UNIVERSIDADE DE SÃO PAULO

STYLE AND GRAMMAR CHECKERS FOR BRAZILIAN PORTUGUESE

M.G.V. NUNES
R. HASEGAWA
S. KAWAMOTO
M.C.F. DE OLIVEIRA
M.A.S. TURINE
C.M. GHIRALDELO
O.N. OLIVEIRA JR.
C.R. RIOLFI
N.S. SIKANSKI
T.B. MARTINS

No. 25

NOTAS

Instituto de Ciências Matemáticas de São Carlos

Instituto de Ciências Matemáticas de São Carlos

ISSN - 0103-2577

STYLE AND GRAMMAR CHECKERS FOR BRAZILIAN PORTUGUESE

M.G.V. NUNES
R. HASEGAWA
S. KAWAMOTO
M.C.F. DE OLIVEIRA
M.A.S. TURINE
C.M. GHIRALDELO
O.N. OLIVEIRA JR.
C.R. RIOLFI
N.S. SIKANSKI
T.B. MARTINS

No. 25

NOTAS DO ICMSC
Série Computação

São Carlos
Mai./1996

STYLE AND GRAMMAR CHECKERS FOR BRAZILIAN PORTUGUESE

RESUMO

Este trabalho reporta o desenvolvimento de um revisor gramatical para o Português do Brasil. Este sistema usa regras empíricas baseadas em marcadores lexicais e retóricos, e é implementado usando-se redes ATN. As regras são implementadas de modo iterativo, sendo que o desempenho das regras é verificado através de testes em um corpus formado por textos reais. A ferramente detecta um grande número de erros, incluindo erros de gênero e número de adjetivos e substantivos, e concordância verbal em sentenças em forma canônica. Entretanto, ela ainda não trata alguns casos de concordância verbal em sentenças longas, para as quais uma análise sintática automática é extremamente difícil. A ferramenta também detecta alguns erros ditos mecânicos, usualmente não detectados por corretores ortográficos. Além disso, ela verifica a adequação de um texto a uma dada audiência, através do cálculo de um índice de legibilidade que foi adaptado do índice Flesch para o inglês.

Style and Grammar Checkers for Brazilian Portuguese

M.G.V. Nunes, R. Hasegawa, S. Kawamoto, M.C.F. de Oliveira and M.A.S. Turine
Instituto de Ciências Matemáticas de São Carlos - USP
CP 668 - 13560-970 São Carlos, SP (Brazil)
e-mail: mdgvnune@icmsc.sc.usp.br

C.M. Ghiraldelo and O.N. Oliveira Jr.
Instituto de Física de São Carlos, USP
CP 369 - 13560-970 São Carlos, SP (Brazil)
FAX: +55 162 713616; e-mail: chu@ifqsc.sc.usp.br

C.R. Riolfi and N.S. Sikanski
Instituto de Estudos da Linguagem, Unicamp, Campinas, SP (Brazil)

T.B. Martins
Departamento de Letras
Universidade Estadual Paulista, Unesp, Araraquara, SP (Brazil)

Abstract

This work reports on the development of a grammar checker for Brazilian Portuguese that uses empirical rules based on lexical and rhetorical markers and implemented as augmented transition networks (ATNs). The rules were implemented in an iterative way, having their performance assessed through tests in a corpus of authentic texts. The tool can already detect a large number of errors, including those related to gender and number of adjectives and nouns, and verb agreement in sentences with canonical order. However, it cannot yet cope with some cases of verb agreement in long sentences for which an automatic syntactic analysis is extremely difficult. The tool also detects some typing errors that are not usually detected by a spelling checker, and verifies the adequacy of a piece of text for a given audience by calculating readability scores from an adaptation to Brazilian Portuguese of the well-known Flesch scale.

1. Introduction

Grammar checkers have become popular over the last few years with the wide availability of personal computers and increasingly sophisticated desktop software packages. For English, for instance, checkers have been available since the early eighties [Cherry-82] and they are now an integral part of most commercial wordprocessors [Bolt-92]. Generally, these tools employ a lexicon in which words are tagged according to their syntactic function in a sentence and, in some cases, also according to their possible semantic functions. Based on such a categorization, rules are then applied for detecting non-grammatical constructions, normally within a symbolistic approach [Suri-93].

In the case of Brazilian Portuguese, on the other hand, only recently interest arose from software and computer companies in the development of similar tools oriented towards Brazilian writers. In addition to the considerably smaller economic appeal, there are several practical difficulties to be faced in developing grammar checkers for Brazilian Portuguese. First of all, no tagger is available which could be used as an off-shelf tool. Furthermore, studies of the current use of Portuguese are considerably more scarce than for other languages

(English in particular). There are no updated data on word frequency and dictionaries of word usage are not available.

We have recently instituted a project for developing writing tools for the wordprocessor REDATOR of Itautec/Philco S/A. This grammar checker has already been made commercially available and is now being further improved. This paper is an account of the problems faced along the way and of the solutions adopted. Theoretical aspects involved in natural language processing are only treated within the context of our development efforts. Yet, the project encompasses research into a broad spectrum of linguistic aspects in order to fill in the gap arising from the lack of fundamental studies on language usage, in addition to the specific implementation of rules for checking grammar and their integration into the wordprocessor. The implementation of the grammar checker has followed three basic principles: i) it is mostly error-driven; ii) it is based on extensive testing on a large corpus of texts; iii) it is iterative. We shall elaborate upon each of these principles in Section 2 where the development approach is discussed more thoroughly.

