Preâmbulo ao aconselhamento ortográfico para o português do Brasil — uma releitura baseada em utilidade e conhecimento lingüístico

Jorge Marques Pelizzoni

Orientação: Profa. Maria das Graças Volpe Nunes

A meu pai.

Agradecimento

a você que lê e crê que essas palavras
todas as palavras são o meu reflexo
lá do verso do outro lado de cada
página a olhar para você e agradecer
por tudo e tudo que eu lhe devo
agradecer e então que eu lhe
agradeço mais um pouco pelo que a
gente nem
sabe o quê

Resumo

Neste trabalho, fazemos uma releitura crítica do problema do aconselhamento ortográfico para o português do Brasil, entendido como a correção interativa de erros ortográficos em palavras isoladas, e reagimos. Em primeiro lugar, identificamos um parâmetro de qualidade importante — utilidade — e argumentamos que as soluções correntes o negligenciam. Procuramos meios de maximizar esse parâmetro, um dos quais justificamos ser o embasamento lingüístico, e daí propomos uma arquitetura genérica de sistema corretor interativo centrado em utilidade. Continuamos o trabalho levantando conhecimentos lingüísticos relevantes e fazendo considerações úteis ao desenvolvimento de um conselheiro ortográfico segundo o modelo proposto.

Abstract

In this work, we review the situation of spell-checking for Brazilian Portuguese from a (very) critical point of view and react. We identify an important quality factor — utility — and argue that it has been neglected by current solutions. We search for ways to maximize that factor, one of which we justify to be massive employment of linguistic knowledge, and hence propose a generic architecture for utility-centered interactive error correctors. Then we go on to gather relevant linguistic knowledge and give useful pointers on the development of a spell-checker according to our model.

Sumário

Índice de tabelas i

Índice de figuras iii

Símbolos fonéticos v

Preâmbulo vii

Considerações sobre formatação viii

Capítulo I O Direito de nascer — Utilidade 1

Contexto 2

Motivação: utilidade e conhecimento lingüístico 5

Objetivo 12

Capítulo II Correção de erros ortográficos em palavras isoladas 15

II.1 Genealogia 15

II.2 Diferentes aplicações e suas especificidades 17

II.3 Padrões de erro ortográfico 18

II.3.1 Aspectos subjetivos 19

II.3.2 Erros simples 20

II.3.3 Comprimento das palavras 21

II.3.4 Erros na primeira letra 21

II.3.5 Influência do teclado 22

II.3.6 Regras heurísticas e tendências probabilísticas 22

II.3.7 Listas de erros comuns 24

II.4 Técnicas 24

II.4.1 Mínima distância de edição 26

II.4.2 Chaves de similaridade 29

II.4.3 Regras 31

II.4.4 Análise de n-gramas 32

II.4.5 Redes Neurais 34

II.4.6 Técnicas probabilísticas 35

Capítulo III Da Origem e fim dos não-vocábulos — Breve reflexão filosófico-metodológica . 37

Argumento de utilidade 38

Realismo ou O não-atributo ou ainda O papel do embasamento lingüístico 41

Medida de utilidade 42
Medindo e maximizando utilidade e propaganda do paradigma reverso 44
Reversão: otimismo, profundidade, intenção, gatos & microondas 46
Revertendo erros ortográficos 48
Um conselheiro (qualquer) centrado em utilidade segundo o paradigma reverso 50

Capítulo IV Nossos sistemas de escrita 55

Não é fácil, não! 55
Caos aparente: leitura precária vs. expressividade 57
Divergências e inconsistências — esclarecimento e crítica 60
Ensinando o computador a ler 61
Nenhuma palavra é acentuada até que se prove o contrário 62
A conspiração das vogais: encontros vocálicos e tuiuiús 64
Do fonológico e do fonético 70
Made in Taiwan ou Os ditongos e hiatos de "Paraguai" 74
Âncora fonológica 77
Âncora étimo-morfológica 78
Âncora fonética 79

Capítulo V Alguns Erros naturais 81

V.1 Deturpação fonológica 81
Deturpação fonológica neutralizável 82
V.2 Erros de classificação — Um estudo de caso em morfologia 83
Classificação: uma operação potencialmente confusa 83
Aplicação à morfologia 88
De "~~di~~" para "dei" 90
V.3 De "~~saldo~~" para "saúdo" 93

Capítulo VI Conclusões e Trabalhos futuros 95

Referências bibliográficas 97

Índice Remissivo 107

Índice de tabelas

Tabela I: Sugestões de correção geradas por quatro spell-checkers para alguns erros comuns.....6

Tabela II: Evidências da ausência de tratamento morfológico em quatro spell-checkers. 9

Tabela III: Interseção entre a taxonomia de Kukich (92) e as demais. 26

Tabela IV: Definição de e(x, Cx) e flexões correlatas segundo cada hipótese. 92

Índice de figuras

Figura 1: Reversão[38] do processo de secagem/sacrifício do gato. 47

Figura 2: Arquitetura genérica de reversão centrada em utilidade.....50

Figura 3: Fluxo de certeza na divisão silábica de "tuiuiú" e "~~tuiuiu~~". 68

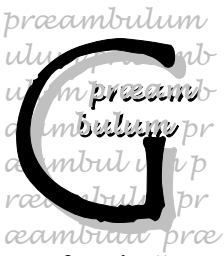
Figura 4: Amostra do formalismo gramatical utilizado. 89

Símbolos fonéticos

Obs.: todos os símbolos fonéticos deste documento pertencem ao *International Phonetic Alphabet* (IPA). Apresentam-se a seguir apenas os símbolos não-óbvios.

Símbolo	Significado
[]	fone ou forma fonética
/ /	fonema ou forma fonêmica
ˈ	acento tônico. Ex.: “bolo” = [ˈbo.lu]
.	fronteira entre sílabas. Ex.: “bola” = [ˈbɔ.la]
ʒ	“j”, como em “**j**eito”
ʤ	“d” como em “**d**ia”
e	“e” fechado, como em “p**ê**lo”
ɛ	“e” aberto, como em “p**é**lo”
j	“i” assilábico. Ex.: “fui” = [fuj]
ɲ	“nh”
o	“o” fechado, como em b**o**lo
ɔ	“o” aberto, como em “b**o**la”
ʃ	“x” como em “**x**adrez”
ʧ	“tch”, como em “**tch**au”
w	“u” assilábico. Ex.: “vou” = [vow]
x	“r” velar surdo, como em “fica**r**”
ɣ	“r” velar sonoro, como em “ca**rr**oça”
R, S, N, L	**arquifonemas**, que se realizam como fones diferentes em função do contexto fonético ou ainda do dialeto considerado. Os símbolos são auto-explicativos.

Preâmbulo



rego para alguns mas latina para a etimologia, "preâmbulo" é de uma polissemia pródiga e aqui especialmente providencial. É comum ter uma acepção "v. *prefácio*" na maioria dos dicionários; e é obviamente essa a que toma ali no canto superior esquerdo desta página e que o Aurélio (96) parafraseia "texto ou advertência, ordinariamente breve, que antecede uma obra escrita, e que serve para apresentá-la ao leitor". E gostaríamos de acrescentar aí: "e em que se permitem certas licenças, para deixar autor e leitor bem à vontade". Pois bem, é exatamente à guisa de advertência que inserimos este preâmbulo *em vez de uma introdução*: é que a introdução propriamente dita ocorre no Capítulo I, que, não sem boas razões, alonga-se um pouco além do usual para introduções. No entanto, há coisas que precisam ser ditas logo, para que não se deixe o leitor a conjecturar demais; e resolvemos dizê-las logo aqui.

Em primeiro lugar, dizer que o aconselhamento ortográfico, entendido como a correção de erros em palavras isoladas (consideradas fora de contexto), costuma ser visto como *um caso perdido menor*. Admitimos que o problema não tenha solução definitiva porque uma tal solução simplesmente não existe no nível da palavra isolada; no entanto, isso não justifica a crença mais ou menos generalizada de que, sem superar a cegueira contextual primordial, o problema não seja mais digno de pesquisa. Desbancar essa crença, ou seja, conquistar a este trabalho o próprio direito de nascer, é a grande missão do Capítulo I e o motivo por que seja tão longo, contenha uma das contribuições mais valiosas de todo o projeto e não deva ser chamado meramente "Introdução".

Em segundo lugar e por fim, já adiantar que, como *não* é praxe em Computação, este Mestrado está todo em papel, aqui; ou seja, a esperada implementação de um conselheiro ortográfico está toda no futuro (próximo). Tal é o motivo da escolha do título desta dissertação, de onde "preâmbulo" paira com propriedade, enigmática, polissêmica. Prontamente acode o Aurélio, lembrando mais uma acepção, bela e auspiciosa: "palavras ou atos que precedem as coisas definitivas". A que o Houaiss (01) responde: "por extensão, palavreado vago que não vai diretamente ao fato". Porquanto a polissemia não ajude, cabe ao leitor decidir, segundo o

contexto.

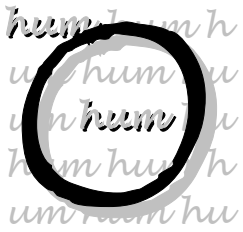
Considerações sobre formatação

Neste documento, os erros ortográficos intencionais, freqüentemente usados para ilustrar a argumentação, serão grafados em fonte ~~tachada~~; neologismos e palavras de validade discutível, por sua vez, serão marcados com sublinhado ondulado. Pretende-se, dessa forma, distinguir esses erros ou "excentricidades" de maneira bem clara e não-ambígua e minimizar possíveis danos aos bons hábitos ortográficos do leitor (e do autor!).

Muitos termos e expressões estarão em **negrito** por corresponderem a entradas do *índice remissivo* consideradas especialmente interessantes. Dessa forma, esperamos agilizar o uso desse índice, que avaliamos como um recurso bastante oportuno num trabalho como este, abundante em definições (ora inéditas, ora pouco familiares), referências cruzadas e seções longas. Não por coincidência, portanto, muitos dos termos em negrito ocorrem em meio às suas respectivas definições, implícitas ou explícitas. Reciprocamente, no índice remissivo, os números de página em negrito correspondem a ocorrências especialmente interessantes, em quantidade e qualidade de informações, tratando-se freqüentemente de definições.

Quanto aos estilos de título, usaremos numeração apenas para seqüências de seções tratando de assuntos paralelos — sem grande dependência entre seções adjacentes — mas subordinados a um tópico superior. Por outro lado, entendemos que a ausência de numeração torna mais clara a dependência "cronológica" ou implicação natural entre seções adjacentes. Ou seja, por meio da presença ou ausência de numeração no título de uma seção, enfatizamos ora hierarquia e paralelismo, ora seqüência, como o(s) aspecto(s) que melhor situam a seção em questão em relação às demais.

Capítulo 1 O Direito de nascer — Utilidade



termo **aconselhamento ortográfico** se refere ao processamento de um texto em uma dada língua natural com os objetivos de (i) identificar os erros cometidos quanto à ortografia ("palavras" que não constam do léxico dessa língua[1] ou usadas em contexto impróprio) e (ii) sugerir alternativas prováveis e ortograficamente corretas a cada erro identificado, geralmente classificadas[2] segundo alguma estimativa da probabilidade de serem realmente a correção pertinente. Neste trabalho, no entanto, adota-se a acepção mais restrita e usual que, de acordo com Kukich (1992), pode ser expressa como *correção interativa de erros ortográficos em palavras isoladas*, uma vez que o contexto de enunciação não é considerado e é certamente uma entidade externa, provavelmente um usuário humano, que decide qual das sugestões se aplica em cada caso[3].

---

[1] Uma palavra que não consta (ou não pode constar) do léxico de uma língua não é uma palavra (!), no sentido formal do termo (veja Monteiro [86], por exemplo). Informalmente, entretanto, usa-se "palavra" de modo mais livre, o que pode ser observado quando dizemos, por exemplo, "essa palavra não existe" em vez de "isso não é uma palavra". Ou seja, "palavra" parece de ordinário denotar algo que poderia ser expresso mais formalmente como "cadeia de caracteres, devidamente delimitada, que representa ou pretende representar um *vocábulo*".

[2] A classificação das alternativas pode ser representada tanto pela ordem em que elas são exibidas quanto por se associar a cada alternativa um ou mais índices numéricos, absolutos ou relativos. O emprego de mais de um índice é, na verdade, apenas uma especialização da geração simultânea de múltiplas classificações, cada qual segundo critérios próprios, ou seja, focalizando características distintas das alternativas produzidas.

[3] Vale observar que isso não torna desnecessário classificar as sugestões dadas. Incumbido da escolha final entre possibilidades de correção, um usuário certamente achará muito incômodo não encontrar as alternativas corretas freqüentemente como a primeira ou entre as primeiras apresentadas. De fato, a *precisão da primeira sugestão* (do inglês "first-guess accuracy") é uma das mais importantes medidas de desempenho de sistemas de aconselhamento ortográfico, sendo crucial em sistemas não-interativos.

De fato, a maioria dos sistemas que realizam esse tipo de processamento — também conhecidos como **conselheiros/corretores ortográficos**, ou ainda ***spell-checkers*** — não são sensíveis às informações disponíveis nos níveis da sintaxe e acima, ou o são de forma extremamente limitada. Diante da frase "Não se ~~molh~~ ao atravessar a cachoeira!", por exemplo, o spell-checker típico deverá sugerir "molho", "molhe", "molha" e talvez até "mol" como prováveis alternativas a "~~molh~~", mas não será capaz de determinar com segurança qual das três possibilidades de correção realmente procede.

Contexto

Apesar dessas severas limitações, spell-checkers são úteis em alguns cenários bastante abrangentes, cabendo destacar os seguintes:

i. **na edição de textos em geral, em língua natural**, uma vez que a maior parte dos usuários de processadores de texto costumam cometer erros ortográficos, devidos tanto às dificuldades intrínsecas à própria ortografia da língua, quanto a fatores extralingüísticos: teclado, hábitos de digitação, má visualização do texto digitado, pressa, problemas de visão, estresse, etc. Nesse cenário, eis alguns dos usos mais comuns que são feitos dos spell-checkers, ou de sistemas que os embutem:

a) *detectar*, mais do que corrigir, pequenos erros de digitação, que geralmente passam despercebidos mesmo por um revisor humano mais cuidadoso e muitos dos quais resultam num vocábulo inválido;

b) poupar esforço de digitação na correção de palavras, mesmo quando o usuário verifica que acabou de digitar errado uma palavra. Refiro-me aqui mais especificamente a editores de texto que, à maneira do Microsoft® Word (Microsoft, 97 e 99), vão marcando o texto quanto a erros de ortografia à medida que o usuário o vai digitando e permitem, praticamente sem sair do modo de edição, visualizar alternativas de correção e selecionar uma delas, que imediatamente substitui a cadeia previamente marcada;

c) minimizar erros ortográficos reais, cuja correção muitas vezes pode escapar à competência do usuário. A maioria dos usuários, caso digite "~~eassarola~~" e seja avisado

de que não existe palavra com essa grafia, provavelmente não precisará de um spell-checker para tentar "caçarola" como correção. Entretanto, considerem-se usuários típicos que cometessem um ou mais dos seguintes erros ortográficos: "~~saberam~~" ("souberam"), "~~proporam~~" ("propuseram"), "~~nódca~~" ("nódoa"), "~~atrapalhamento~~" ("atrapalhação"), "~~reaveram~~" ("reouveram") e "~~uso capião~~" ("usucapião"). Estariam eles capacitados a produzir as hipóteses de correção pertinentes, sem ajuda externa?

d) verificar a validade de uma palavra ou de sua grafia, talvez "inventadas" (inferidas) pelo usuário.

