

COMPUTER-BASED ANALYTIC GRADING FOR GERMAN GRAMMAR INSTRUCTION

by

DAVID R. LEVINE

*Intermetrics, Inc.
Cambridge, Massachusetts*

1. AIMS OF THE GRADING-ANALYSIS PROGRAM

GRADING, or response evaluation, forms a central part of any instructional system. Many computer-assisted instruction (CAI) projects have only limited evaluation capabilities, usually handcrafted for each question. As a result, these programs often could not deal successfully with multiple-word, structured responses, such as complete sentences; proper evaluation was assured by restricting the curriculum to a short-answer model, which not only resulted in a distorted pedagogy, but also allowed only a narrow channel for information about the student.

The project reported here represents an investigation by the author in 1971-1972 into a more powerful approach to grading. It provides a system capable of dealing with large, structured responses, in which the order of elements may vary widely. The general area of concern is language teaching, particularly German grammar instruction, and the goal is the analysis of responses consisting of complete sentences. Grading is based on built-in grammatical knowledge and general algorithms, augmented by a structural description of the expected response. The resulting analysis is thorough and comprehensive; it verifies adherence to rules, as well as checking for the production of expected orthographic forms. In addition, the analyzer attempts to infer the mislearned rules underlying observed wrong responses.

This project provides a working demonstration of sophisticated grading analysis for a rich, though restricted, domain. It operates upon natural language, but purely with reference to grammatical issues. Variations in meaning or vocabulary are outside its scope.¹ The program as reported

This article originally appeared as Technical Report No. 199 (Psychology and Education Series), 1973 Institute for Mathematical Studies in the Social Sciences, Stanford University. The research reported in this article was partially supported by National Science Foundation Grant for Basic Research in Computer-assisted Instruction, NSF GJ-443 to Stanford University.

¹ This restriction need not have an adverse pedagogical impact. Traditional classroom pattern-drill practice, for instance, specifically discourages such variation.

here does not contain logic for dealing with spelling errors; the concern has been first to achieve a comprehensive grammatical analysis.

2. OPERATION OF THE PROGRAM

2.1 *Rationale of the Program*

The first step in analysis consists of recognizing the words and identifying the structure of the sentence. As a general problem in natural-language processing, the general analysis of an arbitrary sentence poses a very difficult task. Moreover, in an instructional setting, one must allow for the presence of errors that will distort or obliterate information needed for a general parse. Still, the use of general subject-matter information seems essential if one is to achieve comprehensive analysis of complex responses.

A compromise position has been taken. The instructional paradigm includes the notion of an expected response. The grading program is given two inputs, a description of the expected response and the sentence actually constructed by the student. Using its built-in knowledge of German grammar, the program analyzes the student sentence for conformity with both the expected response and the applicable rules of German grammar and produces a comprehensive diagnostic report. For this project, it is assumed that the student will actually produce something quite close to the requested sentence. There may be errors, and substructures out of order, but the student must have used essentially the expected vocabulary and be attempting to express the expected meaning. This assumption, incidentally, is not pedagogically unreasonable: A student is not told to "say something in German" but, rather, is given a specific stimulus, often to produce a variation on a given pattern. The program therefore receives assistance in the form of a fairly complete description of the expected response. Both the constituent vocabulary and structure are given.

2.2 *Example of Grading Analysis*

A hypothetical example will demonstrate how the grading program operates. The expected sentence is *Jetzt will er dem Maedchen die Tuer aufmachen* (Now he will open the door for the girl). Explanatory notes in the example are given in square brackets.

Description generated by human expert:

(STMT (VERB WOLLEN (AUF . MACHEN))
 (SUBJECT ER)
 (OBJECT ACC DIE TUER)
 (OBJECT DAT DEM MAEDCHEN)
 (PREDMOD (ADVERB JETZT)))

[The program receives a description of the expected response, containing both the vocabulary and grammatical structure. This response will be a statement with a pronoun subject (*er*) and a double verb; there are two objects—one accusative, one dative—and a predicate modifier. The computer knows that *er* is a pronoun and that *die Tuer* is a feminine singular noun phrase without having to be told specifically.]

Sentence input by student:

JETZT WILL ICH MACHN DIE TUER AUF DER MAEDCHEN

[Although there is a valiant attempt at the sentence, this response has quite a few errors, including a change to first-person subject (*ich = I*).]

Analysis developed by the program:

NOW PROCEEDING WITH ANALYSIS
VERB MACHEN . . APPEARS AS MACHN
MISSING -E ON INFINITIVE ENDING

[The stem is *mach-*, and there is no interpretation under which an *-n* ending is correct. Since the verb is expected to occur as an infinitive, the program complains about a specific malformation.]

