

XIV Conferencia Latinoamericana de Informática
17avas. Jornadas Argentinas de Informática
e Investigación Operativa
Buenos Aires, Setiembre 1988.

CORREÇÃO DE ERROS DE ORTOGRAFIA

ATRAVÉS DA FONÉTICA

em textos escritos em Francês

por

COURTIN, Jacques, KOWARSKI, Irène, DUJARDIN, Danielle,
GENTHIAL, Damien, DE LIMA, Vera Lúcia Strube

Laboratoire de Génie Informatique
IMAG - Campus
38041 Grenoble cedex FRANCE

1 - INTRODUÇÃO

Motivação

Torna-se cada vez mais importante, para um sistema que realiza o tratamento de textos escritos em língua natural, a disponibilidade de uma análise eficaz a nível léxico, sintático e mesmo semântico. A detecção e a correção de erros de Ortografia aparecem aí como problemas fundamentais, uma vez que permitem passar à verificação da estrutura sintática e ao exame do texto buscando seu significado.

O problema da detecção e correção de erros de Ortografia tem sido amplamente analisado, tanto no que tange às grafias contendo um único erro de omissão, inserção, substituição ou transposição, como proposto por Damerau [DAM 64], como no que diz respeito às grafias contendo vários erros, segundo os estudos baseados na distância de Levenshtein [LEV 66].

Até o momento, todavia, os sistemas vêm propiciando uma correção normalmente fundada sobre uma única técnica, e especialmente para palavras contendo apenas um erro ortográfico. Raramente é proposta uma análise a nível sintático ou semântico.

Procuramos elaborar um sistema combinando várias técnicas de detecção/correção, com possibilidades, inclusive, de detecção e correção a nível sintático e semântico, para uma aplicação específica.

Atualmente, no sistema PILAF, já implementamos técnicas de detecção/correção a nível de palavra, bem como um verificador de regras de sintaxe a partir das estruturas de dependência associadas a uma frase. Os trabalhos em andamento têm por meta a detecção/correção de erros a nível sintático e semântico, propondo diferentes arquiteturas de detecção/correção para uma aplicação.

Objetivos

Neste artigo, descrevemos o método de correção de erros de Ortografia através da Fonética concebido e implementado no sistema PILAF.

O método aqui apresentado foi proposto por Courtin [COU 77] e compreende a tradução fonética de uma palavra mal ortografada, seguida da reconstituição de sua correspondente correta, utilizando estruturas de dados disponíveis a nível léxico. Um método de correção visando a composição fonética da palavra foi sugerido, também, por Peterson [PET 80], para a língua inglesa. Os componentes fonéticos, podendo se relacionar a diferentes cadeias gráficas, permitiriam calcular grafias alternativas para a palavra mal ortografada. Para a língua francesa, a alternativa de

substituição de cadeias gráficas já foi analisada em trabalhos como o de Maret [MAR 87].

Com a finalidade de proporcionar uma compreensão global do assunto, apresentamos aqui uma descrição geral do sistema PILAF - parte Morfologia e parte Sintaxe - inclusive os métodos de correção de erros de Ortografia disponíveis no sistema.

Em seguida, a correção de erros de Ortografia através da Fonética é detalhada em todas as suas etapas, desde a tradução fonética da palavra errada, até o enunciado de grafias alternativas.

Como conclusão, são abordados aspectos considerados importantes na proposição de arquiteturas de correção de erros de Ortografia, bem como a continuação dos trabalhos no sistema PILAF.

2 - PILAF

O sistema

O sistema PILAF (Procédures Interactives Linguistiques Appliquées au Français) é um software de tratamento interativo da língua natural. Analisando uma frase ou um texto, a partir da construção de estruturas de dependência e da pesquisa num dicionário, PILAF efetua a detecção de erros em dois níveis - Morfologia e Sintaxe. Os erros detectados podem ser corrigidos por diferentes mecanismos encontrados no sistema.

A análise Morfológica PILAF trata os erros detectados a nível de palavra, e propõe um conjunto de possibilidades corretas que podem substituir a palavra errada. A proposição de correção é feita utilizando os diferentes métodos de correção disponíveis no sistema PILAF.

O módulo de análise sintática detecta e corrige os erros a nível da construção da frase: os erros de linguagem.

Os níveis de correção

Consideremos, por exemplo, o tratamento da seguinte frase escrita em francês:

Les jirraphes mange.

