Tratamento Automatizado da Língua Natural

– rumo à correção automática? –

Profa. Dra. Vera Lúcia Strube de Lima
Instituto de Informática
Pontifícia Universidade Católica do Rio Grande do Sul

1. INTRODUÇÃO

No início dos anos 50 começaram a aparecer os primeiros trabalhos de pesquisa relativos ao *tratamento informatizado das línguas naturais**. Voltados eminentemente à tradução automática, os esforços engajados no tratamento das línguas naturais permitiram a análise de outras características da língua, como a sintaxe e a semântica.

Paralelamente, aparece uma outra linha de pesquisa no universo do tratamento das línguas naturais, a qual se populariza rapidamente, permitindo a inserção de seus resultados na maioria dos sistemas avançados de tratamento de textos e interfaces usuário. Trata-se da correção ortográfica [GREANIAS 84].

O desenvolvimento de ferramentas de tratamento de erros num texto se transforma em objeto de pesquisa em vista de algoritmos rápidos e eficientes, que possam ser acoplados a sistemas mais completos de tratamento e manipulação da língua natural.

Tenha-se em mente, contudo, que os primeiros passos neste sentido já datam do início do século: antes mesmos de viabializar-se a ferramenta *computador,* cientistas da língua já se propunham a trabalhar métodos e estruturas de dados favorecendo um tratamento automatizado da língua natural.

---

* Em inglês, Natural Language Processing (NLP) ou, em francês, Traitement Automatique des Langues Naturelles.

As ferramentas de tratamento de erros em um texto permitiram, numa primeira etapa, o tratamento de erros a nível léxico.

Mesmo sendo inicialmente voltados à "combinação" de cadeias de caracteres em geral (e não especificamente à correção ortográfica), alguns dos métodos desenvolvidos também puderam ser utilizados na correção de erros de ortografia em textos escritos em inglês, francês ou outras línguas [DAMERAU 64, WAGNER 74, POLLOCK 84, entre outros).

Certas línguas, porém, apresentam características particulares, como a acentuação, ou a variedade de possibilidades gráficas a partir de um modelo fonético, que demandam estudos específicos em vista do tratamento de erros (nesta situação se enquadram o francês e o português: *hássido* é uma grafia foneticamente equivalente a *ácido*).

Vários são os programas comercialmente disponíveis no que se refere ao tratamento de erros léxicos em textos escritos em língua natural. Contemplando os idiomas francês e inglês, para a linha de computadores PC, podemos citar WORD da MICROSOFT, MANUSCRIPT de LOTUS, entre outros.

No que se refere ao português, programas de tratamento de erros a nível léxico, inseridos ou não em ferramentas de tratamento de textos, começam igualmente a se popularizar. é o caso do ROT (Revisão Ortográfica de Textos), que aceita textos em vários formatos de edição, ou o FACIL, programa de edição de textos que inclui opções de revisão ortográfica.

Sem considerar-se o tempo de resposta (fator que seguidamente desencoraja o usuário a utilizar revisores automáticos de texto), podemos observar, primeiramente, que certas destas ferramentas ou não detectam todas as palavras erradas presentes no texto, ou assinalam como erradas palavras corretas.

Alguns tipos de erros, como os erros de natureza fonética, não são suficientemente tratados por estes programas (como é o caso do exemplo *hássido*).

Erros causados pela utilização de uma desinência incorreta (exemplo: *atora* em lugar de atriz) são raramente corrigidos.

Erros relacionados a palavras compostas são dificilmente detectados.

Quanto à correção de erros a nível sintático num texto escrito, constatamos que não há grande variedade de produtos comerciais disponíveis.

O mercado atual é caracterizado por ferramentas que dominam o universo da palavra, enquanto que ainda não pudemos vivenciar uma evolução significativa (no que se refere a ferramentas comercialmente disponíveis) a nível da estrutura da frase e de seu significado.

A correção de erros de sintaxe num texto exige a disponibilidade de informações a nível morfológico, sintático e mesmo semântico, que só são obtidas através de uma análise morfológica, sintática e semântica de amplitude considerável. Além disso, é normalmente necessário limitar o subconjunto de interesse no domínio da língua natural, a qual se apresenta ainda como um campo vasto demais para uma verificação/correção sintática exaustiva.