SUMMARY REPORT FOR NOUN PHRASE INFLECTION ON DIE TUER
BEST WE CAN DO GUESSING INTENT IS: ((NOM ACC) S F)
CONFIDENCE LEVEL: VERYHIGH

[The first unit of the sentence to be completely identified is the noun phrase *die Tuer*. The inflectional analysis has tried to determine, from an examination of the response, what grammatical interpretation to assign. In this example, the pair reflects an inflection that would be correct for either nominative or accusative case, singular, and feminine. "Confidence level" is an internal performance assessment about the guess, not a comment about the state of the student's knowledge.]

GENDER REPORT FOR NOUN: TUER ::
STRONG CORRECT = F

[A specific report for the noun, based on the hypothesis arrived at in the previous step.]

CASE REPORT FOR OBJECT: DIE TUER :: WEAK CORRECT
EXPECTED ACC , COULD BE (NOM ACC)

[For this noun phrase, the inflection for gender is clearly for feminine. The response does not really indicate whether the student intended nominative or accusative case, as the form *die* is the same for both.]

SUMMARY REPORT OF NOUN PHRASE INFLECTION ON DER
MAEDCHEN

BEST WE CAN DO GUESSING INTENT IS: (DAT S F)
CONFIDENCE LEVEL: HIGH

[The computer's informed guess, as before.]

GENDER REPORT FOR NOUN: MAEDCHEN :: INCORRECT
SHOULD BE N IS F

[Based on the above hypothesis. Note that the gender error is attributed to the noun, even though only the article is inflected.]

CASE REPORT FOR OBJECT: DER MAEDCHEN ::
STRONG CORRECT = DAT

[Notice that the program has again used its hypothesis and has thus been able to establish correct performance for case.]

PRONOUN SUBSTITUTION .. SHOULD BE (3 NOM S M PER)
IS ((WORD ICH (I C H)) : (1 NOM S * PER))

[The program has successfully identified the subject pronoun, even though the error has resulted in a totally different form than that expected. The message has given the internal representation of the inflection characteristics for the two forms, from which one can see that they differ only in person.]

CASE REPORT FOR SUBJECT: ICH ::
STRONG CORRECT = NOM

VERB ENDING CHECK .. OK!

[Even though the subject was not the expected one, the program can check agreement between the actual form and the actual verb and also credit the student for proper (nominative) case of the pronoun.]

PHRASE DIVISION DONE, SENTENCE TYPE REPORT:
NO CONFIDENCE IN SENTENCE TYPE, PROBABLY CONFUSED
LAST STRAW WAS: SEPARABLE-PREFIX CONFUSION

[This sentence is pretty badly mangled, and the word-order check finally gives up trying to match it to a German pattern.]

OTHER COMMENTS:

DI POSITION ERROR, NOT SEPARATED FROM FV
FRONT FIELD ERROR (FV-2): 2 ELEMENTS IN FRONT FIELD

[Specific rules that were violated include those governing the placement of the Dependent Infinitive, and the front field in a Finite Verb 2nd clause.]

END OF RUN

The grading-analysis program is divided into two phases, the matching algorithm and the diagnosis of the student's response. We discuss these in the next two sections.

3. PHASE ONE: THE MATCHING ALGORITHM

3.1 *Structural Matching*

In the first phase, the expected response provides basic guidance to the matching algorithm where the student response is measured against the

expected response word by word. The program works through the sentence looking for the expected major (lexical) words, temporarily ignoring other information like function words (articles, prepositions, etc.). The major words provide fairly reliable match keys for several reasons:

1. Lexical words (nouns, verbs) are less highly inflected than many function words and are thus easier to recognize reliably.
2. Such words form a natural core of the structural subunits of the language.
3. As meaning-carrying words, they are unlikely to be omitted.

Each instance of a successful match provides a foothold from which the program can work to apply its general analysis rules. By referring back to the descriptive information, the program determines what structure surrounds the word it has just found. If this structure includes function words, such as articles or prepositions, the program looks for them in the immediate neighborhood of the key word. The word at the expected location is scrutinized to determine that it does indeed belong to the expected class of words (e.g., article, preposition), and if so, the match is made.

Implicit in this strategy is a second structural assumption, that the student will keep the clause elements of the sentence internally intact. In other words, the assumption is that the constituent words will appear in the proper order; the article in a noun phrase, for example, must appear (if at all) immediately before the noun.²

One further restriction was included in the working system: The sentence must not contain two instances of any lexical word. Thus, sentences like *The red car hit the blue car* cannot currently be handled. With this restriction, the analyzer can move directly from an identified lexical word to the correct element of the structure description, without the need to examine additional context.

Nonmajor words may appear more than once in a sentence, presenting a potential problem of ambiguity. The structurally guided search for function words ensures that each is matched and identified with the proper structure, regardless of inflection. Special handling is required, however, for free-standing nonmajor words.