Na enumeração acima, os dois últimos itens, (c) e (d), são de especial interesse, uma vez que é muito provável que decorra desse tipo de uso alguma espécie de **aprendizado por parte do usuário**, mesmo que de forma inconsciente e incidental (Gupta, 98; Desmarais, 98; McClurg & Kasakow, 89). Nesse sentido, um comportamento interessante para um spell-checker é aproveitar o interesse e a receptividade do usuário nessas ocasiões, avisá-lo da existência de informações relevantes à solução do problema em questão (no momento exato em que estas se tornam relevantes!) e facultar ao usuário rápido acesso a essas informações. Essa possibilidade é de especial interesse para nós e uma motivação adicional para algumas de nossas decisões de projeto;

ii. **no processamento mais avançado de línguas naturais**, como subsistemas aos quais se podem requisitar prováveis alternativas a uma palavra lexicalizada ou não, possivelmente dadas dicas sintático-semânticas. É exatamente funcionando como componente de preprocessamento que um conselheiro ortográfico (lá denominado "*isolated word recognizer*") aparece num diagrama de blocos proposto por Kukich (1992, pág. 431) para um "sistema hipotético de reconhecimento de palavras/detecção de erros (ortográficos) baseado em expectativas e de larga escala que exploraria muitas das fontes de conhecimento lingüístico disponíveis para os humanos" (pág. 431), ou que se poderia chamar de um "conselheiro ortográfico ideal". Diferentemente desse sistema, em que o conselheiro seria invisível externamente, existem sistemas em que sua presença é facilmente notada pelo usuário, como acontece com o pacote RLP (Itautec, 99) de apoio à escrita, que analisa cada frase em duas etapas, uma de verificação ortográfica e outra de revisão gramatical, nessa

ordem;

iii. **no acesso a bases de dados**, para evitar que consultas sejam malsucedidas graças (i) a erros de ortografia tanto nos argumentos de busca quanto nos dados da base ou ainda (ii) a grafias variantes, ambos tão comuns em se tratando de nomes próprios, por exemplo;

iv. **na edição de código-fonte em linguagens de programação,** caso bem distinto de (i) por diversas características, bastando mencionar as seguintes:

   a) léxico geralmente muito menor, contendo um número pequeno de entradas predefinidas (palavras-chave) e sendo ativa e arbitrariamente estendido pelo(s) autor(es) do texto;

   b) sintaxe e semântica simples, não-ambíguas e formalmente definidas.

   Num ambiente de programação qualquer, dispor de um conselheiro ortográfico específico para a linguagem usada e abarcando projetos compostos de múltiplos arquivos varia de "apenas conveniente", no caso geral, a "quase indispensável", no caso particular de certas linguagens. *Prolog*, por exemplo, é notório pela alta probabilidade com que erros de digitação básicos passam despercebidos pelo compilador/interpretador e acabam por ocasionar erros de execução sérios e de difícil depuração.

Como se pode depreender da enumeração acima, conselheiros ortográficos podem ser úteis para usuários diferentes de léxicos diferentes. No eixo dos léxicos, os elementos variam, só para citar alguns grupos de relevância, entre artificiais e naturais e quanto ao conjunto de línguas abrangidas. No eixo dos usuários, por sua vez, distribuem-se falantes nativos do léxico tratado ou não, mais ou menos familiarizados com o mesmo e apresentando diferentes graus de letramento e níveis de desempenho na norma gramatical vigente e em digitação, entre outros. É razoável, pois, esperar que os tipos de erros ortográficos cometidos, sua distribuição probabilística e sua relevância variem com o par *(usuário, léxico)* considerado. Ou seja, é razoável esperar que conselheiros ortográficos e técnicas de aconselhamento ortográfico desempenhem melhor que outros em condições específicas e apresentem queda de performance fora dessas condições ideais, a não ser que sejam capazes de perceber as variações vigentes e a elas reagir.

Motivação: utilidade e conhecimento lingüístico

Na avaliação do desempenho de conselheiros ortográficos, parâmetros do tipo "**precisão das *n* primeiras sugestões**" são extremamente usuais, representando a probabilidade de, detectado um erro ortográfico qualquer, a correção pertinente estar entre as *n* primeiras alternativas sugeridas ao usuário. É evidente a relevância desse tipo de parâmetro; e não surpreende, pois, que figure freqüentemente entre os de maior peso e mais persuasivos, sobretudo para valores pequenos de *n* (*n*=1, por exemplo). Mais que isso, a literatura consultada, ao tratar da qualidade das respostas geradas pelos spell-checkers citados/descritos, praticamente não demonstra outra preocupação que comparar/maximizar estimadores de tais parâmetros. De fato, não há muito mais o que se considerar em se tratando de sistemas não-interativos de correção automática de textos; por outro lado, defende-se neste projeto, esses parâmetros podem ser, no mínimo, insuficientes e, no máximo, ilusórios, na avaliação de spell-checkers que interagem com usuários humanos. Basta considerar o caso hipotético de um spell-checker cuja precisão da primeira sugestão é de 85%. Em princípio, esse é um indício bastante favorável a que o desempenho desse sistema seja considerado muito bom; mas *e se* a capacidade de correção do spell-checker e a do usuário falham de forma sistematicamente coincidente? Ou seja, *e se*, toda vez que a correção pertinente é gerada, o usuário também está apto a gerá-la e, toda vez que o sistema falha, o usuário também se vê incapaz de gerar a hipótese certa de correção? Nesse caso, pode-se dizer que, embora a precisão das sugestões seja alta, sua **utilidade** é baixa; ou seja, a utilidade do módulo de correção do spell-checker como um todo é baixa, e o sistema está sendo usado bem mais como um detector do que como um corretor de erros ortográficos.

Fica fácil sentir quão útil é a saída de um spell-checker ou quanto cuidado se dispensou a esse aspecto quando ela é **analisada de forma "reversa"**, ou seja, gerada uma lista de sugestões de correção $[C_1, C_2, ..., C_n]$ para um dado erro ortográfico E, assumir cada C_i como a correção pertinente e especular quão fácil seria para o usuário chegar a C_i (ou algo próximo o suficiente para que o spell-checker pudesse dar melhores sugestões) sem ajuda externa e partindo apenas da identificação de E como erro. Para exemplificar esse tipo de análise, tomem-se as saídas de diferentes spell-checkers para o erro "~~estrupo~~" ("estupro"), apresentadas na Tabela I. Considerando-se apenas as possibilidades menos absurdas, se o usuário pretendesse ter usado as palavras "estrepo", "estripo", "estropo" ou até "estupor", é quase certo que ele dispensaria

qualquer ajuda externa para efetuar a correção necessária, visto tratar-se de um mero erro de digitação e as cadeias envolvidas apresentarem acentuado contraste em pronúncia. O mesmo vale para todas as outras sugestões (especialmente "estros"!), exceto "estupro". Se essa fosse a palavra originalmente pretendida, o erro em questão muito provavelmente teria uma causa cognitiva e seria de mais difícil recuperação por parte do usuário. Com isso em mente, é no mínimo curioso que "estupro" tenha sido sugerida por apenas um dos spell-checkers e, então, apresentada em último lugar. A Tabela I ainda apresenta os resultados obtidos para os erros "~~reaveram~~" ("reouveram", 3ª pessoa do plural de "reaver" no pretérito perfeito do indicativo) e "~~peneu~~" ("pneu"), entre outros, para que o leitor possa reproduzir o tipo de análise exemplificado e pressentir a relevância do problema.

Tabela I: Sugestões de correção geradas por quatro spell-checkers para alguns erros comuns.

~~Erro~~ (correção)	**Word 2000**[4] (Microsoft, 99)	**DTS** (DTS, 98)	**Gramática Eletrônica** (Lexikon, 97)	**Word 97** (Microsoft, 97)
~~estrupo~~ (estupro)	estopo, estripo, estrompo, estropio, estros	Estrio	estropo, estupor, **estupro**	esturro estrepo estripo estrumo
~~reaveram~~ (reouveram)	reabram, reagiram, reaverá, reaverás, reaverão	regeram, **reouveram**, reaviam, reaverás, reaverá	reaver —**ou**—[5] reaverá, reaverão, reaverás, reaverem, reaveriam	reaverão, reaverem, reaveriam, reatarem

[4] O conselheiro ortográfico embutido no Word 2000 é parte integrante da ferramenta de revisão gramatical desse editor de texto, a qual, por acaso, foi projetada pelo NILC, laboratório no âmbito do qual este projeto é desenvolvido. Temos acesso, por isso, ao código-fonte das rotinas de aconselhamento ortográfico usadas no Word 2000, as quais, entretanto, não foram desenvolvidas pelo NILC.

[5] O sistema *Gramática Eletrônica* oferece dois modos de correção, um mais rápido e outro mais "completo", ativados respectivamente pelos botões "semelhante" e "semelhantes". O segundo modo tem uma abrangência maior (cobrindo adicionalmente flexões verbais, por exemplo), gera mais sugestões e aparentemente aplica heurísticas mais refinadas. Por isso, nos exemplos apresentados, quando o modo rápido de correção desse sistema não foi bem-sucedido, incluíram-se os resultados obtidos pela invocação do segundo modo, devidamente precedidos do separador "— **ou** —".

~~peneu~~ (pneu)	painel, pediu, penal, pendeu, pendo, ..., **pneu** **(28º lugar!)**	penei, penou, pene, penes, peneis, ..., **pneu** **(13º lugar!)**	peleu, pene, penes, pinéu, **pneu**	penou pendeu pene penei penem
~~possue~~ (possui)	poço, poços, poções, pospõe, posse, ..., **possui** **(10º lugar!)**	possuo, possuis, **possui**, possuís, possuem	posse — **ou** — pelouse, pontue (x3)[6], porque, porquê	poste, posse, possua, **possui**, possuo
~~entitular~~ (intitular)	intitula, intitulai, intitulam, **intitular**, intitulara	intitulas, intitula, intitulam, intitulara, intitulará, **intitular**, ...	- — **ou** — enfistular, entijucar, estimular, estipular	entabular, entijucar
~~entitulados~~ (intitulados)	intituladas, intitulado, **intitulados**, intitulamos	intitulamos, intitulado, **intitulados**, intituladas, intitulador	entijucado, estipulado — **ou** — -	entabulados, entijucados

O que também assusta, ainda na Tabela I, é o fato de um conjunto de sistemas comerciais líderes de mercado deixar tanto a desejar diante de tipos de erros tão comuns e previsíveis, dado o conhecimento disponível em gramáticas e na literatura de Lingüística em geral, dirigidas ao público leigo ou especializado. Só a título de demonstração: a tendência de se confundir "~~estrupo~~" com "estupro" tem até nome — hipértese (de Almeida, 92, pág. 66); "~~reaveram~~" decorre simplesmente de se considerar o verbo "reaver" como regular, o que é compreensível, uma vez que as formas verbais resultantes não soam nada mal; e "~~peneu~~"/"~~pineu~~" se deve a uma

[6] O modo mais avançado de correção do sistema *Gramática Eletrônica* freqüentemente apresenta sucessões de sugestões idênticas em suas listas de alternativas de correção.

transcrição fonética justificável, já que, na fala, uma vogal é propriamente inserida no encontro "pn" de "pneu" (Câmara Jr., 70, pág. 57), no registro formal inclusive.

Em resumo, o conceito de utilidade e o conhecimento lingüístico disponível foram aparentemente negligenciados no projeto dos spell-checkers em questão. De forma geral, eles não parecem muito úteis na correção de erros explicados pela morfologia ("~~reaveram~~"), como um todo, ou pela fonética e fonologia (demais itens da Tabela I), quando o problema vai além dos casos mais óbvios de homofonia (por exemplo, troca de "ç" por "ss", "x" por "z" ou "ch", etc.). Em específico, as listas de correções obtidas para "~~entitular(dos)~~" e "~~possue~~" evidenciam um tratamento fonético-fonológico precário, sendo especialmente intrigante que "intitular(do)" e "possui", quando chegam a ser apresentadas, sejam preteridas em favor de outras obviamente menos pertinentes e que implicam erros de digitação de probabilidade duvidosa. Vale observar que esse "ruído" que surge nas listas de sugestões, especialmente notável nas geradas para "~~entitular(dos)~~", deve-se em grande parte à apresentação de flexões demais para uma mesma palavra. O raciocínio é simples: se o erro é "~~entitulados~~" e a sugestão de "intitulados" não for *útil* o bastante, então (i) apresentar essa alternativa *em meio* a "intituladas", "intitulado" e "intitulada" será não só inútil quanto prejudicial e (ii) nenhuma dessas flexões será mais útil por si só[7]. Esse tipo de raciocínio já penetra nos domínios da morfologia e, como se pode concluir ainda de "~~entitular(dos)~~", *talvez* tenha sido considerado nos sistemas *Gramática Eletrônica* e *Word 97*, que parecem "respeitar" algumas das terminações de palavra mais características.

No entanto, a ausência de tratamento morfológico *adequado* por parte de *todos* esses sistemas fica patente quando se analisam os resultados apresentados na Tabela II. Para começar, vale notar que nenhum dos spell-checkers teve sucesso na correção dos erros lá presentes, todos provavelmente com razões morfológicas. Ambos "~~reaveu~~" ("reouve", 3ª pessoa do singular de "reaver" no pretérito perfeito de indicativo) e "~~proporam~~" ("propuseram") desmentem a possível capacidade de tratamento desse tipo de erro pelo sistema DTS, sugerida pela correção bem-sucedida de "~~reaveram~~"(Tabela I). O erro "~~assimilamento~~" é uma tentativa morfologicamente

[7] Em tempo: não se induza disso a falácia "se flexões de uma mesma palavra, então redundantes"! O ponto aqui é reconhecer conjuntos de flexões cujos elementos *jamais* poderiam ser confundidos entre si pelo usuário, o que não acontece, por exemplo, com *{venderão, venderam}*.

plausível, mas **bloqueada** pela preexistência de "assimilação" e decorrente de serem praticamente idênticas as propriedades morfológicas e semânticas dos sufixos "-mento" e "-ção". Por outro lado, é discutível o bloqueio de "pré-câmara" por "antecâmara", mas seria conveniente que os spell-checkers ao menos aconselhassem o uso desta última forma, já estabelecida, enquanto a primeira provavelmente nem consta de seus léxicos internos[8]. Finalmente, "~~transandar~~" ("tresandar") é um caso curioso, talvez motivado por analogia com o trio de sinônimos "transpassar", "trespassar" e "traspassar", onde se evidenciam as variantes "tres-" e "tras-" do prefixo "trans-", muito embora, diz o Dicionário Aurélio Eletrônico (Aurélio, 96), a verdadeira origem do "tres-" em "tresandar" seja o vocábulo "trás"[9].

Tabela II: Evidências da ausência de tratamento morfológico em quatro spell-checkers.

~~Erro~~ (correção)	**Word 2000**	**DTS**	**Gramática Eletrônica**	**Word 97**
~~reaveu~~ (reouve)	ravel, reagiu, reavei, reaver, reavia	regeu, reaveis, reaverá, reavei, reaver, ...	Reaver — **ou** — reative (x3), reaver, rebateu	reavei, reaver
~~proporam~~ (propuseram)	procuram, profiram, proporá, proporás, proporão	proporás, proporá, proporão, proporias, proporiam	propor — **ou** — preparam, proporá, proporão, proporás, proporção	proporão, proporem, proporiam
~~assimilamento~~ (assimilação)	-	-	assinalamento — **ou** — (idem)	assinalamento
pré-câmara (antecâmara)	(aceita)	(aceita)	preaca — **ou** — pré-datara, pré-datará	-

[8] Os casos de aceitação de "pré-câmara" parecem se dever a algum tipo de processamento morfológico não muito criterioso, o que é evidenciado pelo fato de os sistemas envolvidos igualmente aceitarem, por exemplo, "~~pré-senhorita~~", "~~pré-destinar~~", "~~pré-ter~~", "~~pré-haver~~", "~~pré-bonito~~", "~~pré-seu~~" e "~~pré-ninguém~~".

[9] De acordo com essa fonte, o verbo "tresandar" teria um sentido original de "fazer andar para trás".

~~transandar~~ (tresandar)	-	transadas, transando	transnadar, transumanar **— ou —** transladara (x2), transladará, transnadar, transumanar	transnadar, transnadara, transnadará, transnada, transnadai

Os exemplos presentes nas duas tabelas acima resumem classes inteiras de erros em cuja correção os spell-checkers testados não apresentaram bons resultados e que parecem poder ser satisfatoriamente tratadas com uso de conhecimento lingüístico abaixo do nível da sintaxe. Outras classes com problemática semelhante já foram identificadas e não serão aqui apresentadas e comentadas dado o caráter motivador deste capítulo.