Still regarding the approach followed in the computational implementation, we chose to use the philosophy of ATNs (Augmented Transition Networks) [Woods-70]. ATN's provide a suitable model for natural language analysis, because of their capability to be adapted according to the nuances of prose expression [Miller-87]. As the implementation of the rules is mostly error-driven, ATNs are usually associated with non-grammatical constructions, even though ATNs have been also created which contain syntactic features of correct sentences, especially those possessing high frequency rhetorical markers. Section 3 provides an overview of the types of error detected by the grammar checker, and how they are detected. In addition to checking grammar, the tool also provides readability scores, which have been used for decades for English and other languages, and are, to our knowledge, for the first time applied to Brazilian Portuguese. After a survey on various types of readability scores, we chose to adapt Flesch scores to Brazilian Portuguese. The results of this work is presented in Section 4. Section 5 presents the conclusions and discusses further work to be undertaken.

2. The Development Approach

Potential users for the grammar checker include secretaries and clerical officers. The language register to be considered, therefore, is at the same time restricted - since specific areas such as science and literature do not need to be covered - and broad because the checker must deal with colloquialisms and problems from spoken language interference. Furthermore, the checker is to be used under completely open conditions where any particular type of sentence (syntactically) may occur. As Procter [Procter-93] points out, grammatical templates encoded in rule-based systems often cover only a small proportion of the text held in a corpus. We decided therefore to identify contexts in which automatic language processing might be achieved. The starting point was an investigation into the most frequent types of error made by the target users, for only then proceed towards the design of rules to detect those errors.

2.1 Error analysis and testing with a corpus

Since our target users are secretaries and clerks working in office environments we had to analyse texts[1] produced by high school students (equivalent to students who have obtained O-levels in Britain) and first-year University students. The analysis of 208 pieces of text showed that the most frequent errors are associated in decreasing order of importance with: spelling, *crase* (contraction of the preposition "*a*" with the definite article "*a*", as will be explained later), punctuation, noun and verb agreement, inadequate use of words and improper use of idioms (interference from spoken language). These results are illustrated in Figure 1. It must be stressed this study is by no means comprehensive and served the sole purpose of identifying high frequency errors. As discussed in the next Section, we envisaged a tool which can cope with a large number of errors in all of these categories. The obvious exceptions are the spelling mistakes which are dealt with by the wordprocessor's spelling checker.

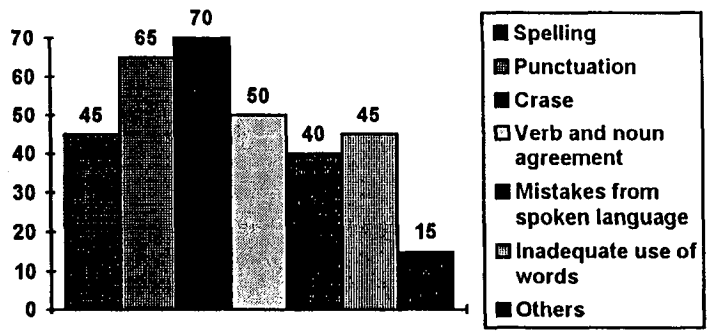

Figure 1: Results of the error analysis.

In deciding which errors should be addressed by the grammar checker, we had to consider those which were amenable to an automatic computer treatment. Also, a survey was carried out on the most important grammars of contemporary Brazilian Portuguese and a number of discrepancies were noted as to whether some linguistic constructions should be considered erroneous. Such discrepancies appear because grammarians tend to take different views about various linguistics facts. For example, it takes longer for some of them to accept incorporation of colloquialisms and neologisms - of common usage in speech - into grammar. Traditional grammars for Brazilian Portuguese still employ classical texts of Portuguese - such as those by Camões (roughly equivalent to Shakespeare for English) to provide examples of language usage. Yet, not surprisingly there is enourmous difference between classical and contemporary Brazilian Portuguese. Therefore, for the implementation of the rules decisions regarding erroneous structures were supported by linguistic studies on the written language use accepted by high quality newspapers and magazines.

Another important principle adopted was the application of extensive testing of the rules on a corpus of authentic texts extracted from books, newspapers, magazines and other publications, and also on a subcorpora

[1]The term text is referred to here as any piece of written material, usually with at least one coherent paragraph.

entirely dedicated to unrevised texts produced by target users. While the latter type of text was used in identifying the most important errors to be detected, the corpus containing error-free texts was used for checking the generation of false errors by the rules. By **false error** we mean an unnecessary intervention of the system which would induce the user to replace a correct linguistic construction either by a correct alternative or by an incorrect one. This was an essential step taken in the development procedure, since the user should not be interrupted with unnecessary messages, even though in some cases he/she could actually recognise it as false error.

The corpus of texts currently available possesses 1.5 million words and takes 60 MBytes of hard disk memory. It is heterogeneous because it is aimed at covering contemporary use of Portuguese, but it does not include literary texts which possess specificities and linguistic constructions unlikely to be employed by target users. This is a major difference between the corpus developed in this project and those corpora utilized in preparing dictionaries [Bidderman-93, Borba-93, Sinclair-87].

