ABSTRACT

An analysis of errors in an approximately 400 page corpus of German essays by American college students in second-year language courses shows that syntax is the most problematic area, followed by morphology. These two categories together make up 53% of errors. Errors are divided into five categories and two types. The study, a contribution to the error analysis element of a syntactic parser of German, indicates that most student errors (80%) are not of semantic origin, and therefore, are potentially recognizable by a syntactic parser.

KEYWORDS

Writing aids, courseware development, parsing, ICALI, error analysis, error typology.

INTRODUCTION

Empirical study of “learner’s language” is needed in order to provide data to designers of ICALI programs, who can then orient their error recognition and analysis efforts toward the most frequent and significant of student errors. This study was undertaken in order to provide such data to ICALI designers, specifically the designers of Syncheck, a comprehensive syntactic parser of German now being implemented as a writing aid for students of German (Sanders and Sanders 1989). Analysis of the “Miami corpus” of 360 pages of German essays written by American college students in second-year language courses shows that approximately 80% of the errors are in principle recognizable by a purely syntactic parser.¹
THE CORPUS AND THE TOOLS

The body of text used in this project is made up of student writings collected during the years 1988-1990. The essays were written by Miami University students of second-year German. In all, 349 student compositions were collected. These comprise the 400 page “Miami corpus,” of which 360 pages (313 essays) were included in the research corpus for computerized analysis (see Table 1). The compositions were written on topics such as “My Last Trip,” “Visiting a Museum and Describing a Picture,” “Description of an Advertisement from German Periodicals,” “My Apartment/Room,” “Detective Story,” and short stories assigned as course reading.

The essays are usually one page long, occasionally longer - two or three pages. All of the headings, titles, and footnotes, such as bibliographical information, sources, etc., were deleted after the scanning procedure. The compositions were anonymous; they were numbered for identification.

PROCEDURE

The student compositions were entered into a computer database with an optical scanner. For the input, the scanner Microtek MS-SCSI/G and scanning program OmniPage™ 2.12 for a Macintosh IIcx computer were used. Despite the fact that optical scanning is much faster than manual entry, it is still rather slow — it takes about three minutes to scan one page of good-quality text, and about twenty minutes in total to correct the unrecognized or ill-recognized characters in one medium-length essay. The computer indicates the unrecognized characters with a tilde on the screen. These included handwritten umlauts, capitalized “B” instead of “ß,” etc. Some of the essays (about 7%) were handwritten, typed or word processed in italics or script font, hand corrected by the instructor, or in typewriter print that was too faint. These problems made the text partly or totally unscannable. About 10% of the pages were totally unscannable, and in the eventual database analysis these were rejected. Scanning is probably the only method of obtaining the computerized data. Manual input, i.e., retyping of these compositions, would be time consuming or even, as a practical matter, impossible because the typist would have to reproduce, with a high degree of accuracy, not German or English but a phenomenon in between — a kind of metalanguage. The second row in Table 1 indicates the actual numbers of the research corpus.
The scanned essays were saved on micro-discs in ASCII format (Macintosh) and later transferred to an IBM format for further processing.

**METHODS OF ERROR ANALYSIS**

Hendrickson (1979) proposed a taxonomic schema that provides a visually clear and unusually straightforward method of categorizing errors. Of the available taxonomies, it seems the most suitable for syntactic parsing, in spite of the fact that this schema is intended for use in the evaluation of students' oral production of language. His schema, slightly modified, provides the basis for the categorization of syntactic errors used here. The category “Lexicon” was changed to “Semantics,” “Morphology” and “Syntax” left the same, and the category “Orthography” was divided into two independent categories: “Spelling” and “Punctuation,” the latter reflecting the importance of the comma for German syntax.

