Extended Spelling Correction for German

Ralf Kese, Friedrich Dudda, Gerhard Heyer, Marianne Kugler
TA Triumph-Adler AG
TA Research EF
Fürther Str. 212
D-8500 Nuremberg 80
e-mail: ralf@triumph-adler.de

Abstract

A conservative extension of standard spelling correction systems for German is discussed which goes beyond normal checking of isolated single words by taking multi-words, linguistically motivated non-words, as well as contexts into account.

1 Motivation

As indicated by Maurice Gross in his COLING 86 lecture (Gross, 1986), European languages contain thousands of what he calls "frozen" or "compound words". In contrast to "free forms", frozen words - though being separable into several words and suffixes - lack syntactic and/or semantic compositionality. This "lack of compositionality is apparent from lexical restrictions" (at night, but: *at day, *at evening, etc.) as well as "by the impossibility of inserting material that is a priori plausible" (*at {coming, present, cold, dark} night) (Gross, 1986).

Since the degree of frozenness can vary, the procedure for recognizing compound words within a text can be more or less complicated. Yet, at least for the entirely and nearly entirely frozen forms, simple string matching operations will do (Gross, 1986).

However, although this clearly indicates that at least those compound words whose parts have a high degree of frozenness are accessible to the methods of standard spelling correction systems, the problem is that these systems at best try to cope with (some) compound nouns (Frisch and Zamora, 1988) while they are still ignorant of the bulk of other compound forms and of violations of lexical and/or cooccurrence (Harris, 1970) restrictions in general.

As Zimmermann points out (Zimmermann, 1987) with respect to German forms like "in bezug auf" (= frozen) versus "mit Bezug auf" (= free), compounds clearly are out of the scope of standard spelling correction systems due to the fact that these systems check for isolated words only and disregard the respective contexts.

Following Gross and Zimmermann (Gross, 1986; Zimmermann, 1987), we propose to further extend standard spelling correction systems onto the level of compound words by making them context-sensitive as well as capable of treating more than a single word at a time.

Yet even on the level of single words many more errors could be detected by a spelling corrector if it possessed at least some rudimentary linguistic knowledge. In the case of a word that takes irregular forms (like the German verb "laufen" or the English noun "mouse", for example), a standard system seems to "know" the word and its forms for it is able to verify them, e.g., by simple lexicon lookup. Yet when confronted with a regular though false form of the very same word (e.g. with "laufte" as the 1st/3rd pers. sg. simple past ind. act., or with the plural "mouses"), such a system normally fails to propose the corresponding irregular form ("lief" or "mice") as a correction alternative.

Following a hint in (Zimmermann, 1987), we propose to enhance standard spelling correction systems on the level of isolated words by introducing an additional sort of lexicon entries that explicitly records those cognitive errors that are intuitively likely to occur (at least in the writings of non-native speakers) but which a

standard system fails to treat in an adequate way for system intrinsic reasons.

Considering that most of these errors give a clear indication of a writer's knowledge of the orthography of a language, and that therefore their correction may be of particular importance to him, the system also is explicitly intended to exemplify how more complex linguistic phenomena that are of importance to a user may yet be treated by simpler means, thus achieving in the sense of an engineering approach (Heyer, 1990) a satisfactory trade-off between theoretical costs and practical benefits.

2 Overview of new Phenomena for Spelling Correction

As there are irregular forms which nevertheless are well-formed, i.e.: words, there are also regular forms which are ill-formed, i.e.: non-words. Whereas words usually are known to a spelling correction system, we have to add the non-words to its vocabulary in order to improve the quality of its corrections.

On the level of single words in German, non-words come from various sources and comprise, among others, false feminine derivations of certain masculine nouns (*Wandererin, *Äbtin), false plurals of nouns (*Thematas, *Tertias), non-licensed inflections (*beigem, *lila(n)es) or comparisons (*lokaler, *minimalst) of certain adjectives, false comparisons (*nahste, *rentabelerer), wrong names for the citizens of towns (*Steinhagener, *Stadthäger), etc. Some out-dated forms (e.g.: Preißelbeere, verkäufst, abergläubig) can likewise be treated as non-words.