2.2 Computational Implementation

The starting point for this project was a lexicon with around 1 million words used by a commercially available spelling checker [Itautec-93], whose entries were all tagged according to the categorization of a usual dictionary. Therefore, words with multiple associated meanings and multiple syntactic functions receive several attributes, with no hierarchical classification. Although such a huge lexicon obviously brings advantages due to its comprehensiveness, it causes problems of over categorization. The latter occurs because most dictionaries for Brazilian Portuguese quote all possible meanings and syntactic functions of a given word, with no regard for the actual frequency of use. Thus, one may end up with the categorization of *"humano"* as a verb which, albeit correct, is actually a very obscure and rare use for this word. This type of problem is similar to those encountered in extracting knowledge with machine-readable dictionaries [Ide-93]. These deficiencies had to be corrected and improvements in the categorization aspect of the lexicon are still underway.

The grammar checker comprises 3 modules implemented in C^{++}, and its general architecture is depicted in Figure 2. The Mechanical Module detects mechanical (or typing) errors that usually are not spotted by a spelling checker, such as unbalanced parentheses and undue capitalization. The Statistical Module is devoted to obtaining statistical features of a piece of text, including the readability score to be discussed in Section 4. The most important module, though, is the Linguistic Module represented in Figure 3 which is now embedded into the REDATOR wordprocessor [Itautec-95]. The Grammatical Rule Manager activates the lexical tagger that is responsible for identifying and classifying tokens using information from the lexicon. Because the lexical tagger is largely responsible for the performance of the tool as a whole, it does not resort to the large lexicon for the classification of high frequency words. Instead, these words are classified separately in a list which speeds up the classification procedure. Once a paragraph has been tagged, rules for detecting grammatical errors are applied by activating the ATNs base.

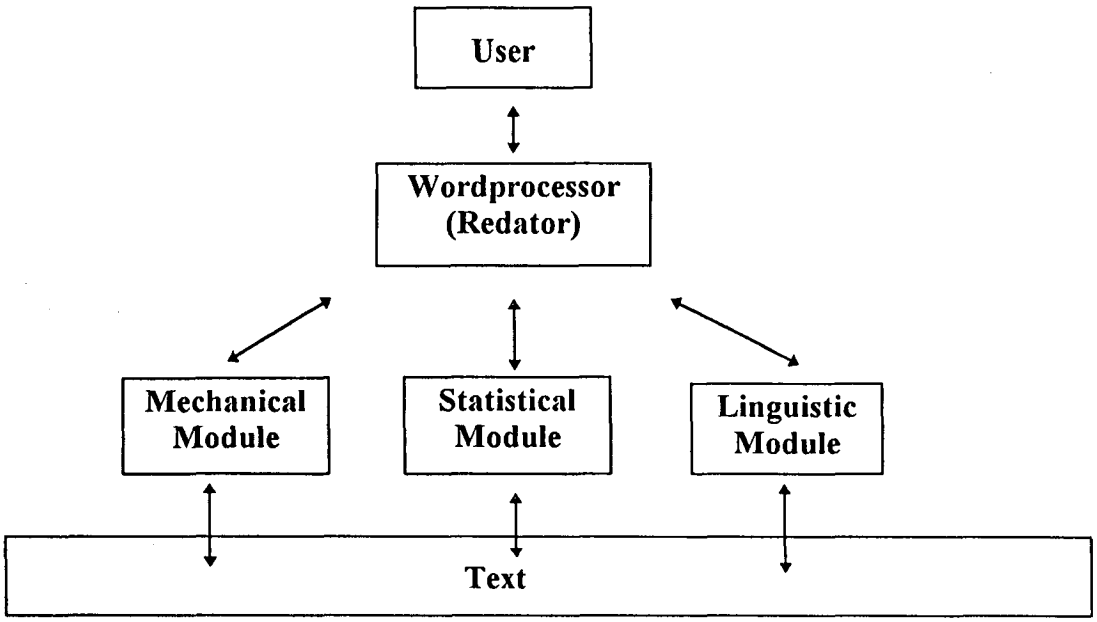

Figure 2: General architecture of the grammar checker.