Sob o ponto-de-vista da pesquisa científica, vários estudos continuam em andamento, buscando a viabilização de um sistema completo de detecção e correção de erros em um texto, sejam estes erros a nível léxico, sintático ou semântico.

No caso do idioma português, é particularmente fundamental a retomada dos estudos que vêm sendo realizados para a língua francesa: estas duas línguas têm características bem semelhantes que permitem visualizar-se o aproveitamento dos resultados obtidos para o francês, fazendo uso dos mesmos no tratamento automatizado do português.

Atualmente, com o advento dos sistemas de 4ª e 5ª geração, que preconizam o desenvolvimento de interfaces homem-máquina em língua natural, fica fortificada a preocupação com o tratamento da língua natural.

A indexação automática, o estudo de novas metodologias para análise e processamento da língua natural, gramáticas, o tratamento de erros em textos escritos, o aprimoramento de técnicas para armazenamento e recuperação de dados em dicionários eletrônicos e a própria tradução automática são ramificações que garantem um interesse especial por esta fatia, hoje denominada tratamento automatizado da língua natural, que se insere tanto no nebuloso espectro da Inteligência Artificial como, de uma forma mais prática do que conceitual, no universo da Lingüística Computacional.

Este artigo visa a prover o leitor com uma visão "panorâmica" no que se refere ao tema *correção ortográfica automatizada*, apresentando um resumo das técnicas e métodos empregados atualmente no tratamento da língua natural, abordando suas vantagens, suas deficiências e sua transposição para o português. Os resultados aqui apresentados, bem como as expectativas atuais em rumo a uma correção orto-

gráfica totalmente (ou quase) automatizada, são provenientes de [STRUBE DE LIMA 90].

2. TIPOS DE ERROS E ESTRATÉGIAS DE CORREÇÃO

2.1 A classificação dos erros em um texto

Os estudos feitos no âmbito da classificação dos erros encontrados em um texto foram dirigidos, inicialmente, aos erros provocados pelo mau reconhecimento de caracteres óticos e aos erros tipográficos no código de programas. Hoje, a correção automática de uma cadeia de caracteres faz parte de inúmeras aplicações – a telemática, a automação de escritórios, a leitura ótica, a assistência à depuração de programas, a Biologia Molecular, a Genética, a criação de interfaces homem-máquina para sistemas de 4ª e 5ª geração.

Em meio a estas aplicações encontramos a correção de textos escritos em língua natural, sobre a qual centraremos as tipologias de erros a serem apresentadas a seguir.

Do ponto de vista lingüístico clássico, durante a análise de um texto podemos encontrar erros distribuídos em três níveis diferentes: o *nível léxico, o nível sintático* e o *nível semântico,* ou seja, respectivamente, erros a nível da palavra, da construção ou do sentido. Vários autores têm enunciado seus pontos-de-vista no que se refere à tipologia dos erros de ortografia. Alguns destes pontos-de-vista, particularmente interessantes em se tratando da verificação correção ortográfica, serão mencionados aqui.

O sistema DECOR (equipe TRILAN, laboratório LGI da Universidade Joseph Fourier de Grenoble) trata os erros sob o aspecto "o mais lingüístico possível", classificando-os segundo os algoritmos disponíveis para corrigí-los.

A nível léxico, podemos distinguir [COURTIN 87] os *erros de ortografia* (como a transposição das letras **e** e **r** em **aeroporto**), os *erros fonéticos* (como a substituição de **x** por **ss** em **mássimo**) e os *erros de geração* (como o uso de uma desinência incorreta no plural *calçãos*). Os *erros tipográficos* (como a transposição das letras **o** e **p** em **oprta**) são igualmente incluídos nesta categoria.

Os erros produzidos no interior dos elementos constituintes das palavras compostas também poderiam ser tratados a nível léxico.

A nível sintático, DECOR considera a correção no que diz respeito às regras de construção da frase e à concordância entre seus componentes. Por exemplo, a frase índio comer hambúrguer pode ser considerada como sintaticamente errônea, em vista do erro de concordância do verbo (*comer*) com o sujeito (*índio*).

