correctly on purpose, or by mistake. This problem also encompasses that of alternative conceptualizations of domain knowledge. When the student incorrectly gives an adjective ending, it could be due to either a rule error or a fact error. If the student does not know the gender or the case of the noun, they have made a fact error. If the student has correctly determined the case, gender and article, and still gives the adjective ending incorrectly, they have made a rule error; they do not know the underlying grammatical principle that determines the adjective ending. It is also possible for a student to make a rule error and a fact error simultaneously.

2. Constraint-Based Modeling and German Adjectives

CBM[7] is a relatively new approach to domain and student modeling, based on the theory of learning from performance errors [8]. It models the domain as a set of state constraints, where each constraint represents a declarative concept that must be learned and internalized before the student can achieve mastery of the domain. Constraints represent restrictions on solution states, and take the form:

IF <relevance condition> is true for the student’s solution,
THEN <satisfaction condition> must also be true

The relevance condition of each constraint checks whether the student's solution is in a pedagogically significant state. If so, the satisfaction condition is checked. If it succeeds, no action is taken; if it fails, the student has made a mistake, and appropriate feedback is given.

The student model consists of the set of constraints, and information about whether or not each constraint has been successfully applied each time it is relevant. Thus the student model is a trace of the performance of each individual constraint over time. Constraints may be grouped together, giving the average performance of the constraint set as a whole over time, which can then be plotted as a learning curve [9,11].

3. Experiment Design

We hypothesized that forcing the students to supply information about their problem-solving process and providing feedback based on that information would enable the system to give them better instruction, and thus they would be better able to learn the domain. We tested this hypothesis by building two versions of an ITS for German adjectives, where the two systems differed in the interface used and the underlying domain/student model (constraints).

The tutors were developed using WETAS [10]. WETAS is a shell that can be quickly adapted to provide basic functionality for an ITS. It provides student modeling, student management, and other features. The developer must supplement this with the problem set, the necessary constraints and, if desired, an interface. The problem set comprised of 55 problems, which was identical for both tutors. Some were obtained from existing sources [12,13], however, most problems were written especially for this ITS. An example of one of the problems in the tutor is

"Die ? Blumen gefallen mir. (bunt)" (I like the colorful flowers)

The two tutors shared a very similar interface. In the center of the screen was an area for the student to answer the question. Below the problem, a selection box allowed the student to choose the desired feedback level, and a button to submit their answer for feedback. Feedback messages appeared at the bottom of the screen. The problem was displayed in the form of a sentence. A gap was left where the adjective should be, and the adjective to be inserted was given in brackets at the end of the sentence. This was a format the students were familiar with, because it had been used during class and quizzes.

Students using the experimental system were asked to fill in the gender and case of the noun, the article type, and the adjective with its ending. The possible answers for gender, case and article were all given in combo boxes. This ensured that there would not be problems with students referring to the same concept by a different name, or misspelling names. Below the combo boxes, there was a text field for the student to fill in the adjective. Students using the control were only asked to fill in the correct adjective and ending. A textbox for the student to fill in was placed in the correct location in the sentence.

Domain constraints were sourced from a number of German textbooks [13,14,15,16, 17], which contain advice on how students can remember the endings more easily. They typically explain a pattern in the endings, for example that every adjective after the direct article ends in either "e" or "en" (see Table 1). The resulting constraints can be divided

into three groups. The first set of constraints is used for error checking, ensuring that the student has answered the question and used the appropriate adjective. The second set occurs only in the experimental tutor and checks whether the student has specified the gender, case and article correctly. The third set of constraints is the group that checks the validity of the adjective ending.

The experimental tutor has 33 constraints. Six are for error checking; ten are for checking that the student has specified the case, gender and article correctly; the remaining constraints check the adjective ending. The adjective ending is checked for validity with respect to the case, gender and article the student has used; incorrect values for case, gender and article will trigger other feedback messages. In this manner, the system determines whether the student has made an error because they have inaccurate knowledge about the sentence, or because they do not know their adjective endings, i.e. whether they have made a fact error or a rule error.

The control tutor had twelve constraints. Three were for error checking, and the remaining nine checked the ending the adjective has been given. Because the only information available to the tutor is the ending the student has given the adjective, the tutor provides feedback relative to the correct gender, case and article. It is assumed that the student knows this information, but may be unaware of the ending that matches correctly. This means that the tutor considers all mistakes to be rule errors, not fact errors. An example of one such constraint is:

```
(10
; FEEDBACK
"When they are preceded by a 'der-word', all adjectives end
in either -e or -en."
; RELEVANCE CONDITION
(and
(match IS ARTICLE ("D"))
(match SS ANSWER (?something ?*)))
; SATISFACTION CONDITION
(or-p
(match SS ANSWER (?*w2 "e" "n"))
(match SS ANSWER (?*w1 "e")))
"ANSWER")
```

The relevance clause of this constraint checks that the sentence contains a definite article (“D”), and that the student has answered the question. If this is true, the student’s answer must end in “e” or “en”, as all adjectives end in “e” or “en” after the definite article. If the student’s answer does not end in “e” or “en”, the system assumes that they have forgotten the rule, not that they have not realized that the sentence contains a definite article.