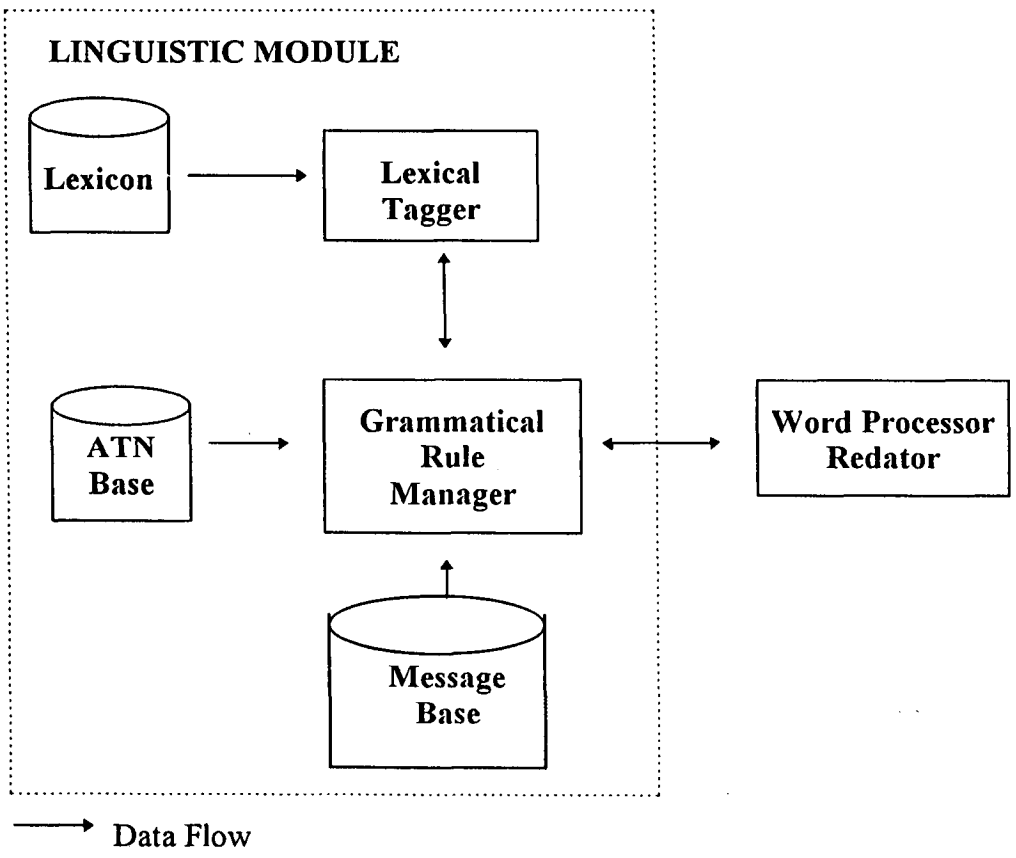

Figure 3: Architecture of the Linguistic Module.

There are essentially two distinct types of ATN in the base. Since the implementation is mostly error-driven, ATNs of the first type represent incorrect linguistic constructions. For the core of noun and verb agreement rules, however, ATNs were created to represent correct constructions, as illustrated in the next Section for a rule on verb agreement. The basic steps for the implementation of ATNs were: i) linguistic rules were created for detecting a given type of error; ii) these rules were then implemented computationally in the form of ATNs. This was followed by extensive testing in the corpus for searching for false and/or non-detected errors; iii) using the results from the testing, the rules were improved upon, and additional testing was performed. That is, steps ii) and iii) were carried out iteratively.

3. Types of Error Detected

We have tried to cover the largest possible variety of errors made by target users, in particular those of high frequency. In this section we shall give examples of the types of error detected by the grammar checker.

3.1 *Crase*

Apart from spelling mistakes, which are adequately dealt with by the spelling checker, the second most frequent type of mistake is associated with the *crase*, as shown in the histograms of Figure 1. *Crase* is the contraction between the preposition "*a*" (which usually means *"to"* in English) and the definite article "*a*" (singular, feminine), as in the example below:

Ele foi à piscina. (*He went to the swimming pool.*)

Detection of the inadequate use of *crase* is reasonably straightforward for there are well established rules which govern it. The most important rule is that a crase should not be used, in general, before a masculine word. Using a set of 20 rules the grammar checker detects practically 100% of the most common errors. The lack of crase, on the other hand, is more difficult to detect, especially because it is difficult to distinguish in a computational setting between the article "*a*" and the preposition "*a*". There are cases, nevertheless, where this distinction is made possible. In order to cover most cases, we are now investigating all possible contexts in which the preposition should accompany the article, thus requiring crase. Corpus analysis is proving essential for this task.

3.2 Punctuation

Errors in punctuation may be quite serious since the meaning of a whole sentence can be changed, especially by excessive punctuation. The survey on the texts by target users clearly indicated tremendous difficulties in applying punctuation rules. Some of the errors are closely linked to context and therefore their detection is extremely difficult. There are, however, two high frequency types of error that are easily detected. The first one is related to the lack of a comma when its use is mandatory. Examples are conjunctions and expressions

such as **mas** (*but*), **porém** (*however*), **isto é** (*that is*), **além disso** (*furthermore*), which should almost invariably follow a comma or another punctuation sign.

The second type of error is critical since it may interfere with comprehension of the sentence, and consists in placing a comma between the subject and its related verb, as illustrated by the example below:

Outra conclusão importante, foi que a condutividade decresceu.

(*Another important conclusion, was that the conductivity was decreased.*)

Correction of this type of error is obviously essential for the analysis of the sentence and the subsequent application of noun and verb agreement rules. This is performed by activating the ATNs associated with the correct, canonical sentence structure. For instance, sequences of the type *[noun phrase (NP) + comma + verb phrase (VP)]*, as in the example above, are almost certainly non-grammatical and will not comply with the ATNs requirements.

3.3 Noun and verb agreement

Noun agreement is an important issue in Portuguese since nouns and adjectives normally possess four different forms (singular (s), plural (p), masculine (m) and feminine (f)) and must usually agree in number and gender. The word *bonito* (beautiful) may take the forms: *bonito* (s,m), *bonita* (s,f), *bonitos* (p,m), *bonitas* (p,f). Our target users generally are aware of the rules for such an agreement but nevertheless they still make a number of mistakes for two main reasons: (i) the words to agree are reasonably far apart, which eventually induces the writer into an error; or (ii) the user simply does not check his/her text, and errors remain which would otherwise be eliminated had the text been revised. The same reasoning applies to the verb agreement rules, where the subject and the verb must agree not only in number but also in person (1st, 2nd or 3rd).