A nível semântico, podemos conceber uma interpretação da frase, que permita sua verificação conforme seu significado. Esta verificação pode ser seguida de uma correção, se necessário. Por exemplo, o tratamento a nível semântico da expressão *concerto de bicicletas* deixa transparecer, certamente, um erro (para corrigí-lo podemos fazer referência a um conjunto de equivalentes fonéticos de **concerto,** que contém como alternativa a palavra *concerto*). Já numa expressão como *concerto de gaitas* dificilmente detectaríamos um erro semântico.

Catach, Duprez e Legis [CATACH 80] apresentam uma tipologia centrada sobre os problemas do ensino da ortografia, classificando os erros de ortografia em seis categorias:

I – erros com dominante fonético;
II – erros com dominante fonogramico;
III – erros com dominante morfogrâmico;
IV – erros relacionados a homófonos;
V – erros relacionados a ideogramas;
VI – erros relacionados a letras não funcionais.

Veronis [VERONIS 88a] usa os conceitos introduzidos por Chomsky (*Aspects of the theory of Syntax,* 1965) considerando a distinção entre os *erros de competência* e os *erros de performance:* enquanto que a competência é o conhecimento das estruturas da língua, a performance diz respeito ao emprego efetivo da língua em situações concretas. Os erros tipográficos, por exemplo, devidos a um toque errado sobre o teclado, são erros de performance, enquanto que os erros fono-gráficos (como **sentrau** por *central*) são erros de competência.

Como o indica Veronis, as noções de performance e de competência podem ser aplicadas aos dois interlocutores, o usuário e o sistema, alternando os papéis de emissor e receptor, mesmo que um deles seja apenas uma máquina. Assim podem ser induzidas quatro grandes classes de erros, que necessitam de estratégias de correção diferentes:

- erros de competência do usuário;
- erros de performance do usuário;
- erros de competência do sistema;
- erros de performance do sistema.

Veronis propõe a aplicação destas noções a todos os níveis – léxico, sintático ou semântico – de tratamento de erros no diálogo homem-máquina em língua natural. Seus trabalhos vêm sendo implantados no sistema ARCHIMEDE – um sistema especialista destinado ao ensino da Geometria Euclidiana Plana – pelo grupo RTC, Marseille.

O sistema VORTEX [LAHENS 86], desenvolvido no laboratório CERFIA da Universidade Paul Sabatier de Toulouse, grupa os erros introduzidos em um texto em duas categorias: erros *ortográficos* e erros *tipográficos*.

Os erros tipográficos são introduzidos durante o processo de "entrada" do texto e estão freqüentemente ligados ao uso do teclado de máquinas de escrever. O sistema VORTEX inclui igualmente neste nível os erros diversos cuja origem independe das dificuldades da ortografia: erros de reconhecimento nas leitoras óticas, erros de transmissão telemática, erros de codificação de dados nos bancos de dados textuais.

Os erros ortográficos compreendem os erros de uso, os erros de concordância e os erros situados no limite entre esses tipos. Os erros de uso são provenientes da falta de conformidade entre o escrito e o oral, e para traduzí-los são definidos grupos de letras que trazem problemas ortográficos (como ç/ss, x/cs, etc).

Os erros tipográficos tratados são os seguintes:

- omissão de uma letra;
- inserção de uma letra;
- substituição de uma letra por outra;
- transposição de duas letras consecutivas;
- interrupção de uma palavra ou inserção de um espaço;
- união de duas palavras consecutivas ou omissão do espaço.

2.2 Métodos de correção a nível léxico

A correção de uma cadeia de caracteres desconhecida no léxico e, portanto, considerada mal ortografada, consiste em propor

m substituição à mesma uma palavra ou uma cadeia de palavras que melhor a corrija, segundo o critério dado.

Esta relação entre a palavra errônea e as palavras que a corrigem leva a uma formalização do conceito de *equivalência* entre duas cadeias de caracteres.

Duas cadeias **a** e **b**, construídas sobre um mesmo alfabeto **A**, serão declaradas *equivalentes* se existir uma função de equivalência **f** tal que **f(a) = f(b)**, sendo **f** uma função que age somente sobre a *forma* das cadeias.

Seja **S** um conjunto qualquer de caracteres, eventualmente igual ou contido em **A**. Sejam **A*** e **S***, respectivamente, os conjuntos de todas as cadeias que podem ser construídas sobre **A** e **S**, inclusive a *cadeia vazia*. Mais formalmente **f** é uma função de **A*** sobre **S***. O problema se reduz então a discutir a escolha da função **f**, levando em conta a aplicação visualizada.