A nível de palavra, PILAF detecta e corrige jirraphes, substituindo-a por girafes. A frase original, então, toma a forma:

Les girafes mange.

A nova versão da frase, após um tratamento a nível de palavra, ainda não está correta. O sujeito (no plural) não concorda em número com o predicado (no singular).

É a análise sintática, então, que se ocupa da detecção/correção desse segundo tipo de erro, propondo alternativas para o acordo sujeito - predicado, como

Les girafes mangent.

ou

La girafe mange.

Os métodos de correção de erros de Ortografia

Os métodos disponíveis no sistema PILAF, para a correção de erros de Ortografia, são quatro:

- a Geração pela Morfologia;
- o método da "clé squelette";
- o dicionário personalizado;
- a correção através da Fonética;

A decisão pela aplicação de um determinado método depende dos tipos de erros encontrados no texto. O algoritmo de aplicação dos métodos de correção depende da arquitetura de correção utilizada, e esta arquitetura é definida analisando-se um conjunto significativo de erros encontrados no texto a ser tratado.

Os métodos de correção de erros de Ortografia utilizados no sistema PILAF serão descritos brevemente, mais adiante, neste trabalho.

3 - O TRANSDUTOR DE ESTADOS FINITOS NO SISTEMA PILAF

Uma descrição geral do transdutor utilizado no sistema PILAF

PILAF utiliza o transdutor proposto e descrito por Courtin [COU 77].

Este transdutor é bastante geral. Ele permitiu, no início, a realização da Análise Morfológica PILAF e, a seguir, outras aplicações tais como a Geração pela Morfologia, a Tradução Fonética etc.

Tomemos, por exemplo, a primeira aplicação do transdutor: a Morfologia.

O objetivo da análise morfológica, utilizando o transdutor, é de segmentar uma frase para obter cadeias de componentes, e determinar, ao mesmo tempo, um certo número de informações lingüísticas sobre as cadeias resultantes da segmentação.

O princípio da segmentação é de determinar as palavras ou grupos de palavras componentes da frase, denominados segmentos. A determinação de um segmento consiste, então, de sua decomposição em prefixo, base, sufixo e desinência.

Para proceder à segmentação, dispomos:

- de um dicionário de prefixos, bases, sufixos e desinências, que permite, por identificação, a decomposição de uma parte contígua da cadeia de entrada, e

- de uma gramática que deve infirmar ou confirmar a concatenação destes elementos na constituição do segmento procurado.

As informações lingüísticas são associadas a modelos lingüísticos. Os componentes do dicionário de prefixos, bases, sufixos e desinências são lingüisticamente descritos através de modelos.

Para cada segmento determinado, é preciso fornecer as informações lingüísticas que serão utilizadas pelo modelo seguinte. Trata-se de fornecer todas as interpretações de todas as decomposições de um segmento.

Para certas aplicações, como a Tradução Fonética, é necessário produzir uma nova cadeia, à partir da cadeia de entrada. A cadeia resultante é produzida pela concatenação de códigos associados aos modelos aplicados durante o processo de tradução.

Os códigos associados aos modelos são denominados códigos de derivação, e só existem no caso das aplicações visando a derivação de uma nova cadeia a partir da cadeia de entrada.

O mesmo transdutor aqui descrito se adapta não só à Tradução Fonética, como também à fase de reconstituição gráfica, no método de correção de erros de Ortografia pela Fonética, empregando-se aí um dicionário e um conjunto de modelos e regras apropriados. As diferentes estruturas de dados utilizadas pelo transdutor nestas aplicações serão descritas quando do detalhamento do método de correção pela Fonética.

4 - MÉTODOS DE CORREÇÃO A NÍVEL DE PALAVRA

A Geração pela Morfologia

Este método consiste em encontrar uma base morfológica dentro da palavra mal ortografada, e calcular todas as formas possíveis para

esta base, a partir dos modelos e regras apropriados.

Por exemplo, seja **bretone** uma grafia incorreta detectada por PILAF. O método da Geração encontra, nesta palavra, a base **breton**. A partir desta base, são geradas as grafias **breton**, **bretons**, **bretonne** e **bretonnes**. Do conjunto das palavras geradas, a opção **bretonne** é filtrada como a mais adequada.

O método da "clé squelette"

Este método é baseado na construção de uma chave de pesquisa, a partir de uma palavra.