Many German verbs have separable prefixes. If a prefix has been separated, then a separate search must be performed to locate it. Most separable prefixes are indistinguishable from prepositions, however, so this search must be delayed until all primary recognition has taken place, in order to minimize the chances of improperly identifying another preposition as the prefix.

Pronouns represent an extra degree of complexity. There may be several pronouns in a sentence, and unlike function words, which are physically part of a larger structure, pronouns stand essentially alone. For

² Such errors would violate the patterns of the student's native language as well as those of German and are thus not commonly observed in language teaching.

proper analysis, it is essential that the student's pronouns be matched with the proper ones in the expected response. It is from this match that the program identifies sentence structure, so an incorrect match will lead to meaningless diagnosis.

Accordingly, the pronoun search is first delayed until the other sentence elements have been identified. In particular, this avoids confusion with definite articles (whose forms overlap those of the demonstrative pronouns), as the articles will have already been claimed. If the sentence contains more than one pronoun, inflectional clues are used to resolve the ensuing ambiguity. The word that is "closest" to the expected response is chosen by the identification logic. Closeness is measured by the number and kind of grammatical categories in which the response differs from the expected value. In matching to a subject pronoun, the agreement between subject and verb is also considered. In particular, if the verb inflection does not agree with the expected value, then the criteria for the subject pronoun are relaxed to ensure matching either the expected value or the form that agrees with the verb as written.

3.2 *Word Matching*

Inflectional changes, both correct and incorrect, constitute the major complication in identifying the words in the student sentence. Happily, most major words inflect, if at all, by suppletion, that is, by adding an ending to an invariant stem. Such words can be located by merely ignoring the ending, though saving it for later checking. For those words that show inflectional changes in the stem as well, the change is usually fairly small and more or less predictable. The few pathological words must just be handled as special cases.

The recognition algorithm uses a constrained pattern match. To find the word "MANN" (*der mann, die maenner: man, men*), the program will generate a pattern that matches $M + \langle \text{one or more vowels} \rangle + NN + \langle \text{any ending} \rangle$. This search pattern will always be able to accommodate a stem-vowel change, even if that would be incorrect for the word in question. Similarly, the program will recognize both singular and plural forms, regardless of the expected value. This approach has proved successful for both nouns and verbs.

Articles are very highly inflected words with a plethora of orthographic forms. Furthermore, the student may well have used a different type of article, for example, indefinite instead of definite. A composite algorithm examines the candidate word, determines whether it represents any sort of article, and records the actual type and inflection for later checking. A word is identified as a definite article if it corresponds exactly with any one of the actual forms of the definite article. The other types of articles have sufficiently long stem portions to permit recognition based on a stem plus arbitrary ending pattern match.

Prepositions show no inflection. The student, however, may have substituted a different preposition. Recognition involves merely checking the word against a complete list of prepositions. A slight bit of extra logic takes care of preposition-article contractions.

3.3 *Miscellaneous Issues*

The preceding discussion has addressed the matching of the student response with the expected response, on a word-by-word basis. There may, however, be some words missing from the response, or there may be a substitution or an addition. All of these lead into difficult areas of natural language. Without vastly stronger semantic tools, the program can do little beyond simply reporting, at a surface level, absence of expected words or the presence of unexpected ones.

Some special cases can be handled. The loss of a function word has far less impact than does the absence of a key word. If a noun is not found, and thus the entire noun phrase is apparently missing, the program will attempt to find, instead, the equivalent pronoun. When successful, this stratagem allows further checking to take place, for example, subject-verb agreement. With a small dictionary, the program might also be able to handle reasonable substitutions of one lexical word for another, at least at the rudimentary recognition level needed to complete the grammatical analysis.

4. PHASE TWO: DIAGNOSIS OF STUDENT RESPONSE

Once the words in the student response have been matched to the words and structure of the expected response, the program proceeds to the diagnosis phase. This phase has the responsibility of checking adherence to the various rules of grammar. The diagnostic analysis makes use of general information about applicable rules of German grammar, augmented with general strategies for coping with errors. The project's goal was to accomplish more than a simple determination of right or wrong; among the many facets of the analysis are the disentanglement of multiple errors, and the extraction of information pertaining to the degree of the student's competence with the specific rules of the subject.

The diagnostic strategy will be discussed first in reference to single-word inflection checking; then multiple-word patterns and word-order rules will be discussed.

4.1 *Single-word Inflection Checking*

The inflection of a single word is often governed by several different rules. The program attempts to report on each one specifically, especially in situations in which only one of several applicable rules has been incorrectly applied. Accordingly, the program first determines which rule-

governed grammatical categories apply to the current word. For each category (e.g., case, number, gender), the correct value is determined, yielding a value-set (e.g., NOM SING NEUT).