Em se tratando de erros a nível léxico, podemos distinguir dois tipos de métodos ou estratégias de correção. Os métodos *locais* ou *estatísticos* se caracterizam pelo emprego de regras estatísticas sobre a construção de cadeias de caracteres formando palavras. Os métodos *globais* (também denominados *combinatórios*) necessitam de um meio de comparação direto entre duas cadeias de caracteres, e permitem extrair do léxico um subconjunto de cadeias nas quais a distância com relação à cadeia de entrada é inferior a um limite dado.

Tratando-se dos métodos globais de correção, o problema geral é:

Seja um léxico **L**, onde cada palavra é uma cadeia de símbolos de um alfabeto **A**.

Seja **e** a cadeia de entrada.

Como extrair de **L** o conjunto de palavras **p** "próximas" de **e**?

As estratégias apresentadas a seguir são estratégias globais de correção de erros.

2.2.1 *Primeiros passos*

Em um artigo publicado em 1960, Charles Blair [BLAIR 60] afirma que, se propusermos à máquina um critério adequado para calcular a "similaridade" entre duas palavras, ela poderá "corrigir" um erro de ortografia substituindo a palavra errônea pelas palavras corre-

tas "mais similares" à primeira. Blair estabeleceu seu próprio critério de similaridade, através de abreviações associadas a cada palavra do dicionário. Seu sistema, mesmo sendo bastante simplificado, serve de inspiração, ainda hoje, a vários outros.

Para corrigir uma palavra errônea, é calculada sua abreviação c, e são buscadas todas as palavras do dicionário possuindo esta mesma abreviação. Se, por coincidência, várias palavras do dicionário correspondem a uma mesma abreviação, são calculadas abreviações mais longas para a palavra errônea e para as entradas do dicionário.

Damerau [DAMERAU 64] apresentou um dos primeiros estudos sobre a correção de erros tipográficos em um texto. Analisando a natureza dos erros de ortografia encontrados, observou que 80% das palavras errôneas apresentavam um único erro, que podia ser a substituição de uma letra por outra, a omissão de uma letra, a inserção de uma letra ou a transposição de duas letras (ou seja, 80% das palavras errôneas apresentavam um único erro de edição).

Damerau propõe um método de correção para este gênero de erro: uma palavra mal ortografada é corrigida a partir de repetidas buscas no dicionário, supondo a cada vez a existência de um dos tipos de erro de omissão, de inserção, de substituição ou de transposição, tentando eliminá-los.

No domínio da similaridade, técnicas mais abstratadas têm sido apresentadas, como é o caso da *distância de Levenshtein*, proposta por seu autor no periódico *Soviet Physics Doklady* nº 8 de 1966. Esta distância é definida como o custo mínimo necessário para transformar uma cadeia de caracteres em outra, com a ajuda de operações de inserção, omissão, substituição ou transposição.

Wagner e Fisher [WAGNER 74] desenvolveram um algoritmo para calcular a distância de Levenshtein no caso em que se excluem as transposições (o algoritmo dado usa programação dinâmica).

Lowrance e Wagner [LOWRANCE 75] propuseram em seguida uma extensão ao algoritmo anterior, a qual leva em conta a transposição entre dois caracteres adjacentes.

Quando a similaridade deve ser calculada sobre um dicionário de dimensão real e em tempo real, os algoritmos de Wagner e Fisher e de Lowrance e Wagner apresentam problemas de custo. Tais problemas foram analisados por Owolabi e McGregor [OWOLAB|88] levando à proposição de uma processo em duas etapas, para cálculo da similaridade entre duas cadeias de caracteres. A primeira etapa

usa uma tabela de n-gramas compacta para selecionar um conjunto de cadeias "grosseiramente similares". A segunda etapa compara estas últimas com a cadeia de entrada. Para realizar a comparação, os autores definiram uma nova medida de similaridade, baseada, como as precedentes, na métrica de Levenshtein.

2.2.2 *Chaves de similaridade*

O cálculo de *chaves de similaridade* tem por objetivo produzir, a partir de uma cadeia **a** dada, uma outra cadeia **c**, denominada chave, que concentre as principais características de **a**.

A chave **c** é obtida de **a** através da aplicação de regras de transformação.