In contrast to the high frequency errors discussed in the previous sub-sections, noun and verb agreement errors may occur in very much open contexts. ATNs were therefore created to represent correct constructions. By way of illustration, we present in Figure 4 the ATN for a simple sentence containing a compound subject of the type *[Noun Phrase 1 + e ("and") + Noun Phrase 2 + Verb]*.

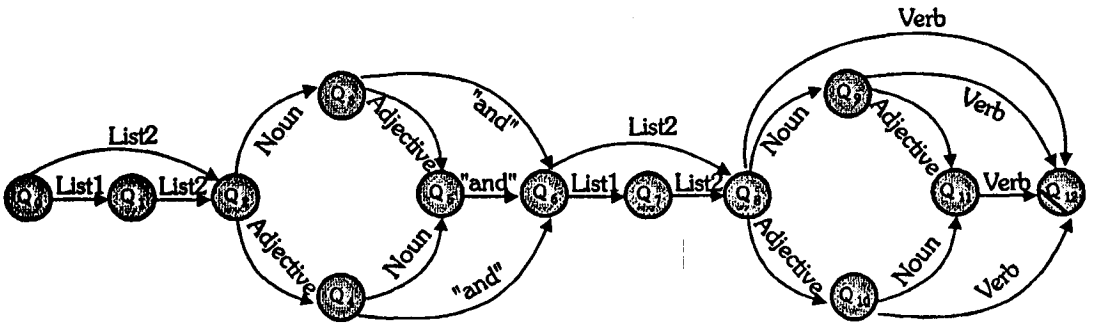

Figure 4: ATN for applying one of the verb agreement rules. In this case the verb must take a plural form.

In Portuguese, noun phrases (NP) may be formed by a noun, an adjective, by bigrams of the type "*noun + adjective*" or "*adjective + noun*" and also include articles, pronouns, etc. Therefore, NP1 and NP2 may include at a given instance words from *list1*, *list2* and bigrams, where:

list1 = (**o, os, a, as, esse(s), essa(s), aquele(s), aquela(s), um(a), algum(a), alguns**, etc.) ("*the*", "*this*", "*these*", "those", "*one*", "*some*");

list2 = (**vários, muitos, diversos, poucos, todo(s), toda(s), cada, meu(s), minha(s), nosso(s)**, etc.) ("*various*", "*several*", "*a number of*", "*a few*", "*all*", "*each*", "*mine*", "*our*").

Longer, subordinate and/or coordinate sentences are sometimes analysed by concentrating on rhetorical markers ("*but*", "*in order to*", "*for*", etc.) and establishing the syntactic features these sentences may possess. Some examples of errors detected are given below:

Os estudantes chegou cedo mas o museu estavam (estava) fechado.
(*The students arrived early but the museum were (was) closed.*)
Os estudantes chegou (chegaram) cedo para apresentar seus trabalhos.
(*The students arrived early for delivering their talks.*)
Foi feito (Foram feitos) vários ajustes na política fiscal
(*Several amendments was (were) made to the fiscal policies*).

The last example shows an error of verb agreement in a passive sentence. Such errors are common because writers tend to consider the singular form as invariable. The first two examples above illustrate not only errors in clauses with canonical order, but also how analysis based on rhetorical markers is employed. In the first example, the appearance of **mas** (*but*) means that only certain types of subordinate sentences can be formed. Using corpus analysis, ATNs were then created which allow one to automatically identify the type of sentence and also

how to apply verb agreement rules. Similarly, in the second sentence the presence of **para** (which may mean "*in order to*" or "*for*") is to be noted. It is even possible, in this case, to define the tense of the verb to follow the marker *para*, as one may expect to occur if the possible sentences containing "*in order to*" in English were analysed.

A comprehensive study is underway for providing ATNs which can cope with almost any type of sentence. Indeed, a large number of errors are already detected by the grammar checker. However, several difficulties remain which are mainly associated with over long sentences that are not easily analysed and with homograph words that may have multiple syntactic functions. Problems of the last type are being minimised by investigating differences in context in which these homograph words appear. With regard to the over long sentences, perhaps the major difficulties are related to incorrect punctuation and to the fact that order of components in Portuguese sentences may vary considerably. Surely, one cannot hope to fully analyse (syntactically or otherwise) a sentence if punctuation is not correct. The way envisaged to deal with this problem is to treat particular cases in very limited contexts.

As far as noun agreement is concerned, the majority of errors are already detected by the checker, even though some difficulties still remain. These are concentrated on the agreement with the article "*a*" which can also be preposition, as already discussed, and on the agreement with homograph words that can be nouns as well as verbs (e.g. **visto** = *visa* (noun) and **visto** = *seen* (verb)). The latter problem may even cause difficulties in the tagging process and therefore verb agreement is also affected. This is being minimised by the identification - through corpus analysis - of the most frequent items to bring such a difficulty.

