AUTOMATIC PROOFREADING OF FROZEN PHRASES IN GERMAN

RALF KESE, FRIEDRICH DUDDA, MARIANNE KUGLER
TA Triumph-Adler AG
TA Research EF
Fürther Str. 212
D-8500 Nuremberg 80
Germany
e-mail: ralf@triumph-adler.de

Abstract

Frozen phrases are introduced as a new level of automatic proofreading in between the standard level of spelling verification of isolated words and the level of grammar checking. The design and the implementation of a corresponding proofreading system are described in detail.

1 Introduction

European languages contain thousands of what Maurice Gross calls "frozen" or "compound words" (Gross, 1986). 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).

Now, these kinds of co-occurrence restrictions (Harris, 1970) determine not only the concrete lexical composition of an individual compound word but also its spelling.

Consider, as an example, the German noun "Bezug" which - like any other noun in German - starts with a capital letter and occurs as a free form - e.g. in contexts like "mit Bezug auf" - such that its co-occurrents can vary freely. There is a single exception to this rule, namely the phrase "in bezug auf", which is entirely frozen, in the sense that it excludes any variation of its parts or structure, and which, by the same token, restricts the spelling of the noun to lower case.

For examples like this we introduce the term "frozen phrases" as referring to (the sub-class of) those frozen words that are compounds of several words. As Zimmermann (Zimmermann, 1987) points out for multi-words in general, frozen phrases are clearly out of scope of standard spelling correction systems due to the fact that these systems check for isolated words only and disregard the respective contexts. Yet, as Gross (Gross, 1986) indicates, at least the entirely and nearly entirely frozen forms can be recognized and compared with the help of simple string operations. Thus, these kinds of frozen phrases are accessible to the methods of classical automatic proofreading.

Following Gross (Gross, 1986) and Zimmermann (Zimmermann, 1987), we propose to further extend standard spelling correction systems onto the level of frozen phrases by simply making them capable of treating more than a single word at a time. This implies that we arrive at a context-sensitive system.

Focusing on individual co-occurrences (Harris, 1970), the proposed extension will be a conservative one, in the sense that it requires just widening the scope of the string matching/comparing operations that classically are used in spelling correction systems. No deep and time-consuming analysis, like parsing, is involved.

Restricting the system that way makes our approach different from the one considered in (Rimon and Herz, 1991), where a 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.

2 Phenomenology

The essential feature of the new level of automatic proofreading is that lexicalized co-occurrence restrictions determine the orthography of whole phrases beyond the scope of what can be accepted or rejected on the basis of isolated words alone.

For 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 System Design

Like conventional spelling correction of isolated words, proofreading of frozen phrases is a lexicon-based process.

However, while in conventional spelling correction "referencing a dictionary of correctly spelled words" (Frisch and Zamora, 1988) is standard, proofreading of the higher-level orthographic features mentioned in 2 above can be done on the basis of a lexicon that encodes the corresponding error patterns directly.

Thus, each entry in the system lexicon is modelled as a quintuple <W,L,R,C,E> specifying an error pattern of a (multi-) word W for which a correction C will be proposed accompanied by an explanation E just in case 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 frozen phrase "alt und jung" meaning "everybody", and another one which takes "Alt und Jung" as the correct free 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.

In its present design, the system is based on the simplest possible working hypothesis, i.e. the assumption that the higher-level or cognitive errors in a sentence can be corrected independently. The intuition behind this assumption is that, normally, cognitive (or orthographical) errors are by far less frequent than ordinary motoric (or typographical) errors (for this distinction see Berkel and Smedt, 1990), and that, as a consequence of this, they occur within distinct contexts such that it is excluded that the correction of one error be conditional upon the correction of another.

According to this assumption, each sentence of the text is processed only once in the following manner: The system reads from a plain text copy, T2, of the original formatted text, T1, processes the errors on T2 one after the other, and finally writes the corresponding corrections to T1, without also writing them to T2.

Now, an abstract counterexample to the working hypothesis can be construed quite easily: Given a sentence containing the sequence of errors

... W1 W2 ...

and given lexical rules

(R1) W1 --> C1

(R2) W2 --> C2 / C1___

rewriting (i.e. correcting) W1 as C1 in any context, and W2 as C2 if preceeded by C1, then, clearly, the system will correct W1 but will fail to correct W2.

For the system to also correct W2, it must be able to take its previous output into account again. That is, it should not only read from T2 but also write to it.

However, to allow corrections to be written also to T2 would mean to stepwise introduce new contexts on T2. As a consequence, the system would have to redo the checking of the whole sentence each time a correction has been made, for this correction might well be a relevant context of some previously encountered expression.

Thus, giving up the working hypothesis would result in having the system take multiple runs through one and the same sentence instead of a single run, and this, of course, would drastically reduce system performance.

Fortunately, we have not yet come across any (significant amount of) data that would justify such a 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.

4 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 orthogonal modules checking basic as well as higher levels (like grammar and style; see (Thurmair, 1990) and (Winkelmann, 1990)) of various languages.

A derived and extended version - covering 3.000 rewriting rules and some 80 explanations - has been integrated into both a proprietary text processing software under DOS and Microsoft's WORD FOR WINDOWS, version 1.1.

This extended version has been combined with a conventional spelling verifier to form a single proofreader for the user. Internally, however, its hidden sub-modules are still totally independent from one another and process a sentence one after the other.

Thus, it may happen that the checkers 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 frozen phrases 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.

The same interdependence seems to exist with respect to higher levels of language checking. At least it can be argued that a grammar checker will profit from integration with a frozen phrases checker: For nothing but an expert for frozen phrases can verify the correctness of (idiomatic) expressions like "ruhig Blut bewahren" or "auf gut Glück", and thus prevent a grammar checker from flagging the missing inflection of the adjective ("ruhig" or "gut") in attributive position.

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 the task for the near future.

5 Conclusion

As concerns the German language, we have shown that there is a well-defined level of automatic proofreading in between the classical level of checking isolated words and the more advanced level of grammar checking. On the basis of work done by Zellig Harris (Harris, 1970) and Maurice Gross (Gross, 1986) we have identified this level as that of so-called "frozen phrases", and we have proposed and implemented an automatic proofreading system that covers a significant amount of the frozen phrases of German.

We take it that a similar approach is feasible for languages other than German as well. Although in comparison with English, French, Italian, and Spanish, German seems to be unique as concerns co-occurrence restrictions 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. In addition, the generalizability of our approach has received some confirmation from case studies carried out for English and Italian.

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.

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.

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.