Ainda quanto aos benefícios de um tratamento morfológico mais cuidado, a capacidade de apresentar "antecâmara" e "assimilação" como alternativas a "pré-câmara" e "~~assimilamento~~", respectivamente, sugere que um tal spell-checker deve estar apto a avaliar se uma forma desconhecida é plausível do ponto de vista de formação de palavras e, em caso positivo, gerar hipóteses de possíveis **formas bloqueantes** (semanticamente "equivalentes", já atestadas e, portanto, preferíveis). Nesse processo, o spell-checker pode concluir (i) que a forma suspeita é, na verdade, uma **tentativa frustrada de neologia** que não respeita os padrões morfológicos de produção vocabular da língua[10], ou, caso contrário, (ii.i) que não há nenhuma hipótese boa o suficiente para se contrapor à forma duvidosa e (ii.ii) que, por conseguinte, esta é *provavelmente válida*, tratando-se de um *neologismo*, uma *flexão* ou *derivação* de uma palavra aprendida em tempo de execução ou, de qualquer modo, uma palavra não prevista na construção do sistema. Dessa maneira, palavras seriam ou automaticamente adicionadas ao léxico do sistema, para evitar

[10] Compare, por exemplo, "~~léxicos-sintáticos~~" com "léxico-sintáticos" ou "~~leiturabilidade~~" com "esticabilidade". Diferentemente de "~~léxicos-sintáticos~~", talvez não seja tão claro o motivo de "~~leiturabilidade~~" ser inaceitável, a saber: substantivos terminados em "-bilidade" são formados por dupla sufixação, acoplando-se "-dade" a uma forma já derivada por meio do sufixo "-vel", o qual, por sua vez, só é compatível com *verbos*. Ou seja, "esticabilidade" se justifica pela seqüência de derivação "esticar → esticável → esticabilidade", em que o padrão descrito acima é seguido; da mesma forma, a aceitação de "~~leiturabilidade~~" só poderia ser justificada por uma seqüência análoga de derivação, que requereria a existência, totalmente absurda, de um verbo "~~leiturar~~".

intromissões distrativas/irritantes, ou simplesmente assinaladas como "provavelmente corretas", esperando o aval final do usuário. Essa última atitude não é necessariamente melhor, mas é definitivamente mais prudente, não deixando de causar uma impressão favorável no usuário quanto à "boa-vontade" e à "esperteza" do sistema em colaborar e aprender.

A propósito, o aprendizado de novas palavras em tempo de execução costuma ser bastante primário, sofrendo de algumas limitações generalizadas e razoavelmente desagradáveis ao usuário. Em primeiro lugar, em todos os sistemas testados, não são inferidas as flexões/derivações padrão, mesmo que regulares, de uma palavra prévia e explicitamente "ensinada" pelo usuário. Ou seja, cada flexão de uma nova palavra tem que ser adicionada separadamente para que seja reconhecida. Em segundo lugar, dois dos sistemas testados (DTS e Gramática Eletrônica) não levam as formas aprendidas em consideração no momento de gerar sugestões de correção, isto é, não são capazes de corrigir erros, mesmo que muito pequenos, cometidos na grafia das palavras novas.

O aconselhamento ortográfico é feito numa depressão profunda e árida (palavra isolada) cravada num paraíso tropical inatingível (contexto de enunciação e a língua propriamente dita). Há escassez de informação, e parece prudente aproveitar racionalmente o pouco de que se dispõe. Como demonstrado acima, um corpo significativo de informações fica disponível ao se analisarem os erros ortográficos à luz de conhecimentos lingüísticos, mesmo que restritos aos domínios da fonética, fonologia e morfologia. Pelo menos no caso do aconselhamento ortográfico para a língua portuguesa, essa possibilidade parece não ter sido bem explorada, ou melhor, posto com mais rigor, suspeitamos haver como explorar essa possibilidade de forma a favorecer sensivelmente o desempenho (utilidade) dos spell-checkers testados.

Finalmente quanto aos conselheiros ortográficos para o português descritos na literatura consultada (Lins et al., 99; Pacheco, 96; Almeida & Pinto, 95; Lucchesi & Kowaltowski, 93)[11] e não testados diretamente, é de acreditar que seu desempenho não seria muito diferente nos

[11] Infelizmente, *ainda* não tivemos acesso direto a (Almeida & Pinto, 95). Tudo o que sabemos desse trabalho provém de (Pacheco, 96) e é, em alguns aspectos, insuficiente. Portanto, pedimos desde já desculpas por qualquer eventual equívoco ou injustiça de nossa parte com relação a (Almeida & Pinto, 95), que terão sido cometidos sem intenção e provavelmente por causa de erro na interpretação do conteúdo de (Pacheco, 96).

quesitos aqui enfatizados. Pelo menos é o que se pode depreender das respectivas publicações, que tinham focos de interesse bastante diversos dos deste projeto: Lucchesi & Kowaltowski (1993) estavam obviamente mais preocupados com aspectos de representação de grandes léxicos (basicamente compactação e acesso eficiente) e, assim como Almeida & Pinto (1995), viam o aconselhamento ortográfico como um subproduto; por outro lado, Lins et al. (1999) e Pacheco (1996) tinham na correção automática de textos seu interesse principal, mas se "distraíram" com o processamento sintático. Maiores detalhes sobre esses trabalhos podem ser encontrados na Seção II.4.1, à página 27.

Objetivo

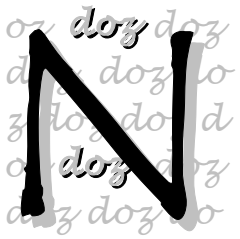
Nesse contexto, o objetivo deste trabalho é rever criticamente a situação do aconselhamento ortográfico[12], em especial para o português do Brasil, parte do que já fizemos neste capítulo, e *reagir*. Como é previsível, a reação parte da combinação de dois elementos básicos: ênfase em utilidade como meta, talvez até em detrimento da precisão, e embasamento lingüístico como meio. Não chegaremos à implementação de um corretor ortográfico, o objetivo original deste trabalho; mas daremos os passos de análise que acreditamos cruciais para um tal projeto de *software*.

Consumado o caráter motivador deste capítulo, vamos terminar de contextualizar nossa proposta no Capítulo II, revisando a literatura referente à correção ortográfica, no que poderemos notar que não há registro de um trabalho semelhante no tocante à ênfase dada ao levantamento e emprego de conhecimento lingüístico, para não mencionar o conceito de utilidade. Devidamente motivados e contextualizados, reagimos a partir do Capítulo III. Este último, em específico, funciona como um verdadeiro plexo metodológico em que (i) admitimos a imponderabilidade da utilidade; (ii) postulamos, em resposta, uma correlação entre utilidade e um par de elementos (relativamente) mais razoáveis — a saber, *(perfil de usuário, reconstituição)* — a que impomos certas condições — *verossimilhança*, *desafio* e *otimismo*; (iii) com base nessa correlação, definimos *medida de utilidade*; (iii) justificamos daí a opção pelo paradigma reverso de correção de erros e o dissecamos; (iv) projetamos uma arquitetura genérica de reversão que nos permite

[12] Vez por outra, não nos furtaremos a criticar a situação de outras entidades transeuntes, por motivo de vocação.

isolar a *gramática de reconstituição* como o elemento a ser enfocado no restante do trabalho e que, grosso modo, responsabiliza-se por gerar explicações para um dado erro ortográfico. Nos dois capítulos seguintes, fazemos apontamentos acerca de uma gramática de reconstituição para o português do Brasil abrangendo três domínios distintos: ortográfico (Capítulo IV), fonético-fonológico (Capítulo IV e V) e morfológico (Capítulo V). Apresentamos, por fim, um Capítulo VI, de conclusões e trabalhos futuros.

Capítulo II Correção de erros ortográficos em palavras isoladas



este capítulo, resumem-se os resultados de uma revisão da literatura relativa ao assunto central deste projeto, isto é, o aconselhamento ortográfico. Na exposição feita a seguir, os tópicos abordados foram priorizados não conforme a sua aplicabilidade neste projeto em específico, mas segundo a sua propriedade em compor o contexto em que este se insere, *principalmente de forma a apresentá-lo em contraste* e, assim, explicitar sua novidade. Dessa forma, questões genéricas de alto nível serão discutidas; e diferentes abordagens à solução do problema, apresentadas.

Embora as fontes consultadas tenham sido numerosas, grande parte do conteúdo a seguir baseia-se diretamente no cuidadoso, abrangente e instigante registro de Kukich (1992) do então estado-da-arte na área de "correção automática de palavras em textos", a qual engloba, com folga, a correção de palavras isoladas. Apesar da idade já considerável daquele documento, pouco foi ou pôde ser acrescentado à sistematização lá apresentada, pelo menos de acordo com o que se observa nos (poucos) trabalhos relacionados mais recentes (Zhao & Truemper, 99; Lins et al., 99; Pacheco, 96; Lucchesi & Kowaltowski, 93), que freqüentemente lhe fazem referência. Em tempo, boa parte das referências bibliográficas que fazemos neste capítulo são *apud* Kukich (1992).

II.1 Genealogia

Da profusão então vigente de estudos correlatos mas significativamente disjuntos, Kukich (1992) abstraiu, com propriedade, a área de correção automática de palavras em textos e a descreveu como tratando de três problemas básicos, em sucessão tanto histórica quanto de complexidade, a saber: (i) a **detecção de *não-vocábulos*[13]**, ou seja, a identificação de cadeias de caracteres que

[13] O termo "**não-vocábulo**" é uma tradução do inglês "*nonword*" usado em (Kukich, 92) e foi preferido a "erro ortográfico" porque este último, em sua acepção mais geral, pode implicar a consideração do contexto de enunciação. Um não-vocábulo é um caso especial de erro ortográfico.

não constam de um dado dicionário, léxico ou lista de cadeias válidas; (ii) a **correção de erros em palavras isoladas**, ou seja, a conversão de não-vocábulos em vocábulos, desconsiderando seu contexto de enunciação; e (iii) a **correção de palavras dependente de contexto**, que lida com erros ortográficos que não necessariamente envolvem o surgimento de um não-vocábulo.

Nesse nível mais alto de sistematização, um primeiro ponto digno de nota é a distinção entre **detecção** e **correção** de erros, havendo, no caso geral, um salto em complexidade considerável entre essas tarefas. Especificamente, técnicas eficientes já foram concebidas para detectar não-vocábulos, mas corrigir uma tal cadeia malformada constitui um problema bem mais complexo.

Em segundo lugar, vale notar que a tarefa de correção pode ser realizada em duas modalidades: **interativa**, caso enfocado neste projeto, ou não, caso de sistemas que corrigem textos de forma (quase) **completamente automática**, como, por exemplo, módulos de pós-processamento em sistemas de reconhecimento de texto (Srihari, 84; Jones et al., 91; entre outros). A diferença entre essas duas modalidades reside fundamentalmente na *verdadeira identidade do agente da correção*, que ora é o usuário, ora o sistema, respectivamente. Essa diferença não deve ser menosprezada, pois determina que aspecto de qualidade deve ser priorizado no projeto do sistema: ora *utilidade* como *colaborador* do usuário (o verdadeiro corretor), ora *precisão* como *corretor*.

Como a idéia de utilidade pareça ser de ordinário ignorada e, portanto, precisão seja freqüentemente o único parâmetro de qualidade considerado, a correção não-interativa é tradicionalmente considerada bem mais difícil, dados seus severos requisitos quanto à precisão da primeira (ou única) sugestão de correção dada. Justiça seja feita, no entanto, as conseqüências dos erros de um sistema não-interativo são, em princípio, mais sérias; e provavelmente é mais fácil maximizar utilidade do que precisão.

Feitas essas distinções básicas, este capítulo prossegue tratando da correção de erros em palavras isoladas. Como ambas as modalidades interativa e não-interativa serão cobertas e em nome de clareza de expressão, o termo "corretor ortográfico" será aqui usado no sentido de "corretor de palavras isoladas, *interativo ou não*". O conteúdo é abordado em três seções: na primeira, apresentam-se as grandes áreas de aplicação dos corretores ortográficos, e discute-se resumidamente a influência das características de uma aplicação em específico sobre o projeto

desses sistemas; na seção seguinte, trata-se em detalhe de *padrões de erro ortográfico*, talvez a característica de maior impacto dentre as citadas na seção precedente; e, na terceira e última seção, diferentes técnicas para corrigir palavras isoladas são descritas.

II.2 Diferentes aplicações e suas especificidades

As características de diferentes aplicações impõem restrições igualmente diferentes ao projeto de corretores ortográficos, e muitas técnicas bem-sucedidas foram talhadas sob medida para explorar as especificidades de suas aplicações. Vale, portanto, revisar algumas dessas aplicações antes de tornar ao detalhamento de diferentes técnicas.

Entre as aplicações mais estudadas, em empate apenas com a edição de textos, figura o **reconhecimento de texto** (Srihari, 84; Burr, 87; Goshtasby & Ehrich, 88; Ho et al., 91; Jones et al., 91). Esse interesse se explica pelo fato de que, embora bons dispositivos de reconhecimento de texto estejam disponíveis comercialmente, eles só apresentam desempenho ótimo sob condições ideais, que incluem texto nítido e impresso em algum tipo padrão. Além disso, mesmo uma precisão de reconhecimento de *caracteres* tão alta quanto 99% acaba por resultar numa precisão de reconhecimento de *palavras* de apenas 95%, uma vez que um erro a cada 100 caracteres corresponde aproximadamente a um erro a cada 20 palavras, considerando uma média de cinco caracteres por palavra.

Outras aplicações para as quais técnicas de correção ortográfica foram concebidas incluem **ambientes de programação** (Sidorov, 79; Spenke et al., 84), ***shells* de linha de comando** (Hawley, 82; Durham et al, 83), **interfaces de recuperação de informações/consulta a bases de dados** (Cherkassky et al., 90; Parsaye et al., 90), **interfaces em língua natural** (Veronis, 88a; Means, 88; Lee et al., 90; Deffner et al., 90), **ensino apoiado por computador** (Tenczar & Golden, 72)**, aprendizado de línguas apoiado por computador** (Contant & Brunelle, 92), **conversão texto-voz** (Kukich, 90; Tsao, 90; Kernigham, 91), **sistemas de apoio à comunicação de deficientes** (Wright & Newell, 91; Demasco & McCoy, 92), **interfaces baseadas em caneta** (Rhyne & Wolf, 91) e até mesmo a procura por formas antigas de palavras em *corpora* em inglês do século XVII (Robertson & Willet, 92).

A maioria das decisões de projeto motivadas pelas particularidades da aplicação em vista surge em resposta a questões relativas a três aspectos principais, a saber:

a) **léxico:** questões relativas ao léxico de um corretor ortográfico e sua construção incluem tamanho (número de entradas), cobertura (línguas e domínios de conhecimento abrangidos), taxa de entrada de novos termos e se todas as flexões e derivações de cada palavra figurarão como entradas distintas ou apenas formas canônicas/analisadas serão armazenadas, implicando algum tipo de processamento morfológico;

b) **interface usuário-computador**, incluindo considerações sobre se é exigida resposta em tempo-real; se o sistema pode solicitar informações ao usuário durante o processamento e, em caso positivo, que tipo de informação se pretende obter ou o usuário é capaz de fornecer; qual a precisão requerida; qual o usuário-alvo; etc. Vale notar que requisitos de precisão geralmente competem com os de tempo de resposta e, também por isso, são freqüentemente aliviados conforme o nível de interação com o usuário;

c) **padrões de erro ortográfico**, tratando, por exemplo, de quais são os erros mais comuns, quantos erros tendem a existir numa única palavra, se os erros tendem a mudar o comprimento das palavras originalmente pretendidas, qual a causa de um dado erro (tipográfica, cognitiva, fonética, etc.) e, de forma geral, se há regras, tendências probabilísticas ou heurísticas que podem modelar/caracterizar devidamente as possibilidades de erro.

Questões relativas a padrões de erro ortográfico são talvez as que tenham tido maior impacto sobre o projeto de corretores ortográficos. Neste projeto, por exemplo, o objetivo de todos os esforços em levantamento de conhecimento lingüístico pode ser resumido como a identificação/indução de padrões de erro úteis para a construção de um conselheiro ortográfico para o português. Dada a sua relevância, esse tópico é discutido na seção seguinte de forma mais detalhada.

II.3 Padrões de erro ortográfico

Padrões de erro ortográfico variam bastante em função da aplicação considerada. Por exemplo, erros de datilografia na transcrição de textos, que se devem sobretudo a deslizes de coordenação motora, tendem a refletir a proximidade entre as diversas teclas (considere-se, por exemplo, a substituição de “b” por “n”, a qual não tem o menor fundamento lingüístico ou cognitivo). Em contraste, erros cometidos por dispositivos de reconhecimento de caracteres são mais

provavelmente causados por confusão entre letras ou seqüências graficamente semelhantes (substituição, por exemplo, de “D” por “O”, “ri” por “n” ou “m” por “iii”). Numa análise mais sutil, mesmo dois modos de entrada de texto similares, como a transcrição de textos e sua redação propriamente dita (edição de *e-mail*, por exemplo), podem diferir sensivelmente quanto à freqüência e distribuição probabilística de erros, graças à maior carga cognitiva imposta pela segunda tarefa. Assim, um certo cuidado deve ser tomado na generalização de descobertas acerca de padrões de erro ortográfico.