Por exemplo, a seqüência **cmpnoal** pode ser uma chave associada à cadeia **compania**.

Neste caso, duas regras de transformação foram aplicadas à cadeia inicial:

1 – tomar as consoantes, na ordem em que aparecem;

2 – concatenar à cadeia obtida as vogais, na ordem em que aparecem, sem repetição.

A técnica das chaves de similaridade é usada para compensar as variações na forma das palavras. Uma das duas cadeias, **a**, sofre um certo número de deformações e é a forma **b**, marcada por erros, que é conhecida. **b** é comparada a uma forma de referência, proveniente de um léxico, e realiza-se a identificação de **a** pela forma de referência equivalente a **b**.

Os resultados publicados por Riseman e Hanson [RISEMAN74] contêm conclusões importantes no que se refere a chaves de similaridade. Os autores enunciam duas propriedades importantes de uma cadeia: a identidade e a *inter-relação entre as letras que a compõem.*

Estas duas propriedades são antagônicas: a chave deve ser suficientemente discriminante, para bem preservar a identidade da cadeia, ao mesmo tempo que deve ser insensível aos erros que se puderam produzir. Quanto mais informações a chave retem, maior é seu poder discriminante, porém maior é sua sensibilidade aos erros.

Pollock e Zamora [POLLOCK 64] desenvolveram um algoritmo de correção de erros de ortografia através do qual, para cada palavra contida no dicionário, é calculada uma chave de similaridade. As

palavras do dicionário são então classificadas na ordem das chaves calculadas. O algoritmo de correção, cuja implementação faz parte do projeto SPEEDCOP (*National Science Foundation* dos Estados Unidos), calcula uma chave associada à palavra mal ortografada, e busca no dicionário as palavras cujas chaves são próximas da chave calculada. Deste conjunto de palavras, uma ou mais alternativas são selecionadas como correção. Estes autores propuseram uma chave específica, denominada "chave-esqueleto" ou *skeleton-key*, calculada tomando-se a primeira letra da palavra e concatenando a esta letra as consoantes da palavra (em ordem de ocorrência, sem repetição), e depois suas vogais (também em ordem de ocorrência, sem repetição).

O aspecto mais discutível desta chave é a importância delegada às consoantes iniciais da palavra. Quanto mais próxima do início da palavra está uma consoante incorreta, mais distante estarão, no dicionário, a chave da palavra incorreta e a chave da correção, o que pode tornar impraticável a correção através deste método.

Assim mesmo, o método da chave-esqueleto é incluído na maior parte dos sistemas de tratamento de erros ortográficos, inclusive naqueles que manipulam textos escritos em português.

2.2.3 Tratamento de erros fonéticos

Mesmo sendo atualmente pouco enfatizados pelas ferramentas de correção, os erros de natureza fonética ou fonogrâmica representam um segmento considerável dos erros de ortografia em textos escritos em francês ou português.

Entretanto, a correção dos erros fonéticos ou fonogrâmicos demanda algoritmos específicos, que permitam a recuperação do maior número possível de equivalentes fonéticos associados a uma certa palavra incorreta. O conjunto completo destes equivalentes só pode ser obtido através de um processo de transdução fonética, que visa produzir todas as cadeias orais associadas a uma certa cadeia escrita, seguidas as regras e modelos fonéticos referentes a um certo idioma. O sistema DECOR, para o francês, provê este gênero de solução.

Em vista, porém, da complexidade do processo de tratamento de erros fonéticos por transdução, a correção de erros fonéticos se limita, na maior parte dos sistemas, a uma simplificação relacionando em tabela grupos de letras que costumam trazer problemas ortográficos (assim ocorre nos sistemas ARCHIMEDE, VORTEX e outros).

2.2.4 Erros de geração

Os erros ditos de geração podem ser corrigidos segundo um princípio de análise morfológica seguida de uma reconstrução. Uma análise morfológica reconhece a raiz r da palavra incorreta, e supõe-se que esta raiz é correta. É reconhecida em seguida a terminação da palavra incorreta, através da qual é obtido o conjunto de valores das variáveis morfológicas "desejadas" quando de sua escrita. Estes valores permitem calcular-se a palavra correta, a partir da raiz r.