An evaluation study of the two tutors was conducted on the 6th of September 2006 at the University of Canterbury, Christchurch. Students enrolled in GRMN115, a beginning German course, used one of the two systems over one class period. The students had been taught adjective endings previously in class, however there was a two-week holiday period between when the topic was taught and when the study was carried out. The class was divided into two even groups. This was done alphabetically by last name. The evaluation took place during one lecture period, a time span of 50 minutes. The students

were first asked to complete a pre-test. They then used the tutoring system for as long as time permitted, or until they finished all 55 questions. Afterwards they completed a post-test. Each test contained six questions. All questions contained sentences of the form:

"Die ? Jacke ist preiswert. (gelb)" (The yellow jacket is affordable)

The student was expected to transfer the adjective (here 'gelb') into the gap in the sentence, and give it the appropriate ending. The final three questions also asked the student to specify the gender and case of the noun present in the sentence, and the type of article preceding the noun. The experiment was carried out in two streams. The control and experimental tutor were used by students from both streams. To allow for any difference in the difficulty of the pre- and post-tests, Test 1 was used as the pre-test for Stream A, and the post-test for Stream B; Test 2 was used as the post-test for Stream A and pre-test for Stream B.

4. Results

23 students took part in the evaluation. 12 students used the experimental tutor and 11 students used the control. Statistics about the system usage can be seen in Table 2. We can see that students using the control system solved more problems with fewer attempts than those using the experimental tutor. This result is unsurprising, because students using the control only had fill in one answer correctly, whereas students using the experimental tutor had to fill in answer values. Students using the experimental tutor also saw more messages. This is also unsurprising; their task was larger so there were more opportunities to make mistakes.

Measure	Control	Experiment
Attempted Problems	52	22
Solved Problems	49	21
Attempts per Problem	2.0	4.0
Seen Messages per Problem	1.5	5.0

Table 2. System usage statistics

Unfortunately, the pre- and post-test were not of comparable difficulty. Over all students, irrespective of which tutor the student used or whether the test was taken as a pre- or post-test, the average score for Test 1 was 83%, and the average score for Test 2 was 65%. This means that the scores for the pre-and post-test are not directly comparable. The reason for the difference in difficulty is that Test 2 contained two questions where the gender of the noun could not be unambiguously determined from the rest of the sentence; the student either knew the gender of the word or they did not. To overcome this, we compared the results for Test 1 only, and compared the outcome for pre- and post-test regardless of which stream the students belonged to. This is not strictly valid because the samples are different; it relies on the assumption that the students in the two streams (and using the same tutor) were comparable, and this cannot be easily measured. Using this assumption, a t-test of the score for producing the correct adjective ending showed

no significant difference between the Test 1 pre-test scores for the two tutors (mean = 4.8 and 4.6 for the control and experimental groups respectively, SD = 0.8 and 1.6, $p > 0.7$). When Test 1 was used as a post-test however, there was a larger difference between the two groups, with the experimental tutor achieving a score of 5.7 compared to 5.0 for the experimental group, although the result is not statistically significant ($p > 0.15$).

We also compared the performance of the two groups in terms of their ability to perform the subtask (determine case and gender). Again there was no significant difference on pre-test score between the control and experimental groups (5.0 versus 4.9). For the post-test, the experimental group again outperformed the control group, scoring an average of 5.7 compared to 4.8 for the control group. The result was statistically significant ($p < 0.05$).

Another method of comparing student performance is via learning curves [9,11]. If the units being measured are being learned by the students, we expect to see a "power law of practice". Learning curves therefore give an indication of the relative performance of samples of students and the quality of the model. Fig. 1 shows the learning curves for the two groups for just those constraints that test for the correct adjective endings (Tutor1 is the experimental group, Tutor2 is the control). The power law fit for the curves for both groups is only average, although it is better for the experimental group (R^2 = 0.63 versus 0.45). Fig. 2 shows the corresponding learning curve for the only those constraints that test the subtask. This latter curve is for the experimental group only since these constraints did not exist for the control group. In this case we see a slightly better power law (R^2 = 0.71), suggesting that the constraints form a fairly good model of what is actually being learned.