II.3.1 Aspectos subjetivos

Quanto à sua causa, os erros ortográficos são por vezes dispostos na seguinte classificação: (i) **erros de digitação**, atribuídos a um deslize motor do autor/datilógrafo, que supostamente conhece a grafia correta da palavra pretendida; (ii) **erros cognitivos**, decorrentes de algum equívoco conceitual ou falta de conhecimento por parte do autor; e (iii) **erros fonéticos**, uma classe especial de erros cognitivos em que o autor substitui a grafia correta por uma cadeia diferente, mas de pronúncia idêntica ou muito próxima, dentro do que o autor entende do sistema ortográfico da língua.

Pode parecer surpreendente a pouca atenção dispensada a esse tipo de classificação pela literatura, que muitas vezes descreve técnicas genéricas para o tratamento de léxicos arbitrários e prefere considerar aspectos mais objetivos dos erros, de computação automática mais fácil (ou possível). Defende-se neste projeto, entretanto, a maior relevância dos erros cognitivos, mesmo os menos freqüentes, e seus aspectos subjetivos. Por esse motivo, pretende-se propor uma versão mais refinada da classificação acima, com a inclusão, entre outras, da classe dos **erros morfológicos**, outra vertente de erros cognitivos.

Erros fonéticos, contudo, às vezes recebem cuidado especial por parte de alguns trabalhos, que tratam de aplicações em que esse tipo de erro é notória e particularmente crítico. Esse é geralmente o caso de sistemas de informação ou consulta a bases de dados em que nomes próprios são usados como chave de busca, tais como cadastros de empresa (Boivie, 81; Oshika et al., 88), catálogos telefônicos/de serviços (Veronis, 88b) e até enciclopédias eletrônicas (van Berkel & De Smedt, 88). A preocupação com erros fonéticos foi a tônica de, por exemplo, (van

Berkel & De Smedt, 88), que empregou a análise de trigramas à transcrição fonética das palavras, em vez de à sua grafia propriamente dita[14]. Nesse estudo, os pesquisadores pediram a 10 sujeitos holandeses que transcrevessem a gravação em fita de 123 sobrenomes holandeses colhidos aleatoriamente numa lista telefônica e observaram que *38% das grafias produzidas estavam erradas apesar de serem foneticamente plausíveis*. Outro registro de estatísticas de erros fonéticos está presente em (Mitton, 87), que reporta que 44% dos erros em seu *corpus* de 925 dissertações estudantis envolviam homofonia.

II.3.2 Erros simples

Especificamente quanto a erros de ortografia gerados por humanos, uma das descobertas gerais mais antigas e festejadas foi feita por Damerau em 1964, tendo norteado o projeto de diversos sistemas desde então. Damerau (1964) constatou empiricamente que aproximadamente 80% de todos os erros ortográficos (em inglês) continham exatamente uma única instância de um dos quatro seguintes tipos de erro, ditos *simples*: **inserção**, **omissão** ou **transposição** (mudança de posição na palavra, como em "~~estrupo~~") de um caractere ou sua **substituição** por outro. Esse estudo forneceu, dessa forma, um *framework* simples e promissor com que tratar erros ortográficos, explicados pela composição de *erros simples*.

A regra dos 80%, entretanto, nem sempre se aplica: para aplicações específicas, erros ortográficos contendo *mais* de um erro simples foram observados em taxas variando de 6% (Pollock & Zamora, 84) a 31% (Mitton, 87). Além disso, vale observar que os erros cometidos por dispositivos de reconhecimento de caracteres não seguem os padrões observados para humanos. A maior parte dos erros, nesse caso, deve-se a substituições, uma significativa fração das quais envolvem seqüências inteiras de caracteres ("ri" por "n" ou "m" por "iii", por exemplo), segundo (Jones et al., 91). Ainda de acordo com esse estudo, os tipos de erro observados variam muito, não apenas de um dispositivo para outro, mas também dependendo das características tipográficas e da nitidez do texto de entrada, entre outros fatores.

[14] A técnica foi batizada pelos autores de "análise de trifones".

II.3.3 Comprimento das palavras

Outra descoberta geral, na verdade um corolário dos resultados de Damerau (1964), é a observação de que grande parte (80%) dos erros ortográficos dista, em comprimento, das respectivas correções de no máximo um caracter, a mais ou a menos. Isso levou muitos pesquisadores, especialmente na área de reconhecimento de texto, a particionar seus dicionários por comprimento de palavra para reduzir o tempo de busca.

Infelizmente, poucos dados estão disponíveis quanto à freqüência de erros por comprimento de palavra. Fato é que essa característica afeta o desempenho de uma técnica de correção, uma vez que erros em palavras curtas tendem a ser de correção mais difícil, em parte porque oferecem ao corretor um contexto intravocabular menos informativo. Para ilustrar esse problema, vale mencionar que um estudo de Pollock & Zamora (1983), tratando de um *corpus* de 50.000 não-vocábulos, relatou que "apesar de os erros ortográficos de comprimento 3 a 4 constituírem apenas 9,2% do *corpus*, eles geraram 42% das correções malsucedidas".

II.3.4 Erros na primeira letra

Existe uma crença generalizada de que poucos erros tendam a ocorrer na primeira letra de uma palavra. Poucos trabalhos documentam estatísticas acerca desse tipo de erro: Pollock & Zamora (1983) reportam uma taxa de 3,3%; Yannakoudakis & Fawthrop (1983a), 1,4%; e Mitton (1987), 7%. Em contraste com esses resultados relativamente baixos, Kukich (1992) observou que 15% dos erros num *corpus* de conversas transcritas (40.000 palavras) foram cometidos na letra inicial das palavras.

Ao se desconsiderarem erros na primeira letra, é possível particionar um léxico em 26 subconjuntos disjuntos, cada qual contendo todas as palavras iniciadas com uma mesma letra, e assim diminuir sensivelmente tempos de busca. Muitas técnicas de correção ortográfica já recorreram a essa possibilidade, não sem incorrer no risco de falhar completamente quando a correção pertinente não se encontra no subconjunto vasculhado.

II.3.5 Influência do teclado

Alguns estudos comportamentais abrangentes sobre datilografia e digitação foram realizados pelo *LNR Typing Research Group* (Gentner et al., 83), cujo objetivo, em vez de desenvolver uma técnica de correção ortográfica, era chegar a um modelo computacional de simulação do ato de digitar. Como parte desse trabalho, Grudin (1983) fez uma análise cuidadosa dos erros de digitação cometidos por seis datilógrafos experientes e oito iniciantes na transcrição de artigos de revista totalizando cerca de 60.000 caracteres em texto. Algumas de suas observações mais interessantes foram as seguintes:

- a maioria dos erros dos datilógrafos experientes consistiam em inserções resultantes da pressão simultânea de duas teclas adjacentes, enquanto a maioria dos erros dos iniciantes eram substituições;
- 58% de todos os erros de substituição envolviam teclas adjacentes; e
- mesmo após normalizar os dados pela freqüência de cada letra na língua considerada, a substituição de uma letra mais freqüente por uma vizinha menos freqüente era mais provável do que o contrário.

II.3.6 Regras heurísticas e tendências probabilísticas

Kukich (1992) cita três estudos abrangentes, todos para o inglês, que dedicaram esforços consideráveis à identificação de padrões em *corpora* de erros ortográficos com o intuito de fundamentar técnicas de correção ortográfica. Apesar de esses estudos terem muito em comum, as informações por eles geradas foram utilizadas para conceber e implementar três técnicas bastante distintas, a serem descritas na seção. Esta seção, por sua vez, contém uma breve revisão das descobertas desses trabalhos.

Yannakoudakis & Fawthrop (1983b) visavam descobrir **regras** específicas que os erros ortográficos tendem a obedecer, com o intuito de projetar um **algoritmo de correção ortográfica baseado em regras**. Nesse estudo, os autores compilaram um *corpus* de 1.377 erros, coletados de uma variedade de fontes, e descobriram que grande parte era coberta por um conjunto de 17

regras heurísticas, 12 das quais relativas ao uso de consoantes e vogais em *grafemas*[15] e 5 das quais relativas a *seqüenciação*[16]. Por exemplo, heurísticas relativas a grafemas incluem as seguintes: (i) a letra "h" é freqüentemente omitida nos grafemas "ch", "gh", "ph" e "rh", como nos erros "~~ag~~(h)~~ast~~" e "~~tec~~(h)~~niques~~"; (ii) é um erro comum duplicar ou, contrariamente, "unificar" consoantes que freqüentemente aparecem duplicadas; e (iii) é um erro comum substituir um grafema menos freqüente por um equivalente mais freqüente, como em "~~acquiesence~~" *vs.* "acquie**sc**ence". Heurísticas relativas a seqüenciação incluem: (iv) um erro ortográfico é mais freqüentemente um caracter *menor*, em comprimento, que sua respectiva forma correta, o que pode ser entendido como uma tendência à simplificação; (v) erros de digitação são causados pela pressão de uma tecla adjacente ou de duas teclas ao mesmo tempo; (vi) erros ortográficos curtos não contêm mais de um erro simples; e assim por diante.

Pollock & Zamora (1983) visavam descobrir **tendências probabilísticas**, tais como que letras e que posições numa palavra estão mais provavelmente implicadas em erros, com o intuito de projetar uma técnica de correção baseada em **chaves de similaridade**. Os autores extraíram mais de 50.000 erros de um total de 25 milhões de palavras em textos científicos e descobriram, entre outros, que: (i) 0,2% das palavras do *corpus* continham erros ortográficos; (ii) 94% desses erros eram constituídos de exatamente um erro simples; (iii) 34% dos erros eram omissões; (iv) 23% dos erros ocorriam na terceira letra das palavras; e (v), com exceção de uns poucos erros bastante freqüentes (substituição de "the" por "~~teh~~", por exemplo), a maior parte dos erros era raramente repetida.

Kernigham et al. (1990) visavam compilar **tabelas de probabilidade para cada um dos quatro tipos de erro simples**, com o intuito de **explorar essas probabilidades diretamente**. Usando o utilitário *spell* do *Unix* e uma técnica simples de geração de correções para testar todas as

---

[15] O **grafema** é a unidade ortográfica: uma seqüência a princípio não-analisável (i.e., cujo valor não corresponde à soma dos valores de suas partes) de letras que representa uma seqüência fônica. Por exemplo, na palavra "chás", têm-se os grafemas "ch", "á" e "s".

[16] Um erro de *seqüenciação* (tradução livre do inglês "sequence production") ocorre quando o autor sabe que caracteres aparecem em uma dada palavra, mas não em que ordem ("receive" *vs.* "~~recieve~~") ou quantas vezes ("transferred" *vs.* "~~transfered~~"). Naturalmente, esse tipo de erro é bem menos característico do português do que do inglês, motivo pelo qual os exemplos precedentes foram dados nesta língua.

possíveis palavras válidas formadas por exatamente uma inserção, omissão, transposição ou substituição operada sobre cada erro identificado, os autores varreram 44 milhões de palavras e levantaram automaticamente mais de 25.000 erros ortográficos para os quais apenas uma sugestão de correção era gerada. A lista resultante de 25.000 pares *(erro, forma correta)* foi então usada para compilar *matrizes de confusão* para cada tipo de erro simples. Por exemplo, eles determinaram que "s" foi erroneamente *inserido* após "e" 436 vezes, "t" foi erroneamente *omitido* após "i" 231 vezes, "a" foi erroneamente *substituído* por "e" 238 vezes, e "it" foi erroneamente digitada como "ti" (*transposição*) 48 vezes. Essas freqüências eram usadas para estimar a probabilidade de ocorrência de cada erro em potencial.

II.3.7 Listas de erros comuns

Outras fontes de dados sobre a natureza dos erros ortográficos são ainda listas publicadas de erros comuns e suas correções, dirigidas ao público em geral, falantes nativos da língua tratada inclusive. Exemplos desse tipo de lista podem ser encontrados em (Webster, 83), para o inglês, e (Faraco & Moura, 94, págs. 47-80) e (Sacconi, 92, págs. 34-39), para o português. Conforme registrado na literatura, muito poucos corretores ortográficos acadêmicos empregam esse recurso, uma das raras exceções sendo a técnica de Pollock & Zamora (1983), que realmente incorporava um passo de consulta a uma lista reduzida contendo palavras notoriamente problemáticas quanto à ortografia.

II.4 Técnicas

Na literatura, encontram-se diferentes taxonomias para as técnicas de correção ortográfica, cada qual evocando os aspectos distintivos considerados relevantes em um trabalho específico. Daelemans et al. (1984)[17], por exemplo, distinguem entre **técnicas estatísticas** e **lingüísticas**, ou seja, que empregam análise estatística e conhecimento lingüístico, respectivamente. Uma distinção mais interessante é feita por Pollock & Zamora (1984) entre **técnicas absolutas** e **relativas**, diferindo na forma de obtenção dos vocábulos candidatos a correção: respectivamente *ou* pela aplicação de operações de transformação (por exemplo, inserções, substituições, etc.) ao não-vocábulo, *ou* a partir da determinação do conjunto dos vocábulos mais parecidos com o não-

[17] *apud* (van Berkel & DeSmedt, 88).

vocábulo dentre todos os constantes do léxico, segundo alguma medida de similaridade[18]. Os termos propriamente ditos cunhados por Pollock e Zamora, pouco intuitivos, não são muito populares, ao contrário dos conceitos denotados: a classe das **técnicas reversas** (em que se tenta "reverter" o erro cometido) é citada com certa freqüência e parece coincidir com a das técnicas absolutas. Essa distinção é tão interessante que a promovemos a divisor paradigmático, postulando dois grandes **paradigmas de correção** mutuamente excludentes e ortogonais às demais classificações: o relativo e o absoluto/reverso.

No entanto, é Kukich (1992) quem propõe a taxonomia que melhor se presta a racionalizar uma revisão abrangente da totalidade das técnicas de correção ortográfica, qual era, afinal, o seu objetivo no referido trabalho. Essa pesquisadora enfoca a ferramenta conceitual básica subjacente a cada técnica, identificando as seguintes: (i) **mínima distância de edição**, (ii) **chaves de similaridade**, (iii) **regras**, (iv) **análise de n-gramas**, (v) **redes neurais** e (vi) **estimativas de probabilidade** (técnicas probabilísticas). Como o objetivo desta seção coincide com o de Kukich (em parte), sua taxonomia será aqui adotada e refletirá diretamente na divisão do texto em subseções, cada qual abordando técnicas baseadas em uma das ferramentas acima enumeradas. Vale observar que todos os conselheiros ortográficos acadêmicos para o português (Lins et al., 99; Pacheco, 96; Almeida & Pinto, 95; Lucchesi & Kowaltowski, 93) podem ser enquadrados como recorrendo a técnicas de mínima distância de edição, sendo, portanto, comentados na subseção correspondente.

A Tabela III apresenta um cruzamento da taxonomia de (Kukich, 1992) com as demais. Nessa tabela, um dado quadrante assinalado significa que as duas classes envolvidas se intersectam. Por exemplo, as técnicas baseadas em chaves de similaridade são também relativas e podem ser tanto estatísticas quanto lingüísticas, a um só tempo inclusive, o que é explicado pela possibilidade de consideração, no projeto de chaves de similaridade, tanto de resultados inferidos estatisticamente quanto de conhecimentos lingüísticos. Por outro lado, dificilmente uma técnica em específico será tanto absoluta quanto relativa, mas ambas as classes em questão são intersectadas pela das

[18] Outra forma de entender essa distinção é associar respectivamente técnicas absolutas e relativas a casamento *exato* e *aproximado* de padrões.

técnicas de mínima distância de edição, o que se explica por esta última incluir técnicas que diferem na forma de obtenção de candidatos a correção.

Tabela III: Interseção entre a taxonomia de Kukich (92) e as demais.

	autor:	(Daelemans et al., 84)		(Pollock & Zamora, 84)	
	classe:	estatísticas	lingüísticas	absolutas (reversas)	relativas
(Kukich, 1992)	mín. distânica de edição			✓	✓
	chaves de similaridade	✓	✓		✓
	regras	✓	✓	✓	✓
	n-gramas				✓
	redes neurais	✓	✓		✓
	técnicas probabilísticas	✓		✓	✓

II.4.1 Mínima distância de edição

Os algoritmos de correção ortográfica mais estudados são, de longe, os que computam a **mínima distância de edição** (MDE) entre um não-vocábulo e as entradas de um léxico. Essa distância, que pode ser medida entre duas cadeias de caracteres quaisquer dadas, foi definida por Wagner (1974) como o número mínimo de operações de edição (inserções, eliminações e substituições) necessárias para transformar uma cadeia na outra. Antes de sua definição por Wagner, no entanto, esse conceito já tinha sido empregado em correção ortográfica pelos pioneiros Damerau (1964) e Levenshtein (1966). Mais tarde, surgiram trabalhos, como (Veronis, 88a), que estenderam o conceito para melhor tratar erros fonéticos, que tendiam a distar das respectivas formas corretas mais do que o desejável.