It's on the level of compounds that words rather than non-words come into consideration again when we look for contextual constraints or cooccurrence restrictions that determine orthography beyond the scope of what can be accepted or rejected on the basis of isolated words alone.

For words in German, these restrictions determine, among other things, whether or not certain forms (1) begin with an upper or lower case letter; (2) have to be separated by (2.1) blank, (2.2) hyphen, (2.3) or not at all; (3) combine with certain other forms; or even (4) influence punctuation. Examples are:

(1) Ich *laufe eis*.
versus
Ich *laufe auf dem Eis*.

Er dürfte *Bankrott machen*.
versus
Er dürfte *bankrott sein*.

(2.1) Sie kann nicht *Fahrrad fahren*.
versus
(2.3) Sie kann nicht *radfahren*.

(2.1) Es war *bitter kalt*.
versus
(2.3) Es war ein *bitterkalter* Tag.

(2.2) Er liebt *Ich-Romane*.
versus
(2.3) Er liebt Romane in *Ichform*.

(3) *Betonblöcke* vs. **Betonblocks*
versus
Häuserblocks vs. **Häuserblöcke*

(4) Er rauchte, *ohne daß* sie davon wußte.
versus
*Er rauchte *ohne, daß* sie davon wußte.

3 Method

According to a distinction made in the literature (Pollock and Zamora, 1984; Salton, 1989), there are two main approaches in automatic spelling correction: While the 'absolute' approach "consists of using a dictionary of commonly misspelled words, and replacing any misspelling detected in a text by the corrected version listed in the dictionary" (Salton, 1989), the 'relative' approach consists of locating in a conventional dictionary with correct spellings words "that are most similar to a misspelling and selecting a correction from these. Generally, the selection method is based on maximizing similarity or minimizing the string-to-string edit distance" (Pollock and Zamora, 1984).

Although there is some use of 'absolute' methods in some systems (Pollock and Zamora, 1984), "referencing a dictionary of correctly spelled words" (Frisch and Zamora, 1988) is standard. On that basis, most of the purely motoric single word errors, or "typographical

errors" (Berkel and Smedt, 1990), can be corrected. Some conventional systems additionally try to cope with a certain subset of cognitive, or "orthographical" (Berkel and Smedt, 1990), errors which "result in homophonous strings" and involve "some kind of phonemic transcription" (Berkel and Smedt, 1990) for their correction.

Since the cognitive errors outlined in 1 and 2 above are non-standard, in the sense that they are neither motoric (by definition) nor phonologically motivated, a straightforward method to correct them is the 'absolute' one of directly encoding error patterns in a lexicon and replacing each matching occurrence in a text by the correction listed in the system lexicon.

Now, in order to treat single (non-) words and compounds in a uniform way, each entry in the system lexicon is modelled as a quintuple <W,L,R,C,E> specifying a pattern of a (multi-) word W for which a correction C will be proposed accompanied by an explanation E iff a given match of W against some passage in the text under scrutiny differs significantly from C and the - possibly empty - left and right contexts L and R of W also match the environment of W's counterpart in the text.

Disregarding E for a moment, this is tantamount to saying that each such record is interpreted as a string rewriting rule

W --> C / L____R

replacing W (e.g.: Bezug) by C (e.g.: bezug) in the environment L____R (e.g.: in____auf).

The form of these productions can best be characterized with an eye to the Chomsky hierarchy as unrestricted, since we can have any non-null number of symbols on the LHS replaced by any number of symbols on the RHS, possibly by null (Partee et al., 1990).

With an eye to semi-Thue or extended axiomatic systems one could say that a linearly ordered sequence of strings W, C1, C2, ..., Cm is a derivation of Cm iff (1) W is a (faulty) string (in the text to be corrected) and (2) each Ci follows from the immediately preceding string by one of the productions listed in the lexicon (Partee et al., 1990).

Thus, theoretically, a single mistake can be corrected by applying a whole sequence of productions, though in practice the default is clearly that a correction be done in a single derivational step, at least as long as the system is just operating on strings and not on additional non-terminal symbols.