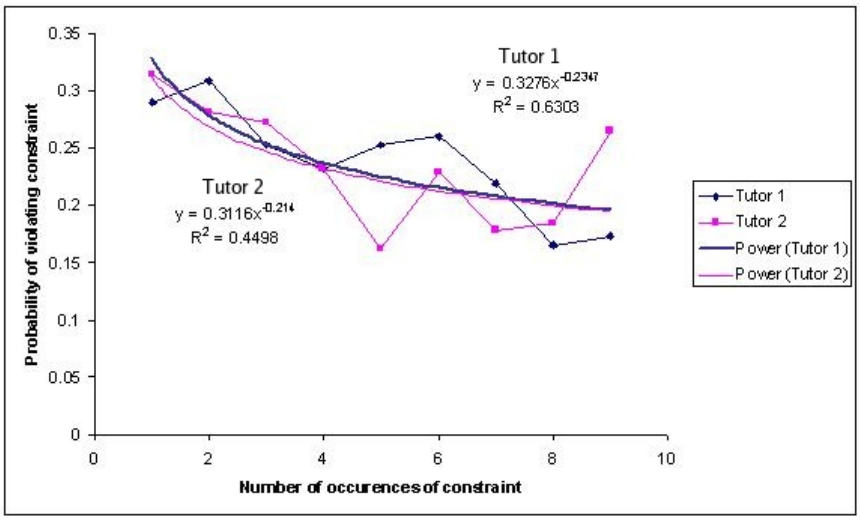

Figure 1. Learning curves for the main task (shared) constraints

From both the learning curves and the pre-/post-tests, there is evidence that the experimental group learned the task of choosing the correct adjective endings better than the control group in the time available. This is despite the fact that the experimental group were slowed by the need to perform the subtask, and so they completed far fewer prob-

lems. This strongly suggests that the additional effort required to perform the subtask was worth it because it allowed the feedback given to better target the current misconception.

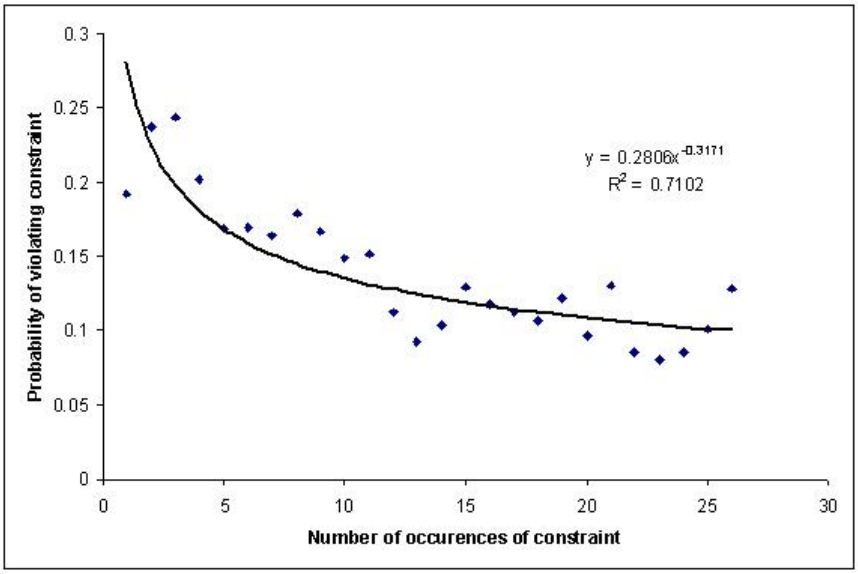

Figure 2. Learning curve for the subtask constraints (experimental tutor)

An alternative explanation is that the subtask itself proved to be useful for learning (and modeling) the main task. Recall that the constraints that were common to both tutors (and the only ones for the control) assumed that the student had made a rule error, i.e. that they knew the gender, case and article, but selected the wrong ending for that situation. (For the experimental group the constraints were subtly different in that they compared the adjective ending to the student's answer for the gender, case and article, i.e. they definitively determined that the problem was a rule error). However, selecting the correct case for an adjective in a sentence requires both that the correct ending be supplied for the situation, and that the situation be correctly interpreted in the first place. The constraints in the control therefore only represents part of the model for this domain, while the model for the experimental tutor is more complete.

Finally, the students were asked to fill in a subjective survey at the end of the study. Responses from were overwhelmingly positive to both versions of the tutor. Comments included "It was good that the mistakes were explained + the grammer rules were also explained." "I liked it and found very useful". Further, the staff from the German department indicated they would like to pursue this technology further, because the students had reacted so positively. They also commented that the results for the formal adjectives test were considerably higher than in previous years, which they attributed to the tutoring systems.

5. Conclusions

Tutoring systems that teach natural languages are susceptible to the problem of ambiguity in student answers, making it difficult to apportion blame appropriately, and thus provide effective feedback. Even a highly constrained domain such as German adjectives exhibits this problem. Requiring the student to supply additional information is often frowned upon because it reduces the correspondence to "real world" problems and may thus negatively affect transfer.