A chave proposta é calculada tomando certas das letras que formam a palavra, segundo critérios bem determinados. A primeira letra é guardada; em seguida figuram todas as consoantes, na mesma ordem em que aparecem na palavra, sem repetição. As vogais são então anexadas à chave, na mesma ordem em que aparecem na palavra, sem repetição.

O método consiste em obter a "clé squelette" para uma palavra mal ortografada, e então recuperar todas as palavras existentes no dicionário que se caracterizem por possuírem esta mesma "clé squelette", ou, eventualmente, palavras possuindo uma "clé squelette" bastante próxima da chave calculada.

Tomemos como exemplo a grafia **aréoport**. Calculando a chave proposta, obteremos a cadeia **arpteo**.

As variantes mal ortografadas da palavra **aéroport** se caracterizam pela mesma "clé squelette", que é também a chave obtida para a palavra **aréoport**. Fazendo-se uma pesquisa no dicionário, as palavras possuindo a mesma chave ou chaves vizinhas são recuperadas, e entre elas se encontra a grafia correta.

O dicionário personalizado

Cada usuário de um sistema de tratamento de textos pode apresentar suas próprias características com relação aos problemas de datilografia e de ortografia.

O dicionário personalizado é uma técnica que permite memorisar, para um dado usuário ou uma dada aplicação, os erros mais habituais, e suas correções adotadas no passado.

A correção pela Fonética

Um usuário que conhece pouco de uma língua, no que diz respeito à

Ortografia, mas que conhece razoavelmente bem a língua falada, pode ter dificuldades com relação à escrita, propondo uma grafia errônea para uma palavra que sabe pronunciar corretamente.

O modo de correção através da Fonética se aplica a palavras contendo um ou mais erros de Ortografia, desde que preservada sua boa forma oral. Por exemplo, seja *eaukureuse* uma forma errônea de escrever, em francês, a palavra *occurrence*: ambas as grafias podem ser pronunciadas da mesma maneira.

O algoritmo de correção através da Fonética calcula todas as cadeias orais possíveis para uma palavra mal ortografada. Utilizando estas cadeias orais, reconstituímos as grafias, usando as bases e as flexões armazenadas no sistema PILAF. As grafias produzidas são verificadas quanto a sua correção, e somente aquelas que constituem palavras da língua francesa são propostas como resultado.

5 - A CORREÇÃO DE ERROS DE ORTOGRAFIA ATRAVÉS DA FONÉTICA

A arquitetura geral de correção

No sistema PILAF, foi concebido um algoritmo para resolver, em três níveis, o problema da correção de erros de Ortografia através da Fonética. Este algoritmo é mostrado na Figura 1.

Num primeiro nível, é efetuada uma tradução fonética, a qual tem por objetivo calcular todas as cadeias orais possíveis para pronunciar a palavra mal ortografada, segundo os princípios da Linguística Francesa.

Num segundo nível, cada cadeia oral é analisada com relação às grafias que permite obter. São então propostas grafias alternativas para a palavra mal ortografada.

A reconstituição gráfica para uma cadeia oral pode produzir uma ou várias cadeias gráficas, que são então verificadas quanto a sua correção, no terceiro nível do algoritmo.

Todos os três níveis que compõem a técnica de correção através da Fonética fazem uso do autômato concebido para a Morfologia PILAF, preservado seu princípio de funcionamento, porém empregando estruturas de dados específicas.