Occurrences of W, L, and R in a text are recognized by pattern matching techniques. An error pattern W ignores the particularly error-prone aspects upper/lower case and word separator (see the examples in 2 above). It thus matches both the correct and incorrect spellings with respect to these features.

Beside wildcards for characters, like "*", a pattern for W, L, or R may contain also wildcards for words allowing, for example, the specification of a maximal distance of L or R with respect to W. Since the types of errors discussed here only occur within sentences, such a distant match has to be restricted by the sentence boundaries. Thus, by having the system operate sentencewise, any left or right context is naturally restricted to be some string within the same sentence as W or to be a boundary of that sentence (e.g.: a punctuation mark).

Any left or right context is either a positive or a negative one, i.e., its components are homogeneously either required or forbidden in order for the corresponding rule to fire. So far it has not been necessary to allow for mixed modes within a left or right context.

In case a correction C is proposed to the user, additionally a message will be displayed to him identifying the reason why C is correct rather than W. Depending on the user's knowledge of the language under investigation, he can take this either as an opportunity to learn or rather as a help for deciding whether to finally accept or reject the proposal.

There are two kinds of explanations, absolute and conditional ones. Whereas absolute rules indicate that the system has necessary and sufficient evidence for W's deviance, there clearly are cases where either W or C could be correct and this question cannot be decided on the basis of the system's lexical information alone. In these cases, a conditional or if-then explanation is given to the user offering a higher-level decision criterion which the system itself is unable to apply.

Take, as an example, the sentence

Dieser Film betrifft Alt und Jung.

which clearly allows for two readings, one which renders "Alt und Jung" as the false spelling of the idiomatic expression "alt und jung" meaning "everybody", and another one which takes "Alt und Jung" as the correct form that literally designates the old and the young while excluding the middle-aged. Thus, substitutability by "jedermann" (i.e.: "everybody") would be an adequate decision criterion to convey to the user.

Although the method described above introduces a new kind of lexical data, its (higher-level) error correction still operates on nothing but strings. No deep and time-consuming analysis, like parsing, is involved.

Restricting the system that way makes our approach to context-sensitiveness different from the one considered in (Rimon and Herz, 1991), where context sensitive spelling verification is proposed to be done with the help of "local constraints automata (LCAs)" which process contextual constraints on the level of lexical or syntactic categories rather than on the basic level of strings. In fact, proof-reading with LCAs rather amounts to genuine grammar checking and as such belongs to a different and higher level of language checking.

Context sensitive spelling checking, as proposed here, can be regarded as a checking level in its own right, lying in between any checking on word level and grammar checking. It thus could complement the two-level checker discussed in (Vosse, 1992) by correcting especially those errors in idiomatic expressions, like "te alle tijden" -> "te allen tijde", which cannot be detected on word or sentence level; compare (Vosse, 1992).

4 A Processing Model

A good model of the system is given by a deterministic multitape Turing machine (Hopcroft and Ullman, 1979) consisting of a finite control with, in effect, three tapes and tape heads. The following description relates to sentence level:

Initially, the input appears on the first tape with each of the tape's cells containing either a word, a blank (symbolized below by a single "B"), or a left or right sentence boundary symbol. Thus, any input sentence can be stored by a finite sequence of cells.

The second tape holds a read-only copy of the initial text. While the first tape will be rewritten, the second serves just as a reference tape. The third tape is also read-only, it holds the finite sequence of lexicon entries.

Consider the following snapshot of the system

```
T1:   B in B Bezug B auf B
             ^
T2:   B in B Bezug B auf B
             ^
T3:   /b/ezug (1in )1auf bezug
```

where "Bezug" has been scanned on the reference tape T2, and a pattern /b/ezug has been found in the lexicon T3 that ignores upper/lower case in the match but requires a lower case "bezug" just in case "in" can be found as 1 word to the left (as is expressed by "(1in") and "auf" can be found 1 non-blank cell to the right on T2.