We examined this problem in the domain of German adjectives by providing two versions of a simple ITS; the control required the students to complete the original task only (and thus suffered from ambiguity) while the experimental group forced them to also complete a subtask that disambiguated their response. The results were not conclusive because of problems with the pre- and post-test difficulties. However, there was evidence from these tests that the experimental group performed better on both the original task and the subtask despite having solved considerably fewer problems because of the additional time needed to complete the subtask. This suggests that far from detracting from the students' ability to complete the main task, the extra disambiguation benefited their learning.

Several questions remain unanswered. First, this study was conducted for a highly constrained problem domain; further investigation is needed on more open-ended domains. The German department at the University of Canterbury has indicated that they would like to pursue the technology further, so it is likely we will conduct further studies for other parts of the German curriculum in 2007. Second, the study made several assumptions that require further exploration. In particular, the assumption that students in the control group always make rule errors (i.e. they know the situation but choose the wrong ending) is highly likely to be invalid; if this were the case, we would expect the students in the control group to perform the subtask flawlessly during the pre- and post-tests, which they clearly did not. In fact, the reverse assumption (that mistakes are caused by misinterpreting the situation) has greater supporting evidence since the learning curve for the associated constraints was stronger. One way to test this assumption might be to have the same constraints, but alter the feedback; instead of telling the student how to work out the ending for a particular set of situations, it could indicate the situations for which the ending they have chosen is correct. This warrants further investigation. Finally, the experimental tutor gave feedback for rule errors if the student submitted an ending that contradicted their supplied case, article and gender combination, even if the ending was correct for the problem. In general it is unclear whether or not feedback about the ending should be provided at all if the subtask has not been completed.

This study has shown that adding extra task requirements to overcome ambiguity in language learning is not always a bad thing, and can in fact be advantageous. This is a positive outcome that encourages us to further explore how constraint-based models may support language learning.

References

[1] Menzel, W., Constraint-based modeling and ambiguity. International Journal of Artificial Intelligence in Education, 2006. 16(1): p. 29-63.

[2] Menzel, W. and Schroeder, I., Constraint-based Diagnosis for Intelligent Language Tutoring Systems. Proceedings IT&KNOWS, IFIP World Congress. 1998. p.484-497.
[3] Zakharov, K., Mitrovic, A., and Ohlsson, S. Feedback Micro-engineering in EER-Tutor. in Proceedings of the 12th International Conference on Artificial Intelligence in Education. 2005. Amsterdam: IOS Press.
[4] Anderson, J.R., Corbett, A.T., Koedinger, K.R., and Pelletier, R., Cognitive Tutors: Lessons Learned. Journal of the Learning Sciences, 1995. 4(2): p. 167-207.
[5] Rogers, M., On major types of written error in advanced students of german. International Review of Applied Linguistics in Language Teaching, 1984. 22(1): p. 1-39.
[6] Juozulynas, V., Errors in the compositions of second-year german students: an empirical study for parser-based icali. CALICO Journal, 1994. 12(1): p. 5-17.
[7] Ohlsson, S., Constraint-Based Student Modeling, in Student Modeling: The Key to Individualized Knowledge-Based Instruction, J. Greer and G. McCalla, Editors. 1994, Springer-Verlag: New York. p. 167-189.
[8] Ohlsson, S., Learning from Performance Errors. Psychological Review, 1996. 3(2): p. 241-262.
[9] Newell, A. and Rosenbloom, P.S., Mechanisms of skill acquisition and the law of practice, in Cognitive skills and their acquisition, J.R. Anderson, Editor. 1981, Lawrence Erlbaum Associates: Hillsdale, NJ. p. 1-56.
[10] Martin, B. and Mitrovic, A. Authoring web-based tutoring systems with WETAS. in International conference on computers in education. 2002. Auckland.
[11] Martin, B., Koedinger, K.R., Mitrovic, A., and Mathan, S. On Using Learning Curves to Evaluate ITS. in Proceedings of the 12th International Conference on Artificial Intelligence in Education. 2005. Amsterdam: IOS Press.
[12] Werner, G., Langenscheidts Grammatik-training Deutsch. 2001: Langenscheidt KG.
[13] Kahlen, L., Interactive German Made Easy. 2006: McGraw-Hill.
[14] Terell, T.D., Tschirner, E., and Nikolai, B., Kontakte: a comunicative approach. 2004: McGraw-Hill.
[15] Webster, P., Schwarz, rot, gold: the German handbook: a practical grammar guide. 1987: Cambridge University Press.
[16] Sparks, K. and Vail, V.H., German in review. 1986: Harcourt Brace Jovanovich.
[17] Dreyer, H. and Schmitt, R., A practice grammar of German. 2001: Max Hueber Verlag.