3.4 Inadequate use of words and improper use of idioms (interference from spoken language)

This is an extremely broad class which encompasses errors associated with often misused words and canned expressions (usually interference from spoken language) and errors caused by specific difficulties with Portuguese. Some of the errors are semantic in nature and may largely depend on context. These could not possibly be treated automatically. Errors amenable to computational treatment included those originating from the inadequate use of (reasonably) fixed expressions and those which possess a semantic marker. The errors in canned expressions were easily detected by applying rules using a pre-compiled list. In our case, 50 expressions are listed. Among those possessing a marker, important errors occur because of incorrect agreement of important verbs, such as *fazer* and *haver*. When employed together with time markers, these verbs mean that a particular period of time has elapsed and must appear in their singular forms. Some examples are given below:

Fazem (Faz) dois meses que ela chegou.
(*She arrived two months ago*)
Estou em Londres a (há) dois dias.
(*I've been in London for two days.*)

Estarei em Londres daqui há (a) dois dias.

(*I will be in London in two days.*)

For the verb **fazer**, it so happens that users normally think that it should agree with the time period, which is not true in this case because the verb is impersonal. In the first example above, they would tend to use **fazem** (plural) instead of the correct form **faz** (singular). For the verb **haver**, on the other hand, confusion often arises because the words **há** (verb **haver**) and **a** (preposition) are homophone. Moreover, the preposition "*a*"should be employed in specific cases, especially when referring to a future time as in the third example above. The verb **haver** also causes problems when applied with the meaning "*to exist*". Again, because it is a so-called impersonal verb it should remain in the singular form which seldom occurs. An example of this type of mistake is illustrated below, where the correct verb form appears between parentheses:

Haviam (Havia) cinco cadeiras na sala.

(*There were five chairs in the room.*)

Another important difficulty faced by writers of Brazilian Portuguese is related to the use of prefixes, whether the prefix should be separated from the main root by a hyphen or should form a single word. Spelling checkers usually do not provide a full treatment of this problem because the prefix and the root may form a set of two separate, valid words. Rules for the use of prefixes were then implemented in the grammar checker which detect the majority of the associated errors. For instance:

neo + **clássico** = **neoclássico** (*neoclassic*)

inter + **nacional** = **internacional** (*international*)

super + **agudo** = **superagudo** (*super acute*)

super + **rentável** = **super-rentável** (*super profitable*)

inter + **hemisférico** = **inter-hemisférico** (*inter hemispherical*)

The checker also verifies the positioning of personal pronouns, which is also a difficulty task for the majority of Portuguese users, as in the following example:

Ele não enviou-me a encomenda. (incorrect)

Ele não me enviou a encomenda. (correct)

(*He did not send me the parcel.*)

4. Calculating Readability Scores

Readability formulas have been extensively used over the last fifty years for estimating the readability of written English texts (for a review see [Klare-74]) with a number of different purposes. They are widely used for assessing the adequacy of remedial reading texts, military training manuals, elementary-school books, and television captioning for the hearing impaired [Royer-90]. After an extensive search into the literature, we were unable to find any report on readability studies for Brazilian Portuguese. This has prompted us to investigate

whether the formulas, originally developed for English, would also be meaningful for the analysis of Brazilian Portuguese texts. Several formulas, viz. Coleman-Liau, Kincaid, ARI and Flesch were employed in 120 passages of textbooks, and the subsequent analysis led to practically the same conclusions. We therefore concentrated on the results from the Flesch scores, obtained from the equation [Flesch-48, Cherry-82]:

$$Reading\ Score = 206.835 - 84.6*syl.per.wd - 1.015*wds.per.sent$$

where *syl.per.wd* is the average number of syllables per word and *wds.per.sent* is the average number of words per sentence.

Our first results showed that Flesch scores were systematically lower for Portuguese texts than for their English counterparts. A direct comparison was then made by calculating the Flesch score for original texts in English from Introductory Physics textbooks and also the Flesch score for the same texts translated into Portuguese. The Portuguese passages scored in average 42 points lower than the English passages. Therefore, for comparing the ranges of Flesch scores, we added 42 points to all Flesch scores obtained for the Portuguese texts. The scores for the various texts fell into four levels of readability, corresponding to the first four years of education (*very easy* - Flesch scores from 75 to 100), to the grades from the 5th to the 8th year (*easy* - scores from 50 to 75), to high school and college (*fairly difficult* - scores from 25 to 50) and to academic texts (*very difficult* - scores below 25). Taken together these results indicate that the application of formulas to predict readability is appropriate in this case and may provide an important tool for helping one to assess the adequacy of textbooks to their intended audiences. It must be borne in mind, however, that formulas such as Flesch's are aimed at predicting readability and not comprehensibility. Other factors that are highly relevant for text comprehension are not contemplated in these readability formulas, and therefore readability scores should be used with extreme caution [Mayo-93, Olson-86, Meade-91, Rush-85].

5. The user interface

The grammar checker was incorporated as an additional function in the wordprocessor REDATOR of Itautec/Philco S/A, whose user interface is illustrated in Figure 5. The only design decision regarding the user interface directly related to the grammar checker lied in the choice of messages to be returned to the users. There are a few types of message that vary as the checker may be sure about an error or may only advise the user with respect to linguistic structures that could (or could not) be incorrect depending on the context. There are two levels of information returned to the user: a first window shows only short messages advising the user on the type of error detected. The user can then request further information on that type of mistake, which is displayed in a second window. Explanations in this latter window do make use of meta-terms, which are avoided in the initial messages. A minigrammar is also being built to be incorporated into the system, so that users may get further information in any particular subject which may have caused difficulties.