The next step is to examine the actual response. Working in reverse through the grammar tables, the program determines the value-set for which the student form would be correct (e.g., "DAS": NOM SING NEUT, also ACC SING NEUT). A response is judged correct if the intersection of the expected and actual value sets is nonnull.

For an incorrect response, a detailed examination of the value-sets yields information about which categories are in error, and those for which the inflection is correct. For each of the latter, the student can be given proper credit. Errors are traced back to the governing rule for that category, thus allowing formation of a hypothesis about where the student's knowledge is deficient. An error in an article, for example, may well be due to confusion about the gender of the governing noun, not to any insufficiency in article inflection skills.

The analysis is rendered more complex by the presence in German grammar of overlapping forms. This means that a response can be "correct" even if the student's intent was actually wrong. There is no way to tell whether a use of *das*, for example, represents an instance of nominative or accusative case. For an incorrect response, the program has to make a choice between the several possible value-sets that characterize the observed form; only then can it proceed with a coherent hypothesis.

Several algorithms have been investigated for choosing among competing explanations of a wrong response. One method involves a "distance" metric. The value-set that characterizes the expected response is compared to the set for each possible explanation of the actual response. The value-set (possible explanation) with the fewer deviations (or the less distant deviations) from the expected values is rated highest. Thus, for example, *der* is the correct form for both NOM SING MASC and DAT SING FEM. If the expected response is NOM SING NEUT, the first explanation deviates only in gender (MASC becomes NEUT), whereas the second would entail changes in both case and gender. The diagnosis will then report a gender error. On the other hand, for a sentence requiring ACC SING FEM, the DAT SING FEM form is closer, with a single error of incorrect case.

Another approach to the choice problem involves a prediction matrix. For each combination of actual form and expected values, the matrix gives a "best" explanation. Thus, *der* occurring in a NOM SING FEM context would be directly linked to a NOM SING MASC explanation. Entirely hand-crafted, the matrix works efficiently but makes no use of the program's general grammar knowledge. A third scheme attempts to combine the strong points of the other two, by incorporating transition probabilities. Through repeated observation, it should be possible to determine the probability that the student will actually produce an accusative form when one is required (and just how likely that form is to be nominative). These

probabilities provide the necessary decision power to choose between the conflicting alternative explanations of a wrong response.

4.2 *Multiple-word Patterns*

Multiword inflections require some additional subtlety, while also presenting additional diagnostic opportunities. Good examples are afforded by the logic for checking subject-verb agreement, and for checking agreement in number between article and noun. The program looks first at the actual inflection of both elements and, as for single-word checking, obtains the value-sets for which these forms are correct. The intersection of the two sets gives direct information about the student's observance of the rules governing agreement. In addition, the actual forms are checked against the expected values, though an error here is probably less serious provided that there is internal agreement. Partial correctness is diagnosed, as for single words.

4.3 *Word-order Checking*

Word-order checking constitutes a third major diagnostic task. A structural approach is used, centering on the notion of "clause element": subject, object, verb, etc. Word-order rules apply both to the order of words within a clause element and to the relative position of the elements within the sentence. The internal ordering does not need checking at this point; as discussed above, the internal order is presumed to be correct and is used to guide the matching routines. External order among the elements is regarded as completely unconstrained during matching; after all clause elements have been identified, a global view of their relative ordering becomes possible.

The word-order analysis must first determine which type of sentence has been produced; then the word-order rules applicable to the type can be checked. Clause type is primarily related to the position of the finite (inflected) verb relative to other clause elements. German allows three possibilities: finite verb in first, second, or last position. The diagnosis also considers other verb characteristics: the position of a dependent infinitive, if present; the state and position of a separable prefix; and whether an introductory relative pronoun is present as expected for a relative clause. Some consistency checking is done during the type determination. Provided there are not too many distortions, the classification is made and the program continues with detailed checking. Otherwise, the program concludes that the student is merely confused.

The actual clause type is compared with the expected type for correctness. The major part of the diagnosis, however, involves internal consistency checks. As always, individual rules are reported separately to localize errors. Concrete rules governing the position of verb, prefixes, and pronoun subjects are verified. Other ordering, such as which element appears before the verb in a statement form, provides semantic nuance only:

Almost any permutation is grammatically legal. Discrimination requires language subtlety outside the range of the current project. The flexible positioning can be checked only by reference to the order given in the description of the expected sentence, with no rule-based explanation offered for errors.

5. CONCLUSION

This project originated from a desire to fashion the computer into a tool for performing sophisticated analytic grading of student responses. In the restricted area of pattern practice and controlled translation, in which the general form and content of the sentence are narrowly constrained, I have produced a computer program that can do an in-depth grading analysis of entire German sentences. Future development will include an expansion of the program's capabilities to additional areas of language. Concurrently, the grader should be embedded in a sophisticated teaching system capable of handling the rich flow of diagnostic information.