Em geral, algoritmos de correção baseados em MDE requerem a comparação de um dado não-vocábulo com cada uma das entradas do léxico, ou seja, têm complexidade de tempo linear no tamanho do léxico, o que é proibitivo em muitos casos. Em resposta a isso, alguns expedientes têm sido concebidos. Por exemplo, com base na hipótese de que omissões simples eram o tipo de erro mais comum, Mor & Fraenkel (1982) obtiveram ganho em desempenho ao adicionar a seu léxico, implementado como uma tabela *hash*, todas as variações de cada palavra obtidas por

exatamente uma omissão (naturalmente, cada entrada do léxico incluía algum tipo de informação de validação). Outros pesquisadores exploraram as possibilidades oferecidas por certos tipos de representação do léxico, como *tries* (Dunlavey, 81), estruturas que podem ser entendidas como uma implementação de autômatos determinísticos acíclicos cujos diagramas de estados são árvores, estritamente.

Cabe ainda citar a vertente bastante diversa das *técnicas reversas de MDE* (Church & Gale, 91; Kernigham et al., 90; entre outros), cujo passo de geração de alternativas de correção consiste basicamente em (i) gerar todas as variações de um dado não-vocábulo obtidas por exatamente uma inserção, uma omissão, uma transposição ou uma substituição (ou seja, pela *reversão* de um erro simples) e (ii) eliminar todas as cadeias assim obtidas que não constem do léxico. No processamento desse passo, dado um não-vocábulo de comprimento n e um alfabeto de 26 caracteres, o número de cadeias cuja presença no léxico deve ser verificada é $26(n + 1)$ inserções + n omissões + $25n$ substituições + $(n – 1)$ transposições = $53n + 25$ cadeias.

Pode-se dizer que três dos quatro **conselheiros ortográficos para o português descritos na literatura** (Pacheco, 96; Almeida & Pinto, 95; Lucchesi & Kowaltowski, 93) empregam tão-somente técnicas reversas de MDE, às vezes relaxadas para tentar a reversão de mais de um erro simples por não-vocábulo e sempre estendidas para também tentar a substituição de grafemas freqüentemente envolvidos em erros fonéticos por outros foneticamente equivalentes.

Pacheco (1996) e Lucchesi & Kowaltowski (1993) exploram a representação do léxico em um autômato finito determinístico acíclico mínimo, extremamente conveniente dadas sua alta taxa de compactação e eficiência em tempo de busca, de complexidade linear no tamanho da cadeia procurada. Isso permite a Pacheco verificar todas as $53n + 25 + x$ cadeias obtidas pela reversão de um erro simples ou *substituição fonética simples* (a que se deve o x da expressão acima). Em contraste, o algoritmo de Lucchesi & Kowaltowski explora um pouco mais a fundo sua estrutura de dados, gerando mutações *enquanto* percorre o autômato e, assim, considerando apenas as transições válidas a partir do estado corrente. Esse expediente aborta prematuramente a geração de outros não-vocábulos e permite ao sistema se dar ao luxo de reverter todas as substituições fonéticas de um não-vocábulo.

É sabido que o sistema de Almeida & Pinto (1995) realiza algum tipo de análise morfológica,

uma vez que seu léxico é implementado como uma tabela *hash* em que se realizam buscas pelos *radicais das palavras*, ou seja, cada entrada dessa tabela está associada biunivocamente a um radical e contém *flags* indicadores de que afixos podem ser aplicados a esse radical, agrupando, dessa forma, todas as palavras que dele compartilham. No entanto, parece que a análise morfológica só é realizada quando se verifica a pertinência de uma cadeia ao léxico (e conseqüentemente se recuperam as informações associadas a essa cadeia), não tendo, pois, efeito sobre a *geração* de alternativas de correção, apenas sobre sua *validação*.

Nenhum registro foi encontrado do critério, talvez inexistente, de classificação das alternativas de correção geradas por esses sistemas.

O trabalho de Lins et al. (1999) é também baseado numa técnica reversa e constitui exceção pelas seguintes características:

- **limitação severa das operações de transformação usadas na geração de candidatos a correção**, as quais se restringem à (i) eliminação da letra *h* inicial, (ii) o acréscimo dessa letra como inicial em cadeias iniciadas por vogal e (iii) a substituição usual de grafemas freqüentemente envolvidos em erros fonéticos. É exata e somente nesse ponto que a técnica usada nesse trabalho diverge do paradigma reverso de MDE; e

- **presença de considerações sobre a classificação desses candidatos**, as quais se resumem à especificação do sentido, ao longo do não-vocábulo, em que cada par de substituição vai sendo tentado, ou seja, se partindo do início para o final ou vice-versa. Por exemplo, as substituições e↔i, g↔j e s↔z são feitas partindo do final do não-vocábulo para o início, porque, segundo os autores, a opção por um dos grafemas de cada par costuma causar dúvidas quando ocorre no final dos vocábulos, exceto no caso específico dos prefixos “ante-” e “anti-”, que é tratado com prioridade. As demais substituições são tentadas a partir do início.

II.4.2 Chaves de similaridade

A técnica de **chaves de similaridade** consiste em ordenar os itens do léxico (em nome da clareza, "palavras") em um índice não segundo a ordem lexicográfica das palavras propriamente ditas, mas segundo algum tipo de ordem, geralmente lexicográfica, *definida entre suas respectivas chaves de similaridade*. A chave de similaridade de uma cadeia de caracteres, por sua vez, é um dado de qualquer tipo, geralmente também uma outra cadeia, computado de forma simples por uma função geralmente não-injetora. Os dois requisitos básicos no projeto dessa função e da ordem a ser estabelecida entre as chaves de similaridade são (i) eficiência nas buscas no índice e (ii) garantir que cadeias ortograficamente similares, ou entre as quais é provável que o usuário se confunda, tenham chaves idênticas ou similares, de forma que estejam (ou estivessem, no caso de não-vocábulos) próximas no índice. Esses requisitos sendo bem atendidos, bons candidatos à correção de um não-vocábulo deverão ser encontrados ao se computar a chave de similaridade desse não-vocábulo, localizar-se um ponto no índice onde uma palavra com essa chave poderia ser inserida (talvez no meio da seqüência, se houver, de palavras que tenham essa mesma chave) e tomarem-se candidatos a partir desse ponto, em ambas as direções, até um possível limite máximo de distância. A lista de candidatos assim obtida pode ser usada como tal ou ainda processada por algum passo final de classificação/seleção, como no sistema *SPEEDCOP* (Pollock & Zamora, 84), que em seguida selecionava o primeiro candidato que diferia do não-vocábulo de apenas um erro simples.

O *SPEEDCOP* usava, na verdade, duas chaves de similaridade, ditas *estrutural* e *de omissão* (tradução dos originais "*skeleton key*" e "*omission key*", respectivamente) e cuidadosamente talhadas, a partir de descobertas estatísticas, para atingir máximo poder de correção no seu domínio de aplicação (textos científicos em inglês), que apresentava baixíssima taxa de erros fonéticos. Ambas as chaves eram obtidas simplesmente por um arranjo, em ordem específica, do conjunto (sem elementos repetidos) de todas as letras que ocorriam em uma dada cadeia. Simples que possa parecer, essa única diferença entre os dois tipos de chave — a *ordem de construção* — tem efeitos drásticos, o que fica mais claro quando se nota que a ordenação entre chaves era lexicográfica e lembra-se que, em índices como aqueles, tão maior é a distância entre os destinos da busca por duas chaves distintas quanto mais inicial for a posição do primeiro caracter em que elas diferem.

Na chave estrutural, a ordem de construção era a seguinte: (1º) primeira letra da cadeia, (2º) consoantes, em ordem de ocorrência, e, por fim, (3º) vogais, em ordem de ocorrência. Essa fórmula foi motivada pela inferência estatística de dois resultados, a saber: (i) que era bem mais provável errar ou omitir as vogais de uma palavra do que suas consoantes ou, o que era menos provável ainda, sua letra inicial; e (ii) que a ordem de ocorrência das consoantes, sobretudo, e vogais tendia a ser preservada nos erros ortográficos.

É fácil perceber que a chave estrutural apresentava desempenho muito ruim nos casos de omissão de consoantes, principalmente à medida que o erro se aproximava do início da cadeia, erros na primeira letra, consoante ou não, sendo absolutamente desastrosos. Por esse motivo, o segundo tipo de chave — a chave de omissão — foi concebido, cuja ordem de construção era a seguinte: (1º) consoantes, em ordem inversa de probabilidade de omissão[19], e (2º) vogais, em ordem de ocorrência. Ou seja, apareciam primeiro as letras cuja omissão era menos freqüente, com o que se tentava evitar que a busca por candidatos de correção fosse desviada para áreas menos pertinentes do índice; e, mais uma vez, as vogais eram colocadas em último lugar, o que demonstra a crença dos projetistas em que estas constituíssem a porção mais "frágil" das palavras. O sistema *SPEEDCOP* só recorria à chave de omissão se a aplicação da chave estrutural não fosse bem-sucedida.

Outros trabalhos que recorrem à técnica de chaves de similaridade foram desenvolvidos, por exemplo, por Bocast (1991), Tenczar & Golden (1972) e Odell & Russell (1918).

A técnica de chaves de similaridade é um exemplo de técnica relativa que, na determinação de uma boa alternativa de correção, evita medir a similaridade entre cada vocábulo do léxico e o não-vocábulo a ser corrigido. O uso de chaves de similaridade pode ainda ser considerado uma variação da clássica técnica dos *n vizinhos mais próximos* em que o espaço de características é, na verdade, uma reta, contando com um eixo único, portanto, sobre o qual se distribuem as chaves de similaridade, cada qual um ponto por si só. Nessa analogia, a única discrepância é o fato de que chaves de similaridade idênticas, cada qual correspondente a uma palavra diferente,

---

[19] De acordo com as inferências de Pollock & Zamora (1984), a ordem de probabilidade de omissão é a seguinte:

RSTNLCHDPGMFBYWVZXQK.

distribuem-se contiguamente, numa configuração mais propriamente de vizinhança do que de coincidência num ponto único, como seria se a analogia do eixo valesse perfeitamente.

II.4.3 Regras

Técnicas baseadas em regras envolvem a representação de conhecimento acerca de padrões de erro na forma de regras de reversão de erros ortográficos. O processo de geração de alternativas de correção, dessa forma, consiste basicamente em aplicar todas as regras possíveis a um não-vocábulo e reter apenas as formas resultantes que constituam vocábulos válidos. Em seguida, cada alternativa é geralmente classificada segundo uma pontuação calculada a partir de índices predefinidos de probabilidade associados às regras aplicadas na geração da alternativa em questão. Esses índices são estimativas da probabilidade de ocorrência do tipo de erro coberto por cada regra.

Yannakoudakis & Fawthorp (1983) desenvolveram, mais ou menos nesses moldes e para o inglês, um sistema de correção ortográfica baseado em conhecimento a partir do conjunto de regras[20] que inferiram na análise de um *corpus* de 1.377 erros ortográficos. Como algumas de suas regras incorporavam conhecimento relativo ao comprimento mais provável da melhor alternativa de correção, seu léxico foi particionado em muitos subconjuntos de acordo com comprimento de palavra e primeira letra. O processo de geração de candidatos não era reverso: subconjuntos específicos do dicionário eram varridos em busca de vocábulos (i) que diferissem dos não-vocábulos de uma ou duas ocorrências dos tipos de erro tratados e (ii) cuja pertinência como correção pudesse ser explicada por pelo menos uma das regras. Nos testes realizados nesse estudo, que usavam o *corpus* inteiro como conjunto de teste, a correção pertinente se encontrava no subconjunto varrido em 1.153 dos casos, ou 75% das vezes. Nesses 1.153 casos, ela era também retornada como primeira alternativa 90% das vezes, resultando numa precisão geral de 68% (90% de 75%).

Em outro projeto, um sistema especialista de correção ortográfica baseado em regras foi desenvolvido para o inglês por Means (1988) tratando de um domínio com alta incidência de abreviações, siglas e jargão. Seu processo de geração tentava reverter erros recorrendo a três

[20] Exemplos das regras empregadas podem ser encontrados na Seção II.3.6.

expedientes distintos, em ordem decrescente de prioridade, a saber: (i) aplicação de um conjunto de regras morfológicas relativas a erros comuns de afixação, tais como não dobrar a consoante que precede o sufixo "-ing", nos casos em que isso é necessário; (ii) aplicação de regras de expansão de abreviações; e (iii), se tudo mais falhasse, geração de todas as variações obtidas pela reversão de um erro simples, tentando-se inclusive suprimir um espaço em branco.

II.4.4 Análise de n-gramas

Técnicas baseadas em **análise de n-gramas** são eminentemente relativas (não-reversas), ou seja, apostam na definição de uma boa medida de similaridade entre cadeias de caracteres e elegem candidatos a correção varrendo o léxico (ou um seu subconjunto específico) em busca dos vocábulos mais parecidos com um dado não-vocábulo. O que diferencia a classe das técnicas baseadas em análise de n-gramas é exatamente que características das cadeias serão considerados para medir sua similaridade, a saber: seus n-gramas!

Qualquer subcadeia de comprimento *n* de uma dada cadeia de caracteres é dito um **n-grama** dessa cadeia. Por exemplo, AMAR contém os seguintes trigramas: #AM, AMA, MAR e AR#. Vale notar, nesse exemplo, o artifício útil do caracter delimitador especial, aqui denotado por "#", que permite que as técnicas baseadas em *n-gramas* sejam sensíveis ao fato, obviamente relevante, de que uma dada seqüência de caracteres inicia ou termina uma cadeia ou palavra.

Exemplos de medidas de similaridade baseadas em n-gramas bastante intuitivas são as funções $2(c/(n + n'))$ e $(c/máx(n, n'))$, onde *c* é o numero de n-gramas em comum entre as duas cadeias consideradas, e *n* e *n'* são os seus respectivos comprimentos. Como era de se esperar, ambas as funções crescem com o número de n-gramas em comum e requerem que comprimentos maiores sejam compensados por mais n-gramas em comum. Ambas as funções provêm de (Angel et al., 83), a segunda sendo proposta para melhor explorar a tendência de os comprimentos de um não-vocábulo e sua respectiva correção diferirem de uma unidade, no máximo.

Distâncias usuais entre vetores — distância de Hamming, produto escalar, distância-cosseno, etc. — podem também ser usadas para medir a (dis)similaridade entre cadeias de caracteres. Para isso, as cadeias devem ser devidamente projetadas em algum espaço de características,

$(x^n + 2x^{n-1})$ de cujas dimensões podem, por exemplo, ser dedicadas aos possíveis *n*-gramas dentro do fechamento[21] de um alfabeto de cardinalidade *x* (a parcela $2x^{n-1}$ se deve aos n-gramas iniciados ou terminados em "#", não contado como elemento do alfabeto). Num tal esquema, a projeção de uma cadeia envolveria atribuir valores, talvez booleanos, a essas dimensões em função do conjunto dos n-gramas observados numa cadeia em específico.

Van Berkel & De Smedt (1988) realizaram um trabalho interessante tanto pela medida de distância (dissimilaridade) empregada quanto pelo fato já mencionado de terem aplicado a análise de trigramas à transcrição fonética das palavras, em vez de à sua grafia, como é usual. Esses "trigramas fonéticos" foram muito propriamente batizados pelos autores como *trifones*. Os autores partiram (i) da concepção de um alfabeto fonético suficientemente grosseiro, de forma que variantes (ou erros) freqüentes de pronúncia tivessem representações idênticas ou muito similares, (ii) do levantamento da freqüência, na língua considerada (o holandês), de cada possível trifone dentro do fechamento do alfabeto fonético concebido e (iii) da implementação de um *arquivo invertido* que permitia obter eficientemente todas as palavras do léxico que continham um dado trifone. Esses recursos eram explorados no seguinte algoritmo de correção de erros fonéticos:

i. dado um não-vocábulo *NV*, transcreva foneticamente todas as suas variantes de pronúncia;
ii. para cada variante de pronúncia V faça:
    ii.i. divida *V* em trifones;
    ii.ii. atribua a cada trifone um *fator de distância* igual à sua freqüência. Normalize esses fatores de distância (de forma que reflitam, de forma absoluta, as proporções entre as freqüências envolvidas);
    ii.iii. fazendo consultas ao arquivo invertido, levante todas as palavras que contenham pelo menos um dos *trifones seletivos* (isto é, aqueles cuja freqüência esteja *abaixo* de um limiar predefinido) presentes em *V*;

[21] O fechamento de um alfabeto é o conjunto de todas as cadeias de qualquer comprimento, inclusive zero, que podem ser formadas com os caracteres desse alfabeto.

ii.iv. para cada candidato a correção *C* assim obtido, compute sua distância a *NV* como a soma dos fatores de distância associados aos trifones em comum entre *C* e *NV*;

iii. classifique todos os candidatos obtidos em ordem crescente de distância.