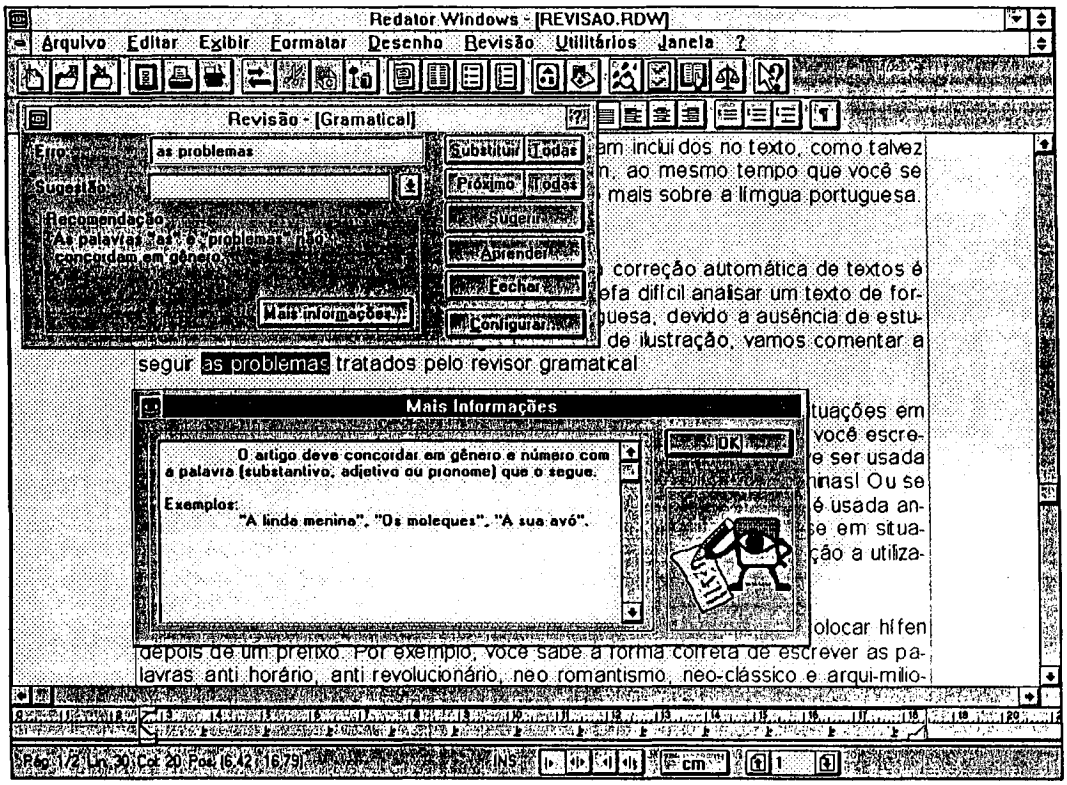

Figure 6: Grammar checker window.

6. Final Remarks

The approach adopted in this project for implementing the grammar checker has been proven adequate for Brazilian Portuguese. Since the implementation is mostly error-driven broad classes of high frequency errors could be detected. Computationally, the implementation of approximately 200 rules in the form of ATNs has also been shown to be very efficient. For example, when running in the batch mode, the Grammar Checker is capable of checking a Ph.D. thesis of 130 pages (roughly 17000 to 20000 words) in only five minutes. In average, for this amount of material less than 10 false errors occur. This is certainly an extremely low value, which was only possible due to extensive testing in a large corpus of authentic texts. As for non-detected errors, so far we have no systematic study for a quantitative measure be obtained.

Further work will be concentrated on solving some of the problems already indicated throughout the paper. These include analysis of long, complex sentences, attempts to identify when the "*a*" is an article or preposition and investigations into homograph words that often cause ambiguities. The corpus of authentic texts is being extended considerably now with the availability of CD-ROMs storing the whole contents of magazines and quality newspapers. This is an important step as using a corpus for testing has been proven essential, in line with recent trends in natural language processing which shows increasing use of corpora as a source of linguistic knowledge (cf. issue of Computational Linguistics, 19 (1-2), 1993, devoted to corpus-based work).

One should stress that the grammar checker is primarily aimed at detecting non-grammatical linguistic constructions, and it is not concerned with the appropriate use of semantic items or with the accuracy of the information being conveyed. Therefore, meaningless clauses such as:

The mouse killed the elephant with its wing.

Shakespeare called in yesterday.

will still be considered entirely acceptable for they are grammatical constructions. It may also be emphasized that non-grammatical expressions originating from non-native speakers were not considered in developing the grammar checker. That is to say, the tool is unlikely to cope with unusual mistakes that may just be created for trying the tool.

Acknowledgements

This project is supported by Itautec/Philco S.A. Research grants from the Brazilian Agencies CNPq and Capes are also acknowledged. The authors are grateful to Benedita A. Chavedar, Álvaro Garcia Neto and Homero Schiabel for helpful comments and discussions, and to Lúcia R. Nascimento for her work with the corpus.