Since the corresponding contexts of /b/ezug can be verified on T2 (by simply moving T2's head ^ to the respective cells, scanning their contents, and comparing these with the relevant information on T3), finally the error "Bezug" is corrected on T1 and a new checking cycle is started with the next word:

```
T1:   B in B bezug B auf B
                       ^
T2:   B in B Bezug B auf B
                       ^
```

Note, as should be clear from the outset, that a previous correction on T1 is not available as a context for any next word under scrutiny, but only the uncorrected words on T2 are.

Thus, if it were counterfactually the case that "auf" had to be corrected somehow whenever it appeared to the right of "bezug" as opposed to "Bezug", and given the input of the above example, our system - though producing "bezug" as the left context of "auf" on T1 - would clearly fail to correct "auf" since it would still be taking any context from T2.

Although one can think of other, more realistic, cases (like, e.g., "daß ich Eis laufte" -> "daß ich eislief") which require two or more correction steps such that at least one of these steps ("Eis lief" -> "eislief") depends on another one ("laufte" -> "lief"), there clearly are other alternatives (like writing clever lexical entries) beside giving up reading from T2.

Giving up T2 would mean to give up the simple working hypothesis that all the higher-level errors within a given input sentence can be corrected independently. As a consequence, the system would become much more complex and, probably, less efficient.

For German, we have not yet faced any (significant amout of) data that would justify a more complex redesign of the system. However, since the data captured in the system's lexicon covers at present some 50 % of the relevant phenomena compared to the Duden (Berger 1985), the ultimate complexity of the system has to be regarded as an open and empirical question.

5 Status of Implementation

A first prototype of the system described above has been developed in C under UNIX within the ESPRIT II project 2315 "Translator's Workbench" (TWB) as one of several separate modules checking basic as well as higher levels of various languages [like grammar and style; see (Thurmair, 1990) and (Winkelmann, 1990)].

A derived and extended ß-release version - covering 3.000 rewriting rules - has been integrated into both a proprietary text processing software under DOS and Microsoft's WINWORD 1.1 under MS WINDOWS 3.0. In each case it runs independently from the built-in standard spelling verifier, although this is not transparent to the user who perceives just one proofreader checking each sentence of a text twice, i.e., on two different levels.

On both these implementations, some problems have received practical solutions to an acceptable degree.

For example, the problem of mistaking an abbreviation for the end of a sentence (because both end with a dot), which could prevent a context from being recognized, is 'solved' by having the sentence segmentation routine always read beyond a known abbreviation. This might result eventually in taking two sentences to be one, but would, of course, not disturb intra-sentential error correction. Nothing, however, prevents the system from stopping at an unknown abbreviation and thereby falling short of a context it otherwise would have recognized. From this it is clear that the system should at least know the most frequent abbreviations of a given language.

Likewise, the formatting information of a text is preserved to a very high degree during correction, as it should be. Nevertheless, there naturally are cases where some such information will get lost as is clear from the simple fact that there can be shrinking productions reducing n differently formatted elements on the LHS to m elements on the RHS, with $m < n$. But these are borderline cases.

What is less acceptable, for each of the implementations mentioned above, is the lack of integration of the checking on the various levels.

Thus, it may happen that the checkers - running one after the other over the same text - disturb each other's results by proposing antagonistic corrections with respect to one and the same expression: Within the correct passage "in bezug auf", for example, "bezug" will first be regarded as an error by the standard checker which then will propose to rewrite it as "Bezug". If the user accepts this proposal, he will receive the exactly opposite advice by the context sensitive checker.

On the other hand, checking on different levels could nicely go hand in hand and produce synergetic effects: For, clearly, any context sensitive checking requires that the contexts themselves be correct and thus possibly have been corrected in a previous, possibly context free, step. The checking of a single word could in turn profit from contextual knowledge in narrowing down the number of correction alternatives to be proposed for a given error: While there may be some eight or nine plausible candidates as corrections of "Bezjg" when regarded in isolation, only one candidate, i.e. "bezug", is left when the context "in____auf" is taken into account.

Thus, there is a strong demand for arriving at a holistic solution for multi-level language checking rather than for just having various level

experts particularistically hooked together in series. This will be a task for the future.