Por exemplo, consideremos a correção da palavra **feijãos**. A base *feijão* é correta e a terminação s permite detectar-se um plural. A base *feijão* é flexionada, então, no plural, obtendo-se *feijões*.

2.3 Erros sintáticos e seu tratamento

Os erros sintáticos em um texto constituem um problema mais delicado. Seu tratamento pressupõe um tratamento morfológico anterior, ou seja, o verificador/corretor sintático deve dispor de um conjunto robusto e consistente de informações lingüísticas a respeito de cada palavra. Seu tratamento supõe freqüentemente um tratamento semântico: é comum chegar-se a situações de impasse numa fronteira imaginária entre Sintaxe e Semântica, onde os traços morfo-sintáticos são às vezes insuficientes para realizarmos uma verificação/correção sintática precisa.

A verificação sintática em um texto concerne normalmente à aplicabilidade de um conjunto de regras que modelam as relações mais fundamentais entre os componentes de uma proposição escrita. Entre estas relações, encontram-se a ligação entre o nome e seu determinante, o nome e seu adjetivo, o sujeito e o verbo, o verbo e seu atributo, bem como a conjunção de coordenação e as conjunções de subordinação.

Certos autores fazem algumas observações no que diz respeito à cobertura de uma análise sintática, nos interfaces em língua natural. Carbonell e Hayes [CARBONELL 84] constatam que, numa aplica-

ção como o tratamento de textos, os erros aqui ditos "sintáticos" são pouco freqüentes: os textos são cuidadosamente preparados e editados, o que elimina a maioria dos erros de sintaxe. O mesmo não é verdadeiro para os sistemas que aceitam uma linguagem produzida espontaneamente pelo usuário, os quais necessitarão uma análise sintático-semântica muito mais robusta.

Segundo Emirkanian e Bouchard [EMIRKANIAN 88a, EMIRKANIAN 88b], os erros de sintaxe são pouco freqüentes em francês. Numa amostragem de 6580 frases retiradas de textos escritos por crianças em escola primária, somente 1,2% foram consideradas frases sintaticamente incorretas.

Quanto à língua portuguesa, esta autora desconhece, até o presente, dados estatísticos a propósito de erros de sintaxe.

Várias equipes de pesquisa trabalham, atualmente, na verificação/correção de erros sintáticos em textos escritos em francês ou inglês, utilizando informações morfo-sintáticas ou, por vezes, informações de caráter sintático-semântico.

De nossa parte, consideramos que o uso de informações semânticas no tratamento dos erros de sintaxe, para um domínio genérico da língua, é uma tarefa que apresentará, ainda por muito tempo, dificuldades.

Seguramente, nem todos os erros de sintaxe podem ser corrigidos sem levarmos em conta informações semânticas. Muitos deles nem poderão ser detectados sem um tratamento semântico... (exemplo: *os cães da vizinha que* **latia** *sem cessar...*).

Por outro lado, a não utilização de informações semânticas permite a verificação/correção sintática de um conjunto muito mais amplo de construções (numa primeira etapa de estudos, fica possível o tratamento de frases como **ela engolia as palavras ao falar...** atribuindo às mesmas, normalmente, estruturas de dependências).

2.4 Erros semânticos

Os erros detectados a nível semântico são erros de competência [VERONIS 88b], provenientes do sistema ou do usuário.

No caso dos erros provenientes do sistema, este último possui uma representação contraditória ou incompleta dos conhecimentos. Os erros provenientes do usuário podem ser classificados em dois tipos: erros conceituais e erros pragmáticos.

Os erros conceituais provêm de uma representação errônea dos conhecimentos, na concepção do usuário. Estes erros aparecem particularmente no domínio da Ensino Assistido por Computador, caso este em que o usuário se encontra em fase de aprendizado do assunto. Por exemplo, a frase **[AB] é a hipotenusa do círculo C** reflete uma representação errônea do domínio pelo usuário (a expressão *hipotenusa de X* pressupõe que o objeto *x* seja um triângulo retângulo).