No desenvolvimento desse trabalho, van Berkel & De Smedt (1988) se inspiraram no sistema *FUZZIE*, projetado por De Heer (1982).

II.4.5 Redes Neurais

O uso de redes neurais (Braga et al., 00) na correção ortográfica constitui um caso curioso de técnica relativa, que se destaca pelo fato de sua medida de similaridade ser definida implícita e automaticamente a partir de conjuntos de treinamento, compostos de pares *(entrada, reposta esperada)* devidamente codificados. A fase de treinamento pode ser entendida, dessa forma, como uma fase de busca pela melhor entre as possíveis definições de similaridade. Claro que nem tudo é tão simples assim: cabe ao projetista, entre outras, a tarefa crucial de eleger que características das cadeias poderão ser levadas em consideração pelas medidas, assim reduzindo, talvez demais, o espaço de possibilidades. Essa tarefa é conhecida como *extração de características*; e, entre as características geralmente "extraídas", figuram os n-gramas (especialmente para $n \leq 3$).

As redes neurais mais utilizadas na correção ortográfica e em geral são as MLP (*Multilayer Perceptron*) (Braga et al., 00), que apresentam uma topologia organizada em camadas sucessivas. Cada neurônio da camada *i* alimenta todos os neurônios da camada *i+1*, os quais, por sua vez, computam sua saída a partir dos estímulos que recebem, ponderados de forma independente. Tais pesos são ajustados automaticamente na fase de treinamento de uma rede MLP, que consiste na aplicação de um algoritmo específico de propagação de erro (*backpropagation*) que (provavelmente) encontra um conjunto ótimo de pesos. A primeira camada de uma rede MLP, dita *de entrada*, apenas *propaga* uma codificação numérica ou booleana das características de relevância do ambiente; por sua vez, a última camada, dita *de saída*, *computa* uma codificação da resposta da rede ao estímulo recebido.

Na literatura sobre a aplicação de MPLs à correção ortográfica, o projeto da camada de saída é consenso: sempre há um neurônio para cada entrada no léxico e uma correspondência biunívoca *a priori* entre neurônios e vocábulos. Assim, a saída da rede é interpretada da seguinte forma:

quanto mais alta é a saída de um neurônio da última camada, tanto melhor é o vocábulo correspondente como alternativa de correção.

Por outro lado, os trabalhos diferem muito no projeto da camada de entrada, ou seja, quanto ao resultado da extração de características. Kukich (1988), por exemplo, usou 450 neurônios na primeira camada, codificando, sem perda de informação, cadeias de até 15 caracteres sobre um alfabeto de 30 letras. Dada uma cadeia, o esquema de codificação era o seguinte: primeiro, todas as entradas eram zeradas; em seguida, o primeiro caracter da cadeia era codificado no primeiro bloco de 30 neurônios, ao se excitar apenas um deles, cuja ordem no bloco estava associada biunivocamente à letra codificada; então, o segundo caracter era analogamente codificado no segundo bloco de 30 neurônios; e o processo se repetia até o fim da cadeia.

Cherkassky & Vassilas (1989a, 1989b) elegeram ora unigramas, ora bigramas como as características a serem codificadas. Na camada de entrada, havia um neurônio para cada possível n-grama (n = 1 ou 2) dentro do fechamento do alfabeto utilizado. Nesse esquema, uma dada cadeia era apresentada à rede ao se estimularem apenas os neurônios correspondentes aos n-gramas observados na cadeia em questão.

Por fim, vale notar a diversidade das características extraídas e codificadas por Deffner et al. (1990), compreendendo características de natureza "n-grâmica", fonética, sintática e até semântica. Exemplos desses dois últimos tipos de característica, bem pouco usuais, eram as booleanas *é_adjetivo* e *é_cor*, respectivamente.

II.4.6 Técnicas probabilísticas

Historicamente, as técnicas baseadas em n-gramas levaram naturalmente às assim chamadas técnicas probabilísticas, nas áreas tanto de reconhecimento quanto de edição de texto. Essas técnicas têm se caracterizado pelo recurso a pelo menos um dos seguintes tipos de probabilidades:

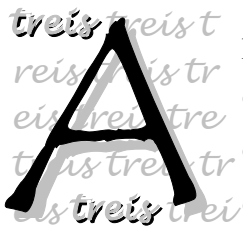
- **probabilidades de transição**, que tratam da probabilidade de um dado caracter (ou seqüência de caracteres) ser seguido por outro dado caracter. Probabilidades desse tipo variam de língua para língua e podem ser estimadas pela análise de *corpora* corretos, não necessariamente anotados;

- **probabilidades de confusão**, que tratam da probabilidade de um dado caractere ser erroneamente substituído por outro dado caractere. Probabilidades desse tipo variam de acordo com a fonte do texto (por exemplo: digitação ou reconhecimento, usuários alfabetizados ou semialfabetizados) e podem ser estimadas a partir de *corpora* contendo erros e anotados quanto à correção desses erros. No caso de textos digitados, por exemplo, pode-se explorar também a proximidade entre teclas.

Várias técnicas probabilísticas podem ser classificadas como relativas, a medida de similaridade sendo definida, nesse caso, como *P(Candidato|Não-vocábulo)*, ou seja, a probabilidade de *Candidato* ser a palavra pretendida dado que *Não-vocábulo* foi observado. Um objetivo freqüente das técnicas probabilísticas relativas é evitar o cálculo da probabilidade condicional de cada entrada do léxico na determinação da mais provável, dado um não-vocábulo. Shinghal & Toussaint (1979b), por exemplo, usavam uma chave de similaridade numérica computada em função das probabilidades das transições observadas numa dada cadeia. As soluções mais freqüentes, no entanto, baseiam-se num algoritmo de programação dinâmica (Nemhauser, 66) chamado *algoritmo de Viterbi* (Forney, 73). Nesse algoritmo, um dígrafo é usado para representar tanto a estrutura do léxico (probabilidades de transição) quanto as características da fonte de texto (probabilidades de confusão) e é percorrido de forma eficiente para encontrar a cadeia de máxima probabilidade. Exemplos de trabalhos inspirados no algoritmo de Viterbi são (Shinghal & Toussaint, 1979a) e (Srihari et al., 1983). Este último trabalho adaptou o algoritmo para percorrer *tries* (Knuth, 73), em resposta ao problema, freqüente em técnicas probabilísticas que dispensam o léxico em tempo de execução, de que nem sempre a cadeia mais provável é um vocábulo válido.

Técnicas probabilísticas também existem na modalidade reversa. Os sistemas de Kahan et al. (1987) e Goshtasby & Ehrich (1988) geravam alternativas de correção considerando, respectivamente, probabilidades de confusão e transição. Esses trabalhos, entretanto, usavam técnicas bem distintas.

Capítulo III Da Origem e fim dos não-vocábulos — Breve reflexão filosófico-metodológica



pesar de seu escopo limitado no auxílio à edição de textos, um bom conselheiro ortográfico é, como ferramenta isolada ou integrada a outras mais complexas, muito bem-vindo nesse cenário, podendo até prestar benefícios educacionais. Contudo, de acordo com o observado no desempenho de alguns sistemas conceituados e em publicações científicas relacionadas, o aconselhamento ortográfico para o português tem sistematicamente subaproveitado boas oportunidades de colaboração com o usuário. Em específico, parece que muitas das dificuldades usuais e já catalogadas do falante nativo do português têm sido negligenciadas ou ignoradas, muito embora expliquem erros de difícil recuperação por parte do usuário. Ou seja, os sistemas falham quando o usuário precisa de uma boa sugestão e esta poderia ser dada, motivo por que, no presente projeto, tais sistemas são considerados *pouco úteis*. Não é por acaso também que o Capítulo II pareça leigo em"utilidade": é que realmente nada foi encontrado de correlato na literatura.

Deste ponto em diante, devidamente motivados e contextualizados, partimos em busca da utilidade via embasamento lingüístico. Via de regra, não haverá mais capítulos dedicados exclusivamente à revisão bibliográfica: no máximo, subseções, próximas do ponto de aplicação do material nelas contido. Visamos, assim, diminuir a densidade do texto e aliviar a carga cognitiva sobre o leitor.

Neste capítulo em específico, trataremos de alguns requisitos circunstanciais adicionais, tomaremos decisões metodológicas fundamentais, elegeremos axiomas e discutiremos as implicações disso tudo para o desenvolvimento do projeto. Aqui o problema sofrerá um primeiro (e crucial) passo de análise/decomposição, de forma que os subproblemas resultantes possam ser atacados nos capítulos seguintes.

Argumento de utilidade

Nossa primeira grande hipótese, *que não chegaremos a comprovar neste trabalho*, concerne exatamente à relação entre os nossos dois "reagentes-título": que seja possível maximizar utilidade por meio de nossa modelagem psicolingüística. Vai até parecer que deixamos "utilidade" de lado. No entanto, a verdade é que se trata de um conceito melindroso, que pretendemos grandeza e que, assim considerado, escapa à mensuração trivial. Resta-nos tentar estimá-la e, para tanto, não podemos esquecer o seguinte: o lixo de uns é o tesouro de outros[22].

Ou seja, utilidade é assim como gosto, uma questão pessoal; e toda e qualquer estimativa de utilidade presume um **perfil de usuário**. Para demonstrar isso de forma cabal, bastam alguns exemplos de análise reversa de utilidade (como à página 5): dado "patinho" como intenção original para "~~patio~~", quão difícil seria a recuperação do erro, sem ajuda? Para um usuário como nosso provável leitor — um *sujeito ideal*[23] — seria trivial, e tratar-se-ia obviamente de um lapso de digitação; mas Cagliari (92) aponta este como um erro comum em crianças em processo de alfabetização, que costumam fazer transcrição fonética, deixar a nasalidade implícita e, portanto, não ser capazes de corrigi-lo sozinhas. Que tal agora "brega" como intenção para "~~preca~~"? Uma suposição absolutamente absurda para o sujeito ideal; mas plausível, e difícil, para as mesmas crianças (Cagliari, 92), que muitas vezes grafam as palavras enquanto as sussurram, e ainda para aquelas pessoas que não pronunciam fones sonoros, para as quais as letras "p"/"c" e "b"/"g" acabam por ter o mesmo valor, o de "p"/"c". O que seria mais útil como sugestão de correção a "~~botano~~": "butano" ou "botando"? Mais uma vez, depende: o sujeito ideal pode até pronunciar "botando" como "~~botano~~", o que tem motivação fonética flagrante[24]; mas, diferentemente de *muitos* outros sujeitos, já tem estabelecida uma relação entre ortografia e morfologia que

---

[22] Tradução livre do inglês "*one man's meat is another man's poison*".

[23] Sem bajulação, já que mesmo dos sujeitos ideais se espera uma miríade de erros, mas também uma familiaridade com certos fatos ortográficos e nenhuma deficiência articulatória, perceptiva ou cognitiva séria.

[24] A única diferença entre os fones [n/m] e [d/b] é a nasalidade.

virtualmente o incapacita para esse erro[25], a não ser como um deslize de pronta recuperação. Assim, temos aqui uma completa inversão de valores: a utilidade de "botando" flutua de um extremo a outro, enquanto a cotação de "butano" se mantém relativamente estável. Por fim, seria "(ele) penou" útil como sugestão a "~~peneu~~"? Jamais, *mas exclusivamente porque estamos desconsiderando perfis de usuário com graves problemas mentais ou que atribuam a certos grafemas valores impossíveis*.

Os exemplos acima têm algo ainda mais interessante a revelar: é notável como a utilidade das sugestões, à primeira vista duvidosa, passa a ficar patente ante os argumentos oferecidos. Em outras palavras, demonstramos utilidade, de algum modo, informal que seja, mas convincente; e, mesmo que a utilidade não seja diretamente apreensível, pelo menos encontramos algo, um tipo de argumento, com que parece ter correlação. Urge encontrar uma fórmula para o **argumento de utilidade**. Analisando esses e outros exemplos, supomos que seus **componentes** básicos sejam dois, a saber:

- **perfil de usuário**, nosso ponto de partida, representando os pontos fortes e fracos do usuário, sobretudo *o que ele não sabe* e *o que ele tende a fazer*. Contudo, como se pode perceber nos exemplos, o que convence mesmo não é uma "foto" da mente do usuário, mas, antes, algum tipo de

- explicação ou **reconstituição** (implícita que seja e como foi, nos exemplos) de como o usuário, segundo sua natureza (perfil), teria procedido para "perpetrar" o não-vocábulo resultante quando, hipoteticamente, pretendia produzir a sugestão defendida pelo argumento.

---

[25] Grafar sistematicamente "~~botano~~" por "botando" é um tipo de erro muito interessante: assumir que o usuário o teria cometido tem graves implicações para um modelo de sua competência ortográfica. Por exemplo, passa a ser provável a grafia de "botar" como "~~botá~~"/"bota", "cantaram" como "~~cantaro~~" e até "falação" como "~~falasão~~".

Entretanto, é curioso notar que a relação fonética entre "~~botano~~" e "botando" é a mesma que a existente entre "~~sambambaia~~" e "samambaia" (vide Nota de Rodapé 24), muito embora assumir o erro presente neste último par não acarrete absolutamente as mesmas, *ou quaisquer*, implicações. É de acreditar, pois, que ambos os erros tenham natureza *muito* diferente: a verdadeira razão de "~~botano~~" ("botando"), postulamos, é o desconhecimento de uma certa relação entre ortografia e morfologia, permitindo que aflore a transcrição fonética daquele segmento geralmente protegido.

Em cada hipótese de correção, a reconstituição vale como prova de que um dado perfil realmente desviaria o usuário da intenção original para o não-vocábulo. Exatamente por isso, um argumento de utilidade só convencerá se for **verossímil**, isto é, se sua reconstituição não contradisser seu perfil de usuário, ou melhor, dele decorrer de forma natural. Verossimilhança é o primeiro dos três **atributos** de qualquer argumento de utilidade que se preze. Nos exemplos, todas as reconstituições são verossímeis[26] porque é o próprio leitor que as deriva mentalmente, a partir dos respectivos perfis de usuário.

Verossimilhança, entretanto, é uma condição necessária, mas não suficiente. É necessário que o usuário também tenha dificuldade na recuperação do erro, *o que não é implicado necessariamente pelo simples fato de este ser verossímil*. Para verificar isso, basta tentar defender a utilidade, para um sujeito ideal, da sugestão de "caçarola" para "~~cassarola~~": é fácil perceber que a reconstituição óbvia aí implícita é perfeitamente verossímil, mas contém um erro de fácil recuperação. Logo, "caçarola" não prima pela utilidade, devendo acabar por ser apresentada, no entanto, devido à falta de candidatos mais úteis. Vale a pena observar que o par *(falação, ~~falasão~~)*, contrariamente às aparências, não guarda a mesma relação que o par em questão: analogamente a *(botando, ~~botano~~)*, subentende uma reconstituição completamente inverossímil, absurda para o sujeito ideal. Em contraste, "falastrão" seria uma sugestão de utilidade memorável nesse caso, exatamente porque permite uma reconstituição de verossimilhança impecável e contendo um erro (muito) difícil.

Este é o detalhe que falta: reconstituições contêm erros, ou melhor, pontos de erro[27]; enquanto perfis de usuário contêm informações sobre o nível de dificuldade de cada tipo de erro, as quais variam de perfil para perfil. Equacionando tudo, podemos identificar o segundo atributo do argumento de utilidade — **desafio**, que representa a dificuldade do usuário na recuperação do(s)

---

[26] Por extensão e em nome da clareza, como *verossimilhança* é não só um atributo do argumento de utilidade, mas também uma relação que se estabelece entre os seus elementos, aplicaremos o termo, adicionalmente e sem receio, tanto a reconstituições quanto a perfis de usuário, *entre outros*. Este é um termo que usaremos com certa liberdade, mas sempre com propriedade: falar que uma entidade interna a um argumento de utilidade é verossímil/inverossímil é dizer que ela não torna/torna o argumento inverossímil.