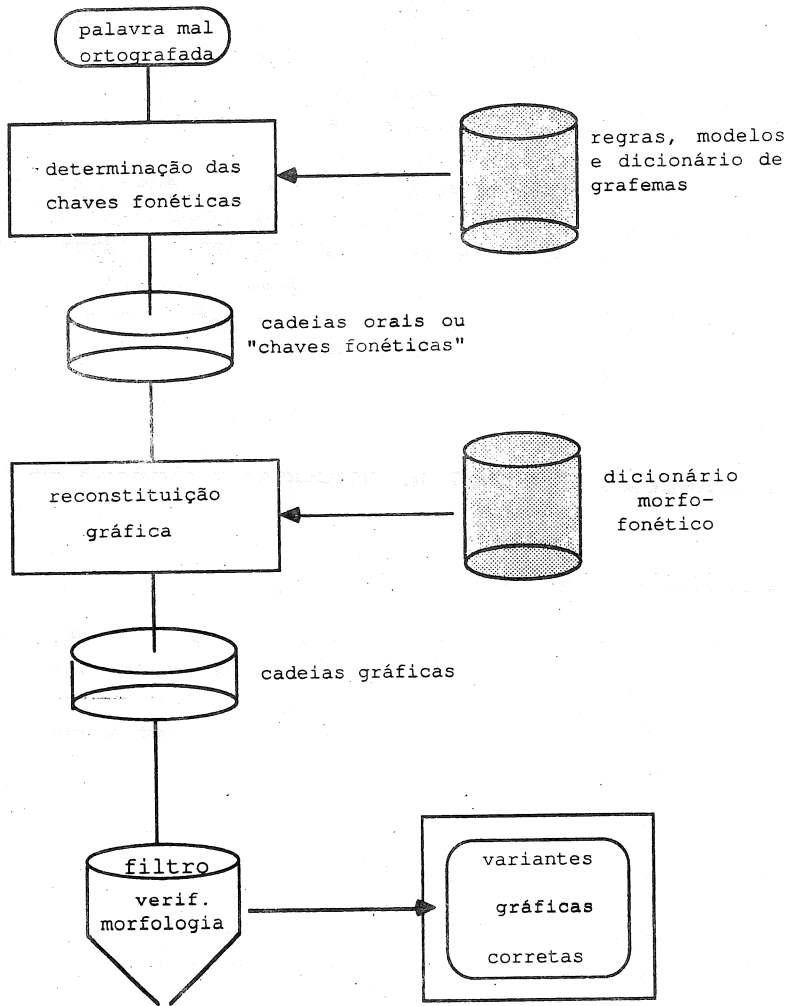


FIGURA 1 - ARQUITETURA GERAL DE CORREÇÃO ATRAVÉS DA FONÉTICA

6 - A CADEIA ESCRITA E A CADEIA ORAL

Grafemas e Fonemas

Um grafema pode ser caracterizado como a menor unidade distintiva ou significativa da cadeia escrita [CAT 80], composto de uma letra ou de um grupo de letras ou sinais auxiliares, possuindo uma referência fônica na cadeia oral.

O conceito de fonema encontra-se mais propriamente ligado ao oral: um fonema é a menor unidade distintiva da cadeia oral.

Considerando-se a correção de erros de Ortografia através da Fonética, o conceito de grafema foi estendido para permitir cadeias escritas mais longas, tratadas aqui como pseudo-grafemas ou, simplesmente, p-grafemas. Por exemplo, utilizamos a cadeia "enou" (concatenação de vários grafemas elementares) como um p-grafema.

Um conceito de pseudo-fonema foi também criado, a fim de permitir cadeias orais mais longas.

Os pseudo-fonemas ou, simplesmente, p-fonemas, são sempre constituídos ou por fonemas elementares, ou por cadeias de fonemas elementares.

A relação p-fonema $\rightarrow$ p-grafema é do tipo 1 : n, $n \geq 1$ e inteiro. Por exemplo, seja o p-fonema "s." definido no sistema PILAF. Ao p-fonema "s." estão associados os p-grafemas "s", "ss", "c", "ç" etc.

A relação p-grafema $\rightarrow$ p-fonema é do tipo 1 : m, $m \geq 1$ e inteiro. Por exemplo, seja o p-grafema "ch" definido no sistema PILAF. Ao p-grafema "ch" são associados os p-fonemas "ch." e "k".

Escolha dos fonemas ligados à correção de erros de Ortografia

A geração de cadeias orais, quando aplicada à correção de erros de Ortografia, obriga a uma formalização e, até mesmo, uma simplificação, com relação a todas as possibilidades de cadeias orais para uma cadeia escrita, na língua francesa.

O problema dos grafemas ligados ao oral, na língua francesa, já foi abordado por vários lingüistas como Catach, Lallias, Demeyère etc. Observando as proposições feitas por estes autores, propusemos um conjunto de 33 fonemas elementares: os fonemas associados às vogais orais, às vogais nasais, às consoantes e às semi-vogais (ver Tabela 1).

Certos fonemas elementares combinados produzem sub-cadeias orais importantes, associadas a p-grafemas de frequência considerável. Por exemplo, podemos citar o p-fonema "y.on.", encontrado frequentemente nas flexões verbais, na língua francesa.