Os erros pragmáticos são provenientes do emprego inapropriado, numa determinada situação, de uma frase ou de uma expressão que, por si só, não é errônea. Tais erros consistem geralmente de uma violação das leis do discurso, ou de uma violação de um procedimento, de um cenário, de um script, próprios ao domínio considerado. Por exemplo: **Quando meu pai era pequeno, eu trabalhava na fábrica de calçados.**

A detecção e a correção de erros a nível semântico permanecem como um vasto campo de pesquisa do qual ainda dispomos de pouquíssimos resultados. O que existe até o presente é o tratamento semântico de subconjuntos da língua, inseridos em aplicações específicas, podendo-se visualizar, para um futuro próximo, o aproveitamento de alguns conhecimentos semânticos buscando aprimorar os mecanismos de detecção e correção de erros.

3. CONCLUSÃO

Nesse artigo, efetuamos uma rápida reflexão no que se refere a métodos de tratamento de erros a nível léxico, sintático e semântico, em textos escritos em língua natural.

Todas as linguagens são geridas por um conjunto de regras de formulação de proposições. Mesmo assim, no universo das línguas naturais, a definição de gramáticas sensíveis aos fenômenos observados ainda não foi totalmente alcançada. Os estudos existentes se limitam aos fenômenos considerados como "os mais importantes", seja na falta considerada como os *fundamentos* da língua, seja em sub-conjuntos bem limitados da língua, utilizados em um domínio específico de aplicação.

Pode-se até mesmo questionar a exeqüibilidade de definição de uma gramática exaustiva, para as línguas naturais.

A perspectiva de chegar-se a um sistema completo de tratamento de erros, o qual possa ser inserido tanto em ferramentas de tratamento de textos como em interfaces genéricos de comunicação homem-máquina, ainda se apresenta como um longo caminho a percorrer. Este caminho, encurtado ou não pelo aproveitamento de estudos efetuados para outros idiomas, há de ser percorrido igualmente por nós, usuários do idioma português.

Nossa língua, empregada hoje por mais de 150 milhões de pessoas distribuídas em três continentes, apresenta ainda inúmeros temas de pesquisas a contemplar.

O processo de evolução tecnológica intensa com o qual convivemos hoje exige resposta de outras ciências, e não só da própria informática, para que se possa fazer uso eficiente de todo o instrumental que nos vem sendo oferecido. As interfaces de comunicação com a máquina, humanizando o diálogo até bem pouco tempo rígido e esquematizado, parecem ser um dos pontos de partida.

4. REFERÊNCIAS BIBLIOGRÁFICAS

[BLAIR 60]
BLAIR C. *A program for correcting errors.* Information and Control nº 3, 1960. p.60-67.

[CARBONELL 84]
CARBONELL, J., HAYES P. *Recovery strategies for parsing extragrammatical language.* American Journal of Computational Linguistics. Vol. 9 nº 3-4, Jul.-Dez. 1983. p. 123-146.

[CATACH 80]
CATACH N., DUPREZ D., LEGRIS M. *L'enseignement de l'orthographe.* Dossiers Didactiques NATHAN, Luçon-França, 1980. 96p.

[COURTIN 87]
COURTIN J., DUJARDIN D., KOWARSKII., STRUBE DE LIMA V. *Détection et correction des erreurs.* Rapport PRC pôle Langage Naturel, Paris-França, Set.-Dez. 19877.

[COURTIN 88]
COURTIN J., DUJARDIN D., KOWARSKI I., GENTHIAL D., STRUBE DE LIMA V. *Correção de erros de ortografia através da fonética em textos escritos em francês.* XIV Conferência Latinoamericana de Informática, 17avas Jornadas Argentinas de Informática e Investigación operativa, Buenos Aires-Argentina, Set. 1988.

[COURTIN 89a]
COURTIN J., DUJARDIN D., KOWARSKI I., GENTHIAL D., STRUBE DE LIMA V. *Vers un système complet de détection/correction d'erreurs en français.* Rapport interne, Equipe TRILAN, LGI-IMAG, Grenoble, Mar. 1989.

[COURTIN 89b]
COURTIN J., DUJARDIN D., KOWARSKI I., GENTHIAL D., STRUBE DE LIMA V. *Interactive multi-level systems for correction of ill-formed french texts.* Second Scandinavian Conference on Artificial Intelligence, Tempere-Finlândia, Jun. 1989. p. 912-920.

[COURTIN 89c]
COURTIN J., DUJARDIN D., KOWARSKI I., GENTHIAL D., STRUBE DE LIMA V. *DECOR – detecção e correção léxico-sintática num texto escrito.* XV Conferencia Latinoamericana de Informática, IX Conferência Internacional Chilena de Ciência de la Computación, Santiago do Chile, Jul. 1989. p. 67-76.