[27] Essa idéia, bem como outros conceitos, ficará mais clara na seção sobre o paradigma reverso de correção de erros (página 46), que é complementar a esta à perfeição.

erro(s) implicados pela sugestão cuja utilidade está sendo defendida. Como verossimilhança, esse atributo também surge da interação entre os dois componentes do argumento de utilidade.

Esse não parece ser o caso do terceiro e último atributo que pudemos identificar — **otimismo**, que será explanado na seção sobre reversão genérica de erros, à página 46. Já adiantamos, de qualquer forma, que o otimismo tem a ver com *a própria possibilidade de correção*, tentando evitar que a reconstituição seja demasiado mirabolante e, portanto, aplicando-se mais especificamente a esse componente. Analogamente ao que aconteceu com a verossimilhança nos exemplos do início desta seção, arriscamos que o otimismo é uma dos pressupostos da inteligência humana e que, por isso, o leitor tenha então derivado reconstituições não só verossímeis, mas também otimistas.

Postulamos daí que uma sugestão de correção seja **útil por hipótese** se e somente se há para ela um argumento de utilidade. Além disso, por extensão e extrapolando, diremos que toda sugestão útil é realista, otimista, desafiante, verossímil e **reconstituível**.

Realismo ou O não-atributo
ou ainda O papel do embasamento lingüístico

Nossos três primeiros atributos do argumento de utilidade têm uma peculiaridade em comum, a saber: *nenhum compromisso com a realidade*. Reparando bem, podemos notar que *são propriedades meramente formais dentro de uma hipótese cosmológica*[28], o que é inevitável, mesmo se estivéssemos tratando *apenas*[29] dos argumentos de utilidade a serem produzidos por humanos. Faz-se necessário um atributo adicional — **realismo**, que, como já sugerimos, é o único que tenta saltar do "plano" formal para o "espaço" real, mas só tenta, porque assumimos que este é incognoscível. O nível de realismo se refere a quanto nosso entendimento (modelagem, mental ou computacional) das diversas entidades envolvidas e do recorte do mundo que nos interessa se aproxima da realidade última desses mesmos elementos, nem que seja de um ponto

[28] Se não foi essa a impressão que o leitor teve, talvez seja recomendável repassar a leitura com esse dado em mente, após terminar de ler esta seção; e nos desculpamos por uma expressão ou outra não tão bem-escolhida.

[29] Pode parecer presunçoso, mas estamos tratando — ou ao menos tentando tratar — de *qualquer* argumento de utilidade.

de vista apenas behaviorista. Ou seja, entenda-se "realista" como "próximo ou equivalente à realidade última". Só para tornar o conceito mais palpável, vamos mencionar que ele tem a ver, por exemplo, com (i) se o perfil de usuário assumido corresponde ao perfil *real* do usuário; (ii) se os aspectos considerados na construção dos perfis de usuário são *realmente* relevantes, suficientes e estão pesados da forma "correta"; (iii) se as reconstituições se aproximam dos processos que o usuário *realmente* usa para produzir vocábulos; etc.

Em todos os nossos exemplos e na vida, o realismo é, a rigor, uma questão em aberto e, em última análise, de fé. Será que aquelas sugestões são *realmente* úteis? Em verdade, só *parece* que sim porque seus respectivos argumentos de utilidade *nos parecem* realistas. Porque as reconstituições lá implícitas apelam à nossa intuição do que acontece quando produzimos palavras escritas, de como a língua (escrita) funciona. Porque acreditamos em perfis de usuário e que aqueles em específico existam. Porque atribuímos autoridade a Cagliari (92).

Por essas e outras razões, talvez realismo seja mais um atributo cosmológico, sendo apenas "herdado" pelo argumento de utilidade. Tal é a postura que adotaremos neste trabalho. Assumimos que, numa cosmologia realista, não haja lugar para argumentos, reconstituições, perfis de usuário, etc. não-realistas. Em outras palavras, tudo será realista por hipótese e construção, na medida do possível. Isso pode soar como se fôssemos ganhar realismo de brinde. Ledo engano. Eis a única função de nosso entusiasmo em levantamento e aplicação de conhecimentos lingüísticos e em modelagem psicolingüística: conferir realismo à nossa cosmologia, compor uma cosmologia realista.

Medida de utilidade

Neste ponto, dispomos de uma definição operacional — e, por incrível que pareça, formal — de utilidade, o que representa um salto qualitativo. Mas não alto o suficiente: o ideal seria dispor de uma definição de *medida de utilidade*, nem que fosse meramente para poder ordenar listas de candidatos a correção. Felizmente, os elementos de que já dispomos se prestam naturalmente a essa extensão. A idéia é simples: transformar os atributos do argumento de utilidade em grandezas contínuas e definir algo como a *força de um argumento de utilidade*.

Em primeiro lugar, assumimos (i) três **funções-atributo** — *verossimilhança*, *desafio* e

otimismo:$\{$*(Perfil, Reconstituição)*$\}$[30] $\rightarrow$ *[0, 1]* — que computam, respectivamente, quão verossímil, desafiante e otimista é um argumento, ou seja, geram índices para cada um de seus atributos; bem como (ii) uma **função-reconstituição**

$$reconstituição: Léxico \times Alfabeto^{*}[31] \rightarrow P(\{Reconstituição\})[32]$$

tal que *reconstituição(Sugestão|* ~~*NãoVoc*~~*)* retorna o conjunto, potencialmente numeroso, de todas as possíveis reconstituições demonstrando como alguém — qualquer um — poderia chegar ao não-vocábulo ~~*NãoVoc*~~ se pretendesse originalmente *Sugestão*.

Basta agora colocar essa (poderosa) biblioteca de funções para trabalhar. Definimos um **critério de utilidade** como qualquer função $f:R^3 \rightarrow [0, 1]$ que serve para gerar um índice único de utilidade a partir de suas três entradas, respectivamente os índices de verossimilhança, desafio e otimismo *de um mesmo argumento*. Segue trivialmente daí a **força de um argumento de utilidade** segundo um critério *f*, a qual fica definida assim:

$$força_f: \{(Perfil, Reconstituição)\} \rightarrow [0, 1]$$
$$força_f(x) = f(verossimilhança(x), desafio(x), otimismo(x))$$

Finalmente, definimos a **medida da utilidade** de uma sugestão de correção segundo um critério *f*, dados um não-vocábulo e um perfil de usuário, como a seguinte função:

$$utilidade_f: Léxico \times Alfabeto^{*} \times \{Perfil\} \rightarrow [0, 1]$$
$$utilidade_f(sugestão|\ \sout{não Voc}, perfil) = máx(força_f(\{perfil\} \times reconstituição(sugestão|\ \sout{não Voc})))$$

Em resumo, dados um não-vocábulo e um perfil de usuário, a utilidade de uma sugestão é igual à força do melhor (= "mais forte") argumento de utilidade que a corrobore.

---

[30] Denota-se por $\{X\}$ o conjunto-universo das entidades que têm um protótipo em *X*.

[31] Denota-se por Σ^* o fechamento de um alfabeto Σ, ou seja, o conjunto de todas as cadeias de qualquer comprimento que podem ser formadas com os símbolos de Σ.

[32] Denota-se por ***P(X)*** o **conjunto das partes** — ou *dos subconjuntos* — de um conjunto *X*.

Medindo e maximizando utilidade e propaganda do paradigma reverso

Sendo em princípio possível medir utilidade, a questão se impõe de como gerar sugestões de correção maximamente úteis. Uma primeira resposta, maximamente ingênua, seria classificar todo o léxico quanto à utilidade, a cada novo não-vocábulo. Flagrantemente inviável, esse protótipo simplista já encerra uma questão fundamental — a de *como* medir a utilidade de uma única sugestão — a que ainda não demos e não vamos dar resposta definitiva. É importante perceber que as definições apresentadas até agora estão mais para "o que" do que para "como" e que pensar em medir utilidade pela definição é, no mínimo, ingenuidade, bastando pensar no cálculo da função *reconstituição* para desistir da idéia.

Ainda quanto ao nosso protótipo ingênuo, vale observar que ele se enquadraria no paradigma relativo de correção de erros, que, como já vimos no Capítulo II, já rendeu soluções engenhosas e admiráveis e, em princípio, poderia muito bem se prestar a maximizar utilidade. No entanto, essa não é a opção deste trabalho, em que adotamos o paradigma reverso para o mesmo fim. Nosso objetivo último, por conseguinte, pode ser expresso como procurar algoritmos que realizem mutações em cadeias de entrada de forma maximamente útil e, por isso, lingüisticamente realista/fundamentada[33], com o intuito de *anular* as mutações "perpetradas" pelo autor das cadeias em questão.

Há uma série de argumentos com que justificar essa opção, eis alguns deles:

- **naturalidade**: como veremos na próxima seção e nas seguintes, o paradigma reverso é perfeitamente compatível com nossos objetivos e as idéias que vimos desenvolvendo neste capítulo. Até a ponto de algumas considerações *parecerem* redundantes, muito embora, advertimos, não haja nenhuma implicação necessária entre elas, na verdade;

[33] Recorrendo a uma breve analogia com a física, *plausibilidade lingüística* aqui se refere a que as mutações sejam projetadas de modo a serem da mesma *natureza*, direção e módulo, mas de sentido inverso, das mutações que os falantes do Português realizam ao gerarem não-vocábulos. Somadas, as mutações do autor e do corretor se anulariam, gerando um vocábulo.

- **potencialidade**: um sistema corretor baseado no paradigma reverso, da forma como entendemos e idealizamos, tem potencial para se tornar uma interessante ferramenta educacional, já que gera explicações (reconstituições) para os erros do usuário e (idealmente, se dotado de realismo suficiente) identifica suas causas profundas;

- **parcialidade**: admitimos uma certa parcialidade pré-definida ao paradigma reverso, plenamente justificada pela oportunidade de alguns recursos poderosos disponíveis no *NILC*. Dentre eles, destaca-se a **biblioteca KLS-GT**, componente do *ReGra*[34], a qual permite acesso eficiente ao amplo léxico desse sistema (aproximadamente 1,5 milhão de vocábulos). Sua funcionalidade inclui, só para citar os itens de maior interesse, (i) a verificação de se uma cadeia é lexicalizada ou não e, em caso positivo, (ii) a determinação de seus traços gramaticais (possíveis classes gramaticais, gênero, número, tempo, etc.), (iii) determinação de qualquer flexão ou, inversamente, da forma canônica de qualquer palavra lexicalizada e (iv) divisão silábica. Além disso, seu código é bastante compacto, bem como o arquivo de dados do léxico (pouco mais de 1,3Mb), e extremamente eficiente, todas as operações citadas sendo executadas *em tempo constante*[35]. A biblioteca KLS-GT praticamente *é* o conselheiro ortográfico do Word 2000 e foi desenvolvida por Tomasz Kowaltowski, Cláudio Lucchesi e Jorge Stolfi, pesquisadores do Instituto de Computação (CCUEC) da Unicamp.

  Essa biblioteca constitui, em conjunto com o referido léxico, a pedra fundamental de nossa futura implementação. Suas características a tornam ideal para suportar alguma técnica reversa de correção de erros ortográficos. Em termos de eficiência, a ordem de complexidade do algoritmo de reversão resultante, qualquer que seja, reduz-se à ordem de complexidade do algoritmo de geração de alternativas de correção propriamente dito, já que o algoritmo de validação das alternativas geradas (acesso ao léxico) é *O(1)*.

Segundo o paradigma reverso, como sugere o próprio nome, *construiremos* sugestões de trás para frente, o que, grosso modo, ocorrerá da seguinte maneira: (i) a partir do não-vocábulo e de um corpo de conhecimentos, informações e *heurísticas* que podemos chamar de cosmologia,

---

[34] http://www.nilc.icmc.sc.usp.br, link *Projetos/Regra*.

[35] Mais especificamente, numa análise de *pior caso*, *O(mín{n, n_máx})*, onde *n* e *n_máx* são respectivamente os comprimentos da cadeia de entrada e da maior cadeia lexicalizada.

derivaremos argumentos de utilidade maximamente fortes *por construção*; (ii) a partir dos argumentos assim obtidos, derivaremos sugestões de correção, maximamente úteis, portanto; e (iii) as validaremos, testando sua pertinência ao léxico.

Terminada a propaganda, vamos nos deter na discussão mais cuidada do nosso paradigma de correção e suas implicações.

Reversão: otimismo, profundidade, intenção, gatos & microondas

Um item malformado passível de correção é muito mais um acerto do que um erro**:** só há esperança se o usuário tiver acertado muito mais do que errado na produção de um tal item. É a partir desse **otimismo** que toda correção se torna possível. Deve ficar clara, daí e deste ponto em diante, a distinção entre **"item malformado"** e **"erro"**: o primeiro é resultado de um processo em que o segundo ocorre como fator de perturbação, espera-se, bem localizada. A visão de "erro" como "uma *operação* equivocada num *processo*", em oposição às usuais "um *defeito* num *produto*" ou "um *produto defeituoso*", é bastante lúcida e oportuna, permitindo uma modelagem adequada da reversão de erros.

Uma primeira conseqüência dessa mudança de perspectiva é a de que não existe algo como *erros de superfície*: antes, todo erro está a uma certa **profundidade**, ou seja, *jamais* está explícito no item malformado a ponto de sua identificação prescindir de algum tipo de inferência, suposição ou análise acerca do processo de produção. Naturalmente, o desastre não é propriamente a idéia de que um **gato** seja instância da classe *Microwaveable*[36], mas o que se faz com o gato *a partir dessa idéia*. Ou seja, nessa situação, a presença de um erro seria sugerida a um sistema corretor hipotético pela insatisfação do usuário ao ver o que aconteceu com seu gato após um bombardeio de **microondas**. Qual foi o verdadeiro erro — o ponto no processo que desencadeou o gato "malformado" — e como ele poderia ser revertido?

A reversão desse erro, é claro, não consistiria na ressurreição do gato. Nem tampouco em informar o usuário de que gatos são assassinados daquela maneira, algo que ele teria acabado de

[36] *Microwaveable* (inglês) = "que entra num forno de microondas e sai em melhor forma".

aprender. Uma alternativa desejável — ou *útil* — seria a emissão, por exemplo, da seguinte sugestão: "Da próxima vez, tente usar um secador de cabelos... ou uma toalha!" No entanto, como chegar a essa conclusão sem que o usuário seja capaz de expressar sua intenção inicial? Essa condição, aqui aparentemente absurda, é bastante realista nas situações análogas vividas por um usuário de computador. Na maioria dos casos, o usuário ser capaz de comunicar ao computador sua intenção original implica ele também ser capaz de se corrigir sozinho[37].

O processamento para chegar à saída desejada não é absolutamente trivial. Uma linha de raciocínio possível (e supersimplificada, mas útil como ilustração) seria a apresentada na Figura 1.

CONCLUSÕES/QUESTÕES	**FATOS RELEVANTES**
	O usuário amava o gato.
LOGO, o usuário não pretendia matá-lo.	
O que pretendia então?	
Esquentar o gato **E** mantê-lo vivo.	
Para quê?	
	Calor seca água **E** o gato estava molhado **E** humanos não gostam de "molhação" **E** tempo é dinheiro.
Para secar o gato rápido.	
O que pode secar o gato rápido **E** mantê-lo vivo?	PONTO DE ERRO
Um ~~microondas~~ secador de cabelos!	

Figura 1: Reversão[38] do processo de secagem/sacrifício do gato.

Para sugerir o uso do secador de cabelos, o sistema sem dúvida teria que ser otimista e inferir a intenção inicial do usuário a partir dos (supostos) acertos do processo. Além disso, provavelmente teria levantado a hipótese de que o usuário se fez a pergunta "O que pode secar o

[37] Complementarmente, a compreensão de uma tal comunicação por parte do computador provavelmente implicaria tanta inteligência (artificial) que tornaria toda a presente discussão e projeto obsoletos. No máximo, o sistema pode inferir e apresentar alternativas de intenção inicial (poucas, boas e claras) para que o usuário selecione a mais adequada.

gato...” e errou no momento — ou *no ponto* — de respondê-la. E esse é, de fato, *o único erro* ou **ponto de erro** num processo, de resto, correto, o que não o impede de culminar em desastre.