Os 33 p-fonemas elementares combinados permitem obter mais de 160 p-fonemas não elementares, associados a cadeias gráficas mais longas. Por exemplo, sejam os p-fonemas "o.in." e "e<.t." definidos no sistema PILAF. O p-fonema "o.in." pode ser associado ao p-grafema "oin ", ou ao p-grafema "oint " etc. O p-fonema "e<.t." é aplicável aos p-grafemas "ette ", "ete ", "ettes ", "etes " etc.

Os p-grafemas idealizados para calcular as cadeias orais de uma palavra são associados a um conjunto de modelos fonéticos, que são validados ou não durante o tratamento da palavra mal ortografada. A validação de um determinado modelo dá lugar a uma sub-cadeia oral: o p-fonema associado a este modelo.

Na determinação das cadeias orais, vários modelos podem ser validados a partir de um mesmo p-grafema, o que permite a produção de várias cadeias orais para uma mesma palavra.

Por exemplo, examinemos a grafia **bo**. Esta palavra, após um tratamento de detecção, é considerada errônea, o que dá origem a um processo de correção. O método de correção através da Fonética, aplicado à cadeia **bo**, calcula, inicialmente, todas as cadeias orais possíveis para esta grafia, observadas as características da língua francesa. Como resultado obtemos, neste caso, uma única cadeia oral: "b.o.".

Passamos então à geração de grafias alternativas, propondo, para a cadeia oral "b.o.", as grafias **beau** et **beaux**, e ainda outras como **beaus**, **bau**, **baux** e **bot**. As grafias geradas são então analisadas quanto à integridade, eliminando-se as cadeias incorretas.

TABELA 1

FONEMAS	PILAF
a appel, arbre, pâte	b boucle, brebis
e< lait, blé, après, manger, tête	ch tâche, couche, cheval
e> recherche, petit	d sud, deux
i ici, vie	f faveur, oeuf
o beau, mot, gauche	g guichet, gant
u tu, sûr	j jouet, genou
ou doux, tout	k kilo, cave, psychopathe, qui
oe oeil, leur	l lourde, balle
	m sommet, mouche
	n nez, annonce
	p pile, coupe
y fille, mention, yoga	r route, arrosoir, finir
w squash, week-end, loin	s passion, nation, cil, ça, son
	t thème, tard
	v vif, wagon
	z maison, zone
in sapin, pain	
an dans, temps, vent, lampe	
on coupon, tombé	
un un, jungle	gn baigner, châtaigne
	ing parking
muet haut, pommes	

7 - OS MODELOS FONÉTICOS

O conjunto de regras, modelos e dicionário

A determinação das cadeias orais para uma palavra é feita pelo transdutor de estados finitos, usando três tipos principais de estruturas de dados: as regras, os modelos fonéticos e o dicionário de p-grafemas.

As regras permitem a análise da palavra reconhecendo a posição de seus grafemas componentes. Nós definimos regras que permitem o reconhecimento do espaço em branco, do início e do fim de palavra e do conteúdo interno aos limites de início e fim de palavra.

Os p-grafemas são armazenados num dicionário, atualmente com mais de mil registros. A cada p-grafema, existe um modelo fonético associado.

Os modelos fonéticos, finalmente, contêm as informações referentes a sua aplicabilidade, num dado estado, pelo autômato. Entre tais informações, encontramos as regras de chegada a este estado, e as regras que validam este modelo. O sistema PILAF se serve de mais de 200 modelos fonéticos.

Quando um modelo é considerado aplicável, o p-fonema associado a este modelo é utilizado pelo autômato para a geração de uma sub-cadeia oral correspondente ao p-grafema analisado.

Tomemos, por exemplo, a grafia **sous**. Nela encontramos os p-grafemas **s**, **ou** e **s**, os quais possuem modelos fonéticos associados. A tradução fonética desta grafia produz duas cadeias orais: "s.ou." e "s.ou.s.". A produção de duas cadeias orais diferentes se deve à existência de várias ocorrências do p-grafema **s** no dicionário de p-grafemas, cada vez associado a um modelo diferente (neste caso, associado ao modelo muet ou ao modelo s).

O código de derivação

Os p-fonemas associados aos modelos fonéticos PILAF são armazenados como códigos de derivação, o que permite ao transdutor de estados finitos calcular uma cadeia oral resultante da tradução.