[COURTIN 89d]
COURTIN J., DUJARDIN D., KOWARSKI I., GENTHIAL D., STRUBE DE LIMA V. *Análise de textos escritos em português com PILAF – uma experiência e seus resultados.* 18avas Jornadas Argentinas de Informática e Investigación Operativa, Buenos Aires, Ago. 1989. p. 9.29-9.46.

[DAMERAU 64]
DAMERAU F.J. *A technique for computer detection and correction of spelling errors.* Communications of the ACM, Vol. 7 nº 3, Mar. 1964. p. 171-176.

[EMIRKANIAN 88a]
EMIRKANIAN L., BOUCHARD L. *Towards a knowledge-based tool for correcting french text.* IFIP European Conference on Computer in Education, Lausanne-Suiça, Jul. 1988. p. 583-588.

[EMIRKANIAN 88b]
EMIRKANIAN L., BOUCHARD L. *Knowledge integration in a robust and efficient morphosyntactic analyser for french.* 12th International Conference on Computacional Linguistics, Budapest-Hungria, Ago. 1988. p. 166-171.

[GREANIAS 84]
GREANIAS E. C. *Computer aids for spelling correction and word selection.* IBM Europe Institute 1984, Davos – Suiça, Jul.-Ago. 1984.

[LAHENS 86]
LAHENS F. *Un modèle stochastique pour la vérification et la correction automatique de textes: le système VORTEX.* Tolouse-França, Nov. 1986 (Tese de Doutorado). 166p.

[LOWRANCE 75]
LOWRANCE R., WAGNER R. *An extension of the string-to-string correction problem.* Journal of the ACM, Vol. 22 nº 2, Abr. 1975. p. 177-183.

[OWOLABI 88]
OWOLABI O., MCGREGOR D. *Fast aproximate string matching.* SOFTWARE – practice and experience, Vol 18 nº 4, Abr. 1988. p. 387-393.

[POLLOCK 84]
POLLOCK J., ZAMORA A. *Automatic spelling correction in scientific and scholarly text.* Communications of the ACM, Vol. 27 nº 4, Abr. 1974.

[RISEMAN 74]
RISEMAN E., HANSON A. *A contextual postprocessing system for error correction using binary n-grams.* IEEE Transactions on Computers, Vol. C-23 nº 5, Mai. 1974. p. 480-493.

[STRUBE DE LIMA 90]
STRUBE DE LIMA, V.L. *Contribution à l'étude du traitement des erreurs au niveau lexico-syntaxique dans un texte écrit en français.* Grenoble-França, Université Joseph Fourier, Mar. 1990 (Tese de Doutorado). 272p.

[VERONIS 88a]
VERONIS J. *Morphosyntactic correction in natural language interfaces.* 12th International Conference on Research & Development in Information Retrieval, Jun. 1988, Grenoble-França.
[VERONIS 88b]
VERONIS J. *Contribution à l'étude de l'erreur dans le dialogue homme-machine en longage naturel.* Marseille-França, Out. 1988 (Tese de Doutorado). 325p.
[WAGNER 74]
WAGNER C.; FISHER M. *The string-to string correction problem.* Journal of the ACM, Vol. 21 nº 1, 1974. p. 168-173.

Geração Traída
Jane Tutikian
74 p. - Série Novelas Exemplares
(Prefácio de Guilhermino César)

Geração Traída, de Jane Tutikian, é o quarto título da Série Novelas Exemplares que a Editora Mercado Aberto está lançando na oportunidade da Feira do Livro de Porto Alegre.

Ficcionista gaúcha bastante conhecida, Jane Tutikian recebeu o prêmio em 1984, na categoria novela, além de ter publicado livros de contos. Em **Geração Traída** a autora retrata o desencanto e a desorientação de uma geração que, tendo-se formado numa época em que os valores sociais estavam mudando rapidamente, encontrou e encontra grandes dificuldades em organizar-se diante do mundo, afundando no pessimismo, que jamais alcança o plano da revolta ou da contestação. Para Guilhermino Cesar, se **Geração Traída** é o produto típico de uma época de crise como a presente, reflete por outro lado, a esperança dos que, mesmo na solidão, conseguem inventar um caminho.