With these changes, then, the schema looks like this:

<table>
<thead>
<tr>
<th>CATEGORY TYPE</th>
<th>SYNTAX</th>
<th>MORPHOLOGY</th>
<th>SPELLING</th>
<th>PUNCTUATION</th>
<th>SEMANTICS</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOCAL ERRORS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Specification of the Problem Area</td>
<td></td>
</tr>
<tr>
<td>GLOBAL ERRORS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 1. Research Corpus**

<table>
<thead>
<tr>
<th>Number of ESSAYS</th>
<th>Scanned</th>
<th>Impossible to Scan</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>313</td>
<td>26</td>
<td>349</td>
<td></td>
</tr>
<tr>
<td>Number of Single Page Essays</td>
<td>277</td>
<td>33</td>
<td>300</td>
</tr>
<tr>
<td>Number of Double Page Essays</td>
<td>35</td>
<td>2</td>
<td>37</td>
</tr>
<tr>
<td>Number of Triple Page Essays</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Number of Pages</td>
<td>360</td>
<td>40</td>
<td>400</td>
</tr>
</tbody>
</table>

**Table 2. Error Taxonomy**
The types of errors, local and global, are represented on the vertical scale. **Local errors** are those which affect only the constituents in which they appear. Categorization of problem areas is straightforward.

**Global errors** affect the structure of the entire sentence, such as a missing essential part of the sentence — subject or verb. Global errors are difficult or even impossible to categorize. Krishna (1975, 45) calls such errors “structural” and also rejects the possibility of their categorization. For example, the erroneous verb form in the construction *er trugt* may have at least three different causes: wrong **tense** form (Simple Past Tense Singular 1st or 3rd person *trug*), wrong **person** form (Present Tense Singular 3rd person *trägt*), wrong **mood** form (Present **Subjunctive** Plural 2nd person *trüget*), or all of them together.

In the text of each essay particular error codes were inserted according to the category, type, and problem area of each error. Thirty-eight kinds of errors were distinguished.

**DATA ANALYSIS**

Using the text retrieval software **WordCruncher® Index™**, the number of sentences, words, and errors in each category and type were counted. The research corpus was reorganized by sorting the text pieces containing a particular error with **WordCruncher® View™**. For example, all morphological local errors with the problem area "Noun Gender" are grouped together in one file.

**GENERAL SURVEY**

In the 360 pages analyzed, a total of 2199 errors in syntax, 1881 errors in inflectional morphology, 1537 errors in semantics, 1130 errors in spelling, and 954 errors in punctuation were found. Figure 1 shows the relevant percentages. Overall, excluding semantic and punctuation errors, 17.2% of the errors were global (i.e., affected the entire sentence), while 82.8% were local (i.e., affected only one word). In general, global errors may not be addressable by computer-based writing aids: they may cause the sentence to be rejected as ill-formed, but the complexity of the error will probably be beyond analysis by the program.
SEMANTIC ERRORS

This category contains errors of meaning, such as wrong word choice, "made-up" words, and errors in pronoun reference, which in principle cannot be recognized by a syntactic parser. Examples of this type of error, all drawn directly from essays in the research corpus, are given below, along with the correct word or form immediately following each example:

- *Es* [the room] *hat noch einen Spiegel, einen Vorhang, ein Regel, Pflanze, und zweie Lampen.*

  (...ein Regal...)

- *Mein Fernsehen ist klein, aber es ist nett.*

  (...Fernseher...)

- *Auch, es gibt einen Rauchendetektiv.*

  (...Rauchdetektor...)

Sometimes it is possible to guess the intention of the author and to substitute the correct word, but often the semantic structures of the words used in the sentence are totally incompatible with each other, and it is hardly possible to interpret the author’s intention, as in the sentence:

- *Was ist das Schlafzimmers Absicht?*
Logical errors or errors caused by misinformation are not considered semantic errors because these errors are not of linguistic origin. For example:

- *Salzburg ist die Hauptstadt des Österreich.*
- *Obwohl ich nie Salzburg gesehen habe, besuche ich Salzburg gern.*

LOCAL ERRORS

Local errors are those which influence just part of a sentence or clause, such as errors in noun or verb phrases. Usually, it is not difficult to identify the problem area of local mistakes. In sum, 34 kinds of local errors were distinguished: eleven in the morphological category, eleven in syntax, five in spelling, and seven in punctuation. Figure 2 shows the frequency of these errors.