A cada modelo fonético corresponde zero ou um código de derivação. Os casos de ausência do código de derivação são devidos à existência de letras mudas ou que podem se comportar como mudas, hífen, etc.

Um mesmo código de derivação pode ser utilizado por vários modelos, desde que existam diferenças com relação às combinações de regras utilizadas.

8 - EXEMPLOS DE FORMAÇÃO DE CADEIAS ORAIS A PARTIR DE UMA CADEIA ESCRITA

A análise fonética da palavra mal ortografada pode produzir uma ou várias formas orais: suas diferentes possibilidades de pronúncia, segundo os modelos fonéticos. O sucesso, após a análise de um modelo, determina a concatenação do código de derivação do modelo à cadeia oral gerada.

Vejamos alguns exemplos:

Exemplo 1

palavra errônea:	vinku
cadeia oral:	v.in.k.u

Exemplo 2

palavra errônea:	okurranse
cadeia oral:	o.k.u.r.an.s.

Exemplo 3

palavra errônea:	peti
cadeias orais:	p.e<.t.i. p.e>.t.i.

Exemplo 4

palavra errônea:	chevale
cadeias orais:	ch.e<.v.a.l. ch.e>.v.a.l. k.e<.v.a.l. k.e>.v.a.l.

9 - CADEIAS GRAFICAS A PARTIR DE UMA CADEIA ORAL

As bases morfo-fonéticas

O dicionário de bases PILAF dá origem a uma estrutura de dados especificamente criada para melhor gerar as cadeias gráficas a partir de cadeias orais: o dicionário Morfo-Fonético ou, simplesmente, dicionário M-F.

Este dicionário é constituído pelas cadeias orais associadas às

bases morfológicas. Existe uma cadeia oral associada a cada base morfológica. Se várias bases idênticas possuem uma mesma cadeia oral associada, o dicionário M-F conterá uma referência, somente, a esta cadeia oral.

Todas as bases têm cadeias orais associadas, mesmo que sejam cadeias mudas.

Por exemplo, consideremos a base savant. Esta base se apresenta duas vezes consecutivas no dicionário de bases PILAF: está definida sob um modelo de adjetivo e também sob um modelo de substantivo. No dicionário M-F, esta base aparece uma vez somente, sob a forma da cadeia oral "s.a.v.an.". Este fato se deve à pronúncia idêntica das duas bases savant.

Consideremos, agora, a base arrêt. Esta base também está presente duas vezes consecutivas no dicionário de bases PILAF. Ela está definida sob um modelo de substantivo, e também sob um modelo verbal o qual permite obter todas as formas do verbo "arrêter".

Neste caso, as duas ocorrências da base arrêt não se pronunciam da mesma maneira. Ao substantivo arrêt está associada a cadeia oral "a.r.e<.", enquanto que à base verbal arrêt está associada a cadeia oral "a.r.e<.t." (o fonema t deve estar presente porque ele é necessário em todas as conjugações do verbo "arrêter"). Ambas as cadeias orais aparecerão no dicionário M-F.

Uma terceira situação a mencionar é aquela que torna possível a formação de diferentes cadeias a partir de uma base morfo-fonética. Uma cadeia oral pode aparecer várias vezes no dicionário M-F, cada vez associada a uma base diferente. Tomemos a cadeia oral "s.e<." como exemplo. Esta cadeia aparece duas vezes no dicionário M-F, associada à base sec e à base cet.

Os registros do dicionário M-F são da seguinte forma:

°	
°	
s.a.v.an.	savant
°	
°	
a.r.e<.	arrêt
°	
°	
a.r.e<.t.	arrêt
°	
°	
°	

A produção de cadeias gráficas

A produção de cadeias gráficas é feita por uma versão simplificada do transdutor de estados finitos. Esta versão não utiliza regras nem modelos, uma vez que opera sobre um dicionário que já contém as cadeias e suas bases associadas: o dicionário M-F.

A cadeia oral é percorrida buscando-se todas as cadeias morfo-fonéticas possíveis. As bases morfo-fonéticas são então substituídas pelas bases morfológicas correspondentes, o que produz variantes gráficas para a palavra errônea. Durante este processo observam-se, para todas as grafias produzidas, as mesmas cadeias orais, ou cadeias orais próximas da palavra errônea.

Se a combinação de sub-cadeias orais não permite uma tradução, com relação ao dicionário M-F, a geração gráfica não produz resultados.