6 Ongoing Work

As an inhouse test revealed, it is very important for users to have the possibility to add data to the system. As a consequence of that we are currently developing a higher-level user dictionary that will accept and support entering context-sensitive (multi-) words of the kind discussed in this paper.

At the same time, data acquisition is still going on. Since, unfortunately, there is (to our knowledge) no ready-made corpus of higher-level errors available, the collection of data is a time consuming process. The best reference book for German seems to be the Duden (Berger 1985), yet it consists of an unstructured mixture of all possible kinds of errors and often presents a paradigmatic example rather than all the members of a given error class.

As concerns languages other than German, we take it that a similar approach is feasible for them as well. Although in comparison with, e.g., English, French, Italian, and Spanish, German seems to be unique as concerns the relevance of the context for upper/lower case spellings in a large number of cases, there are at least, as indicated in (Gross, 1986), the thousands of compounds or frozen words in each of these languages which are clearly within reach for the methods discussed.

References

Dieter Berger, editor. *Richtiges und gutes Deutsch. Wörterbuch der sprachlichen Zweifelsfälle*. Der Duden in 10 Bänden. Das Standardwerk zur deutschen Sprache, volume 9. Bibliographisches Institut, Mannheim, Germany, 3rd ed. 1985.

Brigitte van Berkel and Koenraad De Smedt 1990. Triphone Analysis: A combined method for the correction of orthographical and typographical errors. In *Proceedings of the Second Conference on Applied Natural Language Processing*, pages 77-83, Austin, Texas, February 1988. Association for Computational Linguistics.

Rudolf Frisch and Antonio Zamora. Spelling assistance for compound words. *IBM Journal of Research and Development*, 32(2): 195-200, March 1988.

Maurice Gross. Lexicon Grammar. The Representation of Compound Words. In *Proceedings of the Eleventh International Conference on Computational Linguistics*, pages 1-6, Bonn, Germany, August 1986. International Committee on Computational Linguistics.

Zellig S. Harris. *Papers in Structural and Transformational Linguistics*. Formal Linguistics Series, volume 1. D. Reidel Publishing Company, Dordrecht, The Netherlands, 1970.

Gerhard Heyer. Probleme und Aufgaben einer angewandten Computerlinguistik. *KI-Künstliche Intelligenz* 3(1): 38-42, 1990.

John E. Hopcroft and Jeffrey D. Ullman. *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley Publishing Company, Reading, Massachusetts, 1979.

Barbara H. Partee, Alice ter Meulen, and Robert E. Wall. *Mathematical Methods in Linguistics*. Studies in Linguistics and Philosophy, volume 30. Kluwer Academic Publishers, Dordrecht, The Netherlands 1990.

Joseph J. Pollock and Antonio Zamora. Automatic spelling correction in scientific and scholarly text. *Communications of the ACM*, 27(4): 358-368, April 1984.

Mori Rimon and Jacky Herz. The Recognition Capacity of Local Syntactic Constraints. In *Proceedings of the Fifth Conference of the European Chapter of the Association for Computational Linguistics*, pages 155-160, Berlin, Germany, April 1991. Association for Computational Linguistics.

Gerard Salton. *Automatic Text Processing. The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley Publishing Company, Reading, Massachusetts, 1989.

Gregor Thurmair. Parsing for Grammar and Style Checking. In *Papers presented to the Thirteenth International Conference on Computational Linguistics*, volume 2, pages 365-370, Helsinki, Finland, August 1990. Hans Karlgren, editor.

Theo Vosse. Detecting and Correcting Morpho-syntactic Errors in Real Texts. In *Proceedings of the Third Conference on Applied Natural Language Processing*, Trento, Italy, March/April 1992. Association for Computational Linguistics.

Günter Winkelmann. Semiautomatic Interactive Multilingual Style Analysis (SIMSA). In *Papers presented to the Thirteenth International Conference on Computational Linguistics*, volume 1, pages 79-81, Helsinki, Finland, August 1990. Hans Karlgren, editor.

Harald Zimmermann. Textverarbeitung im Rahmen moderner Bürokommunikationstechniken. *PIK. Praxis der Informationsverarbeitung und Kommunikation* 10: 38-45, 1987.