References

[Bidderman-93] Bidderman, M.T.C. Projeto lexicográfico sobre a língua portuguesa. *Estudos sobre Lexicografia*, Ano VII, N. 1. Araraquara/SP, Unesp, 1993, p. 33-52 (*In Portuguese*).

[Bolt-92] Bolt, P. An Evaluation of Grammar-checking programs as Self-help learning aids for learners of English as a Foreign Language, *Computer Assisted Language Learning*, 5, 1992, p. 49-92.

[Borba-93] Borba, F.S. Roteiro para a montagem de um dicionário de usos do português contemporâneo do Brasil (DUP). *Estudos sobre Lexicografia*, Ano VII, N. 1, Ararquara/SP, Unesp, 1993, p. 7-33. (*In Portuguese*).

[Cherry-82] Cherry, L. Writing tools. *IEEE Transactions on communications, COM-30*, 1982, p. 100-105.

[Flesch-48] Flesch, R.F. A new readability yardstick. *Journal of Applied Psychology, 32*, 1948, p. 221-233.

[Ide-93] Ide, N. and Véronis, J. Knowledge Extraction from Machine-Readable Dictionaries: An Evaluation, *Lecture Notes in Artificial Intelligence, Machine Translation and the Lexicon*, Springer, 3rd International EAMT Workshop, (Ed. P. Steffens), Heidelberg, Germany, April, 1993, p. 19-34.

[Itautec-93] Itautec Informática S/A - Grupo Itautec. *Redator Profissional: Guia do Revisor.* São Paulo, 1993. *(In Portuguese).*

[Itautec-95] Itautec Philco S/A. *Redator Windows.* São Paulo, 1995. *(In Portuguese).*

[Klare-74] Klare, G.R. Assessing readability. *Reading Research Quarterly.* N. 1, X/1, 1974-75, p. 62-102.

[Mayo-93] Mayo, D.J. Beware the jabberwock and other perils of readability formulae. *Transfusion*, 33(10), 1993, p. 884-5.

[Meade-91] Meade, C.D. and Smith, C. Readability formulas: cautions and criteria. *Patient Education and Counseling*, (17), 1991, p. 153-8.

[Miller-87] Miller, P.L. and Rennels, G.D. Prose Generation from Expert Systems, *AI Magazine*, Fall 1988, 37-44.

[Olson-86] Olson, A.V. A question of reading validity. *Journal of Research and Development in Education*, 19(4), 1986, p. 33-40.

[Procter-93] Procter, P. The Cambridge Language Survey, *Lecture Notes in Artificial Intelligence, Machine Translation and the Lexicon*, Springer, 3rd International EAMT Workshop, (Ed. P. Steffens), Heidelberg, Germany, April, 1993, p. 77-84.

[Royer-90] Royer, J.M. The readability dilemma. *Contemporary Psychology*, N. 1, V. 35, 1990, p.41-2.

[Rush-85] Rush, R.T. Assessing readability formulas and alternatives. *The Reading Teacher*, December 1985, p. 174-283.

[Sinclair-87] Sinclair, J.M. (editor) *et alii*. Looking up. An account of the COBUILD Project in lexical computing. London, Collins Cobuild, 1987.

[Suri-93] Suri, L.Z. and McCoy, K.F., Correcting Discourse-level Errors ina CALL System for Second Language Learners, *CALL*, 6(3), 1993, p. 215-231.

[Woods-70] Woods, W.A. Transition Network Grammars for Natural Language Analysis. CACM, Special Issue on *Computational Linguistics*, 13(2), October 1970, p. 591-606.

NOTAS DO ICMSC

SÉRIE COMPUTAÇÃO

024/96 FORTES, R.P.M. - Uma ferramenta orientada a links para avaliação de hiperdocumentos.

023/96 CANSIAN, A.M.; MOREIRA, E.S.; MAURO, R.B.; MORISHITA, F.T.; CARVALHO, A.C.P.L.F. - Um sistema adaptativo de detecção de intrusão em redes de computadores.

022/96 BRIGANTE, W.J.; MOREIRA, E.S. - Utilização de Monitores OLTP no gerenciamento de ambientes de manufatura integrada voltados à produção discreta.

021/95 BEZERRA, L.A.F.; SANTANA, R.H.C.; SANTANA, M.J. - Sistema auxiliar de arquivos baseado em disco WORM para ambientar computacional distribuído.

020/95 NUNES, M.G.V.; HASEGAWA, R. - PROTEMA: intelligent tutoring systems for mathematics.

019/95 OLIVEIRA, M.C.; TURINE, M.A.S.; MASIERO, P.C. - A statechart - based model for hypertext.

018/95 PIMENTEL, M.G.C. - Alternative operations for browsing hypertext.

017/94 ROMEIRO, N.M.L.; CASTELO FILHO, A. - Análise Comparativa de Métodos Numéricos de equações algebrico-diferenciais.

016/94 MAGALHÃES, A.L.C.C.; SIQUEIRA, M.F.; OLIVEIRA, M.C.F. - Operadores de Euler na modelagem por fronteira: conceito, aplicação, estudos de casos.

015/94 ODA, C.S.; MOREIRA, E.S. - ASNMP graphical network monitor with automatic topology discovery.