Analisemos a palavra **arête** como exemplo a corrigir. A cadeia oral "a.r.e<t." é a chave fonética calculada para esta grafia. A reconstituição gráfica, a partir do dicionário M-F, obtém a cadeia "a.r.e<t." como a mais longa cadeia oral contida na chave fonética calculada. A base associada à cadeia oral "a.r.e<t." é então utilizada para produzir a primeira parte da grafia resultante: **arrêt**.

A continuação da grafia resultante é calculada analisando-se a sequência da chave fonética. No caso proposto, a chave fonética foi, integralmente, levada em conta pela análise da sub-cadeia inicial. A geração gráfica termina, então, por adicionar alguns p-grafemas mudos à grafia resultante: **s, x, ent, e, es** etc, segundo regras de lingüística.

Como resultado de correção para o exemplo proposto, temos as grafias **arrête, arrêtent e arrêtes**. Utilizando um dicionário mais completo, as palavras **arête e arêtes** também seriam incluídas no resultado.

Possibilidades de produção

A geração de cadeias gráficas pode ser controlada no que concerne ao número de grafias geradas.

O princípio de tradução utilizado esgota todas as alternativas de combinações de sub-cadeias orais na análise da chave fonética, produzindo, às vezes, grafias com menor probabilidade de corrigir a palavra mal ortografada.

É possível afirmar que a melhor alternativa de correção através da Fonética é certamente a primeira encontrada no dicionário M-F, se este dicionário for suficientemente completo para a aplicação em questão, e convenientemente ordenado.

O algoritmo PILAF de correção de erros de Ortografia através da

Fonética trata esta questão propondo ao usuário a escolha entre a produção da primeira ou de todas as grafias encontradas.

No que concerne às situações em que é possível chegar-se a um resultado incompleto, devido à geração de somente a primeira grafia encontrada, o sistema toma algumas precauções.

Por exemplo, examinemos um caso de ocorrência de várias bases morfo-fonéticas idênticas. Seja **vinku** uma palavra mal ortografada detectada por PILAF. A palavra **vinku** produz uma única cadeia oral: "v.in.k.u."

Enquanto isso, no dicionário Morfo-Fonético, temos:

u.	u
v.in.k.	vainqu
v.in.k.	vainc

Portanto, a fase de recuperação gráfica dá origem às grafias **vainquu**, **vainquue**, **vainquus**, **vainquues**, **vaincu**, **vaincuc**, **vaincus**, **vaincues** etc, calculando um conjunto de alternativas que inclui todas as duas bases associadas ao verbo vaincre.

Esta precaução evita interromper a geração antes de ser encontrada a base vainc, que é a mais adequada na correção da palavra **vinku**.

10 - O RESULTADO: CADEIAS GRAFICAS MORFOLOGICAMENTE CORRETAS

A necessidade de uma verificação

A geração gráfica, mesmo feita a partir de bases morfo-fonéticas, pode produzir um mau encadeamento de bases, introduzindo, entre os resultados corretos, grafias que não constituem palavras. Por exemplo, podemos citar a grafia **vainquu** durante a correção de **vinku**.

É necessário, então, realizar uma verificação quanto à correção das cadeias gráficas calculadas.

Esta verificação é feita utilizando-se a Morfologia PILAF, e propondo como resultado somente as grafias que constituam palavras corretas. A ação de verificação elimina do resultado final todas as grafias

consideradas como incorretas pela Morfologia PILAF.

Uma alternativa

A verificação de integridade das grafias geradas pode ser dispensada se um dicionário de formas da língua francesa fizer parte do sistema. É então possível testar diretamente as grafias geradas, à partir da pesquisa no dicionário de formas [COH 88].

11 - CONCLUSÃO

O módulo de correção desenvolvido trata em três níveis o problema da correção de erros de Ortografia: o cálculo de uma chave fonética de pesquisa, a produção de grafias para a chave proposta e uma verificação de integridade das grafias geradas.

Este algoritmo se aplica bem às palavras contendo vários erros de Ortografia, desde que a forma oral seja conservada. Alguns acentos também podem ser recuperados pelo método, e o mesmo ainda deixa aberta a possibilidade de considerar a palavra mal ortografada como a concatenação de várias palavras, sem separador (por exemplo, *equivalait*: omissão dos separadores em *et qui valait*).