![Figure 2. Frequency of local errors](image)
SYNTACTIC ERRORS

In this category are eleven types of the errors of semantic origin:
1. Preposition + Required Case
   
   • *Es hat ein Bodezimmer mit eine Toilette.*
     (*mit einer Toilette.*)

2. Verb + (Required Preposition) + Required Case
   
   • *Wir haben ein Tepichboden mit gelb, braun und orange Farben.*
     (*einen Teppichboden...*)

3. Word Order in Main Clause
   
   • *Jeden Tag durch das Radio spreche ich mit meinem Vater.*
     (*Jeden Tag spreche ich...*)

The other eight kinds of errors in syntax are as follows:
4. Subject - Verb Agreement

5. Word Order in Dependent Clause

6. Missing Parts of Sentence/Clause (Subject, Verb, or Part of Verb, Object, etc.)

7. Relative Pronoun

8. Reflexive Verb

9. Word Order in Dependent Clause with Additional Infinitive

10. Infinitive Particle *zu*

11. Adverb or Adjective with Required Case and Preposition
MORPHOLOGICAL ERRORS

In this category are errors in the inflectional morphology system of German.

1. Noun Gender

This is the largest group of morphological errors. Erroneous gender identification also causes a wrong inflection in the noun or in the entire noun phrase:

- Während der Wochenende, gehe ich nach Heuston Woods für drei Tagen.
  (Am Wochenende [Neuter]...)

2. Case Endings of Nouns

The most common errors in noun case endings are oblique case ending in the singular, (e)n, and (e)n in all cases in the plural required by a special group of masculine nouns, (e)s in the singular genitive of masculine and neuter nouns, and the dative plural ending n for the majority of nouns.

- Die Idiomen ist gut in die Kapitels.
  (...den Kapiteln.)

- Sie gibt die Miniatur Herr Klotz.
  (...Herrn Klotz.)

- Die Mitte des Gemälde ist der Sonnenuntergang.
  (Die Mitte des Gemäldes ...)

Sometimes the case endings are added to nouns of the wrong category, as in this example where s is added to a feminine noun:

- Das war meiner Grossmutters.
  (...meiner Großmutter.)

3. Number Form of Noun

- Das Zimmer hat zwei Schreibtischen und zwei Stühle - das ist sehr gewöhnlich von die Zimmern von der Universität.
  (...und zwei Stühle)
The remaining eight types in this category are as follows:

4. Ending of Adjective/Ordinal Number in Group Adjective/Ordinal Number + Noun

5. Definite Article/viele or Indefinite Article/viel with Noun

6. Ending of Adjective/Ordinal Number in Group Ein-word + Adjective/Ordinal Number + Noun

7. Ending of Adjective/Ordinal Number in Group Der-word + Adjective/Ordinal Number + Noun

8. Verb Form

9. Indeclinable Adjective or Number Form

10. Adjective/Adverb Form

11. Auxiliary Verb in Tense Formation (haben/sein)

PUNCTUATION ERRORS

In German, punctuation is more likely than in English to be an element of grammar and particularly the comma is likely to be a problem for English speakers who are learning German. The most common kinds of errors are as follows:

1. Missing Comma in Structure Main Clause + Main Clause
   • Meines Bett ist gross und ich schlafe und sitze hier gern.
   (Mein Belt ist..., und ich schlaf...)

2. Missing Comma(s) in Structure Main Clause + Dep. Clause or Dep Clause + Dep Clause
   • Kleine Städte haben auch viel anzubieten, wie ich erfahren habe als ich über Tübingen gelassen haben.
   (... , wie ich erfahren habe, als ich...)
3. Presence of Erroneous Comma
- Ich habe den Braunen Teppich, den Kühlschrank, die Stereoumlage mit dem Plättenspieler, und die Truhe mit meinen Kleiden.
- Von hinter dem Auto man kann ein Auto von dem gehobenen Klasse sehen.

4. Missing Comma(s) in Infinitive Group
- Ich ging ins Gasthaus zuruck um mich schlafen zu legen.
  ( ... zurück, um mich ... )

The remaining three kinds of errors in the category “Punctuation” are:

5. Missing Question Mark

6. Missing Exclamation Mark

7. Missing Quotation Marks

**SPELLING ERRORS**

Five types of spelling errors make up this category, which together account for 14.7% of the errors:

1. The Missing Umlaut
   - *Nachstes* Semester bekomme ich das andere Schlafzimmer.
     (Nächstes...)

2. Minisculization Needed
   - Es interessiert mich, wie Deutschen Schreiberin seinen Geschickten erzählen.
     (... wie deutsche Schriftsteller...)