O sistema funciona atualmente com um dicionário de teste, o qual estamos enriquecendo. A execução do sistema com um dicionário maior permitirá a obtenção de resultados mais completos.

Uma proposição a fazer, observando-se o comportamento autônomo de cada nível de nosso algoritmo de correção através da Fonética, seria a construção de uma arquitetura funcional paralela, para os três níveis do algoritmo desenvolvido. O paralelismo é uma alternativa que se mostra, neste caso, eficaz e, ao mesmo tempo, consideravelmente natural.

A multiplicidade de resultados gerados, atualmente, pelos métodos de correção de erros de Ortografia, nos leva à proposição de um processo de filtragem, eliminando as opções menos prováveis de correção. No sistema PILAF, uma filtragem sintática está em desenvolvimento, tendo por base a pré-análise sintática da frase contendo a palavra errônea. Partindo-se da construção das árvores sintáticas alternativas, certas características morfológicas da palavra mal ortografada podem ser conhecidas, e algumas alternativas sintaticamente nulas podem ser eliminadas do conjunto final de proposições de correção.

Analizando o perfil do usuário final, devemos agora explorar as diferentes formas de ligação possíveis entre os módulos de correção

disponíveis. Os erros de competência, por exemplo, são mais facilmente corrigíveis por um método que se aperceba das más flexões aplicadas a uma base (como o método da Geração pela Morfologia), ou por um método que procure os equivalentes fonéticos para uma grafia. Dificuldades de performance poderiam ser tratadas com o suporte de um dicionário personalizado, ou da "clé squelette". PILAF propõe um conjunto de arquiteturas destinadas a tratar, distintamente, os erros de competência e de performance, bem como o Ensino assistido por Computador.

A vantagem do método desenvolvido, com relação aos já existentes, reside na formação de grafias diretamente a partir de bases, e não a partir de grafemas alternativos, o que reduz consideravelmente o número de cadeias gráficas foneticamente calculadas. Considerando-se a forma como as bases morfo-fonéticas estão armazenadas, sua recuperação pressupõe que a correção mais provável seja a primeira a ser encontrada.

BIBLIOGRAFIA

- CAT 80 CATCH, Nina. L'orthographe française - traité théorique et pratique. Ed. Nathan, Paris, 1980. 334 p.
- COH 88 COHARD, Brigitte. Logiciel de détection et de correction des erreurs lexicales. Memoire présenté en vue d'obtenir le Diplôme d'Ingenieur C.N.A.M.. C.N.A.M., Centre régional associé de Grenoble, USTMG, Mars 1988.
- COU 77 COURTIN, Jacques. Algorithmes pour le traitement interactif des langues naturelles. Thèse d'état USTMG - INPG, Octobre 1977.
- COU 86 COURTIN, Jacques, DUJARDIN, Danièle, Kowarski, Irène. Le système PILAF. Table Ronde Franco-Allemande sur les Langues Naturelles, Décembre, 1986.
- DAM 64 DAMERAU, F. A technique for computer detection and correction of spelling errors. Communications of the ACM, Vol. 7 Num. 3, Mars 1964. pp. 171-176.
- FOU 84 FOURNIER, J. "Sauvetage" de raisonnements en langage naturel. Thèse, Orsay Université Paris-Sud, Juin 1984.
- LAH 86 LAHENS, François. Un modèle stochastique pour la vérification

et la correction automatique de textes: le système VORTEX.
Thèse, Université Paul Sabatier. Toulouse, 1986. 166 p.

- LEV 66 LEVENSHTAIN, V. Binary codes capable of correcting deletions, insertions and reversals. Soviet Physics Doklady, Vol. 10 Num. 8, 1966. pp. 707-710.
- MAR 87 MARET, Dominique. Comparaisons de chaînes de caractères, acces lexicaux tolérants et applications. Thèse de Doctorat, Université de Sciences Sociales de Grenoble, Mai 1987.
- PER 86 PERENNOU, Guy, DAUBEZE, Pierre, LAHENS, François. La vérification et la correction automatique de textes: le système VORTEX. T.S.I. - Technique et Sciences Informatiques, Vol. 5 Num. 4, 1986. pp. 285-304.
- PET 80 PETERSON, James. Computer programs for detecting and correcting spelling errors. Communications of the ACM, Vol. 23 Num. 12, Décembre 1980. pp. 676-687.