3. Capitalization Needed
   - Ich will diese zimmer beschreiben.
     (... Zimmer...)

4. Indiscrimination Between ss/s/ß
   - Er wusste das das licht immer sie scheinen.
     (Ich wußte, daß das Licht...)

5. Erroneous Umlaut
   - Es tut mir leid, aber das ist die Währheit.
     (... die Währheit.)
GLOBAL ERRORS

The global errors may be in morphology, syntax, or spelling. Figure 3 shows the numbers and kinds of global errors found in the corpus. Sometimes it is impossible to identify what is incorrect in a sentence with a global error, as for instance in this example:

- *Vieles Paul Chidlaws Gemälden vorschlagt die Landschaft und das frisches Farben ist sehr impressionistisch.*

Global spelling errors include, i.e. typos, letter transposition, misspelling, separation of compound words into parts, etc., which affect the sentence as a whole:

- *Den Schreibtische gegenüber ist zwei Betten über einander.*
  (... zwei ... übereinander.)

*Figure 3. Frequency of global errors*
CONCLUSION

Analysis of the Miami corpus shows that 80% of errors in the essays of the second-year students of German can in principle be recognized by a syntactic parsing program; the remaining 20% are semantic. The biggest problem in the students’ writing seems to be syntax, especially the use of the verb with the required noun case or the use of the required case with the noun object of a preposition. Inflectional morphology with its much-feared “endings” takes second place. Syntax and morphology together make up 53% of the errors in the corpus.

Noun gender and verb forms are the most frequent errors in students’ writing within the category of morphology (5.7% and 4.3% of all errors respectively). Errors in some other categories, such as the adjective endings or use of cardinal numbers, are often caused by incorrect gender assignment. The elimination of the above mentioned errors in gender and related categories would lead to a reduction in the total number of errors by twelve percent, at least.

In contrast to at least one other study (Rogers 1984), semantic errors made up a small portion (20%) of the total. After adjusting Roger’s classification (op. cit., 27) by assigning her categories “lexical errors” and “complete transfer of English expression” as well as some types of syntactic and morphological errors (e.g., pronoun reference, word formation, etc.) to the semantic category used in my classification, it becomes obvious that at least about 30% of errors in Rogers’ study are of semantic origin. This may be explained by the fact that the Miami University student samples were from second-year students, while the students in Rogers’ study were advanced, with at least four years of learning German in a formal environment, in many cases supplemented by visits to Germany (ibid., 1).

NOTES

1 See also Sanders 1991.
2 Here and in further examples only one error in the phrase or sentence/ clause is highlighted and explained, even when other errors are present.
3 In some cases, this group of errors can be considered as a subcategory of the group “Missing Parts of Sentence” (missing object).
4 The compete documentation of the errors and examples from the corpus are found in Juzulynas (1991), as well as computer listings of the entire corpus by error type.
Without going into the depths of historical origins and types of the German Umlaut, it should be mentioned that two types of Umlaut are distinguished here: the morphological, or formational, Umlaut includes, for example, plural forms like Stuhl-Stühle, or adjective/adverb degree forms like jung-jünger, while the spelling Umlaut, e.g., Universität-Universitäten, ähnlich-ähnlicher, is simply part of the German spelling convention.

REFERENCES


AUTHOR'S BIODATA

Vilius Juozulynas completed his undergraduate education at Vilnius University, Lithuania, in 1988, and received an M.A. in German at Miami University, Ohio, in 1991. This article is based in part on his master's thesis, “An Error Analysis of the Compositions of Second-Year German Students: A Computer-Assisted Study.” He is currently a graduate student in the Ph.D. Program in Germanic Linguistics at Indiana University, Bloomington.

AUTHOR'S ADDRESS

Department of Germanic Studies
Ballantine Hall 647
Indiana University/Bloomington
Bloomington, IN 47405

Phone: (812) 323-1931 (home)
(812) 855-7947 (office)
E-mail: vjuozuly@ucs.indiana.edu