Em resumo e princípio, o procedimento de reversão de erros pode ser entendido como envolvendo duas operações:

(i) a reversão[38] do processo de produção de um item malformado, provavelmente resultando em diversas (re)versões hipotéticas, visto que boa parte do processo, senão todo, costuma se passar apenas na mente do usuário. Boas reversões serão sempre otimistas — contendo alguns poucos pontos de erro, todos plausíveis — consistentes com as circunstâncias conhecidas (ou seja, verossímeis);

(ii) a reconstituição do processo, agora revisado quanto ao resultado das operações em pontos de erro.

O que se vê na Figura 1 é apenas uma possível hipótese de reversão do processo de sacrifício do gato, considerada excelente por ser baseada na suposição de um único erro, maximamente plausível, dado o absurdo da situação. Aplicaremos, a seguir, os conceitos aqui introduzidos ao domínio da reversão de erros ortográficos.

Revertendo erros ortográficos

Especializando as entidades do modelo abstrato de reversão introduzido, temos:

Reversão de erros (abstrata)	**Reversão de erros ortográficos**
item malformado	não-vocábulo
processo enfocado	produção de vocábulos (em português)

Ou seja, um *não-vocábulo* é o resultado de um erro na *produção de vocábulos*. Além disso, um tal erro é dito *ortográfico*. Essas duas *definições* introduzem uma terminologia básica, própria deste projeto, que é inconsistente com uma interpretação mais rigorosa de cada um de seus termos. Em primeiro lugar, “ortográfico” é usado aqui em sentido mais amplo: os erros que

[38] O termo “**reversão**” deve aqui ser entendido como “reconstituição de trás para frente”.

levam a "~~reaveu~~" e "~~mortandela~~", por exemplo, são ditos ortográficos, muito embora não sejam propriamente "de ortografia" (!). Estritamente falando, o termo "**ortografia**" se refere pura e simplesmente à forma correta de *grafar* um vocábulo *dado que ele pertence ao léxico* ou, no caso de neologismos, *assumindo que ele pertença ao léxico*. Naturalmente, essa condição não se verifica para os dois erros em questão, que ocorrem em outro momento que o da grafia propriamente dita dos respectivos não-vocábulos. E, exatamente por "grafia" costumeiramente se aproximar de "converter som em escrita", adotamos o pouco usual "**produção de vocábulos**" para se referir ao processo de chegar até a grafia de um vocábulo *partindo de suas especificações/restrições semânticas e/ou gramaticais*, provenientes da intenção do autor, contexto sintático, etc., ou melhor, *partindo de um conjunto de restrições que garanta selecionar todos e somente os vocábulos (corretos) do léxico que possam ocupar a posição do vocábulo em questão mantendo o significado pretendido*. O grande mérito desta última definição está na simples identificação de uma *entrada* — ou *ponto de partida* — para o processo em foco a qual é tão primitiva quanto um modelo de correção cego ao contexto, como o nosso, pode "deglutir". Outro mérito está em ser subespecificada, aberta, passível de ser refinada com as fases que se mostrarem pertinentes, tais como as de "formação de palavras", "grafia" e até "digitação".

Em contraste, os corretores ortográficos mais simples pressupõem que os erros ocorram apenas na digitação, o nível mais superficial no processo de produção de vocábulos. Na verdade, esses sistemas encaram o processo como partindo da grafia correta do vocábulo pretendido (!), no que reside a essência de sua pouca utilidade. Outros, mais sofisticados e um pouco mais úteis, assumem uma imagem sonora correta como ponto de partida, o que pressupõe uma boa formação morfológica e nenhuma deturpação fonológica.

Como queremos identificar possíveis erros ortográficos em não-vocábulos e, portanto, teremos de proceder a "algum tipo de inferência, suposição ou análise" acerca da produção de vocábulos, torna-se imperativo modelar esse processo. Primeiro, no entanto, contextualizaremos um tal modelo como componente numa arquitetura computacional genérica de reversão de erros. E chamaremos a esse componente "gramática de reconstituição".

Um conselheiro (qualquer) centrado em utilidade segundo o paradigma reverso

Equacionando todos os elementos já apresentados neste capítulo, delinearemos uma arquitetura computacional para conselheiros construídos segundo o paradigma reverso de correção de erros *e* centrados em utilidade. Note-se aí a ausência do adjetivo "ortográfico": estamos extrapolando para sistemas **corretores interativos** (= "conselheiros") **quaisquer**, uma classe genérica apenas quanto à natureza dos itens de que se esperam malformações. Para tanto, pressupomos do leitor um passo trivial de abstração sobre os conceitos de "argumento" e "medida de utilidade" para que se apliquem a situações de aconselhamento quaisquer, ou seja, a tuplas *(Sugestão|* ~~*ItemMalformado*~~*, Perfil)* quaisquer. Basta, para isso, considerar as respectivas seções como um estudo-de-caso para o aconselhamento ortográfico.

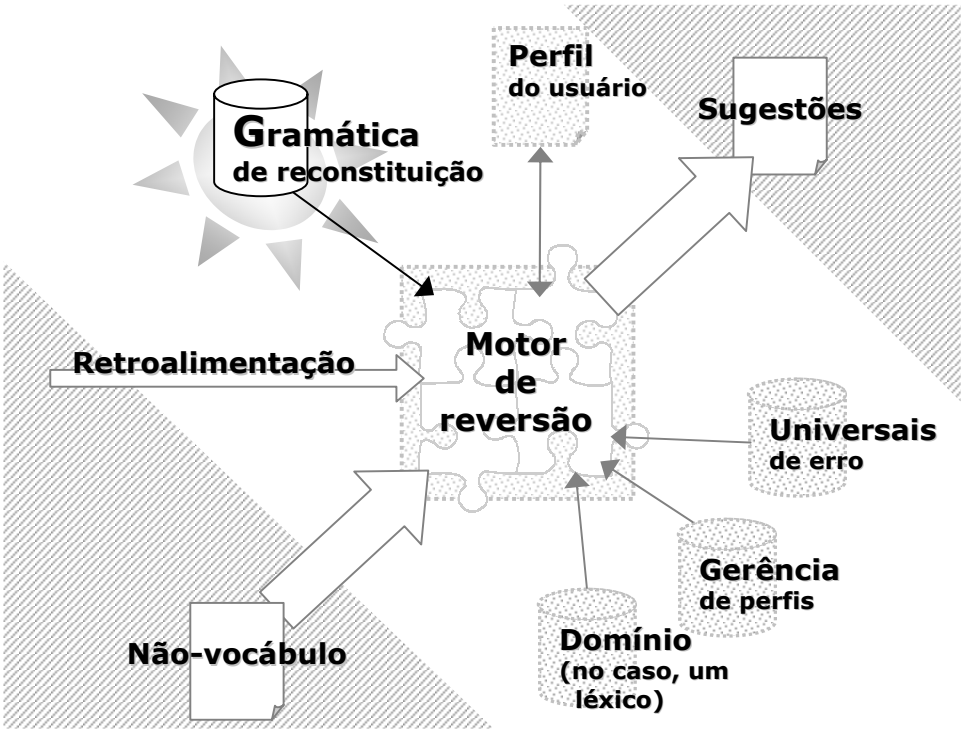


Figura 2: Arquitetura genérica de reversão centrada em utilidade.

Apresentamos, na Figura 2, um esquema de nossa arquitetura, em que figuram seus componentes de mais alto nível. Cabe adiantar que o mais interessante em nossa proposta é a *dissociação entre os diversos tipos de conhecimento*, em especial ao isolar o componente *gramática de*

reconstituição[39]. Adicionalmente, chamamos a atenção para o fato de a arquitetura suportar *adaptabilidade ao usuário*, uma capacidade ideal para sistemas conselheiros centrados em utilidade, visto ser este um parâmetro pessoal.

A **gramática de reconstituição** compreende todo o conhecimento necessário para gerar reconstituições, devidamente anotadas quanto aos pontos de erro; mas não entra no mérito de se toda a sua "prole" é verossímil, desafiante e otimista. No exemplo do gato (pág. 46), uma boa gramática de reconstituição certamente teria de gerar algo como apresentado na Figura 1 (pág. 47), ao mesmo tempo em que poderia perfeitamente levantar a hipótese de que o usuário pretendia confinar o gato até que este se secasse e, na falta de uma gaiola, usou o microondas e, como o bichano tivesse medo do escuro, acionou o microondas tão-somente para acender a luzinha interna... Francamente, a mera consideração desse tipo de disparate tem que ser evitada, o que é responsabilidade do **motor de reversão**.

É exatamente por isso que este é o único item de processamento em nosso esquema, o que lhe permite usar, de (alguma) forma integrada, a totalidade do conhecimento disponível e investir apenas nas reconstituições promissoras. O motor de reversão usará a gramática com "sabedoria", provavelmente numa espécie de *parsing* heurístico *bottom-up*. Por exemplo, uma possível implementação poderia se basear no algoritmo de busca A^* (Russel & Norvig, 95), (i) mantendo a utilidade parcial de diversas reconstituições geradas em paralelo, (ii) estimando heuristicamente o *potencial utilitário futuro* prometido por cada possível próximo passo de expansão das reconstituições correntes, e (iii), com base nesses dados, inibi-las temporária (expandindo primeiro as mais promissoras) ou definitivamente (devido a um *limiar de utilidade*).

Uma responsabilidade adicional do motor de inferência consiste na manutenção de um **perfil de usuário**. Para tanto, aplica conhecimentos relativos à **gerência de perfis** a algum tipo de **retroalimentação** informando quais sugestões foram aceitas pelo usuário em cada caso e até *em que casos o usuário pediu ajuda ou conseguiu se corrigir sozinho*, se isso for suportado pela interface do sistema.

---

[39] A identificação das demais bases de conhecimento nos parece justificável, porém admitimos ser perfeitamente passível de revisão.

A "totalidade do conhecimento disponível", por sua vez , está distribuída entre quatro grandes bases de conhecimento segundo a pertinência de cada item a um dos seguintes domínios:

- **reconstituição** e **gerência de perfis** (já comentados);
- **universais de erro**, isto é, predicados sobre erros válidos para qualquer usuário, incluindo relações de implicação entre erros, tais como a que traçamos entre "~~botano~~" ("botando") e "~~vendero~~" ("venderam");
- **o domínio de aplicação propriamente dito**, considerado fora do contexto de correção de erros. No caso do aconselhamento ortográfico, a base em questão deveria conter, pelo menos, um léxico, para que a validade das sugestões geradas pudesse ser verificada. Outros itens de interesse incluem uma "ontologia gramatical", informações lingüísticas (ao menos morfossintáticas) acerca das entradas do léxico, regras gramaticais, etc.

Em direção a uma gramática de reconstituição: metodologia

Nos capítulos seguintes, vamos fazer apontamentos que acreditamos úteis ao projeto de uma gramática de reconstituição para um conselheiro ortográfico construído segundo a proposta deste capítulo. Naturalmente, nossos apontamentos consistirão em levantamento de conhecimentos e hipóteses referentes ao processo de produção de vocábulos e aos seus possíveis pontos de erro. Antes de tudo, entretanto, cabe fazer algumas considerações metodológicas finais sobre como procedemos para chegar aos resultados apresentados.

Uma rotina que estabelecemos foi a da análise de não-vocábulos segundo conhecimentos diversos, principalmente lingüísticos, e a incorporação das conclusões resultantes na versão corrente de nossas hipóteses. Pode-se argumentar que essa seria uma tarefa inglória, visto o conjunto de todos os possíveis não-vocábulos ser, em princípio, infinito. Não nos abalemos diante de tamanho pessimismo, pois que uma nova versão do conceito de otimismo vem ao nosso socorro. Diz assim:

> Enquanto talvez não exista qualquer ordem no conjunto de todos os possíveis itens malformados, todos os itens malformados passíveis de correção (por qualquer agente, humano ou computacional) obedecem a leis de (mal)formação relativamente poucas: são,

na verdade, bem-formados num sistema (levemente) relaxado.

Daí decorre a esperança de que exista um conjunto limitado *P* de itens malformados, ditos *protótipos*, a partir do qual o conjunto *L* de todas as leis de malformação possa ser inferido. O que tentamos foi encontrar, incrementalmente, aproximações para *P* e *L* com a precisão necessária e embutir os elementos de *L* em nosso modelo na medida em que foram sendo inferidos/descobertos.

Nossos protótipos provêm de listas de erros ortográficos comuns, como as encontradas em (Faraco & Moura, 94), (Sacconi, 92) e (Cagliari, 92), variantes dialetais e infantis clássicas, como "~~ponhei~~" ("pus"), "~~trabaio~~" ("trabalho"), "~~di~~" ("dei") e "~~cabeu~~" ("coube"), bem como de um trabalho contínuo de coleta de novos casos em situações do dia-a-dia e nos meios de comunicação de massa. À primeira vista, talvez se duvide do cabimento de considerar formas como "~~ponhei~~" e "~~di~~". Poder-se-ia argumentar (como já o foi, várias vezes) que estas jamais seriam digitadas num computador. Contudo, fica patente a invalidade desse argumento se considerarmos a disseminação crescente e generalizada do computador e que o conhecimento aqui contido pode até fazer parte de sistemas envolvendo processamento de fala, por exemplo. Acima de tudo, basta mencionar que, pela análise de exatamente essas formas, chegamos a uma lei subjacente de malformação que os "bem-letrados" também costumam aplicar, como veremos[40]. Dessa forma, parece imprudente não só (i) descartar qualquer protótipo antes de se constatar que as leis aplicadas em sua (mal)formação são conhecidas, como também (ii) descartar uma nova lei só porque não se conhece um protótipo natural do usuário-alvo que envolva a aplicação da lei em questão.

Iniciou-se um trabalho de análise de um extenso corpus de não-vocábulos que foi, no entanto, abandonado, devido à constatação de que boa parte se tratava de erros simples no nível da grafia, os demais não contribuindo para o crescimento do conjunto de protótipos. Apesar de o talvez único fruto dessa análise ter sido o resultado de que há uma tendência clara à ignorância ou desprezo das regras de acentuação gráfica, o corpus levantado provavelmente se provará útil quando da avaliação e sintonia fina de uma futura implementação.

---

[40] Existe uma analogia direta entre erros como "~~di~~", "~~constrangiu~~" ("constrangeu") e "~~reaveu~~" ("reouve").

Capítulo IV Nossos sistemas de escrita

Chamamos aqui **sistema ortográfico/de escrita** a qualquer sistema de convenções que permita atribuir uma grafia a todo vocábulo corrente ou potencial de uma língua. Este capítulo trata especificamente do sistema ortográfico adotado como norma culta no português do Brasil, dito *nosso*, e dá conta de uma parte significativa fase ortográfica da produção de vocábulos. No final, procederemos à identificação de pontos de erro no uso desse sistema, o que resultará, colateral e virtualmente, numa aproximação (imprecisa) do conjunto dos vários sistemas ortográficos prováveis e correntes nas mentes dos falantes/"escreventes", aqui referidos simplesmente como *usuários*[41]. Uma meta ambiciosa, sem dúvida, mas possibilitada por tratarmos aqui, em oportunidade única neste projeto, de uma entidade artificial, quase extralingüística.

Não é fácil, não!

É comum ter-se a impressão de que o nosso sistema de escrita seja simples, "fonético", em que "se escreve como se fala". Como veremos a seguir, isso não corresponde absolutamente à verdade — identificados muitos dos tipos de regra vigentes, bem como explicitadas, de forma razoavelmente realista, algumas regras principais, ficam patentes pontos que desmentem essas crenças, a saber:

i. poucas que sejam, as regras se expressam com primitivas bastante abstratas (sílabas, ápices, onsets, codas, morfemas, tonicidade, etc.). Quanto menos familiares essas abstrações, tanto maior será o caos aparente do sistema;

ii. certas regras envolvem linhas de raciocínio não-triviais, tais como recursividade e provas por contradição. Quanto menos familiares as modalidades envolvidas, tanto mais as regras

[41] A rigor, no sentido de "usuários do sistema de escrita", mas também "usuários em potencial de nosso *spell-checker*".

em questão serão substituídas por outras mais particulares e, portanto, mais numerosas, desconexas, "esquecíveis" e cheias de exceções. Por exemplo, as regras mais elegantes para a acentuação gráfica envolvem uma prova por contradição ("Se essa palavra não fosse acentuada, como seria lida?"). Dessa forma, a prova da necessidade de acento gráfico em "Piauí", por exemplo, requer uma tentativa de leitura de "~~Piaui~~" ([pi.a.ˈuj]), que envolve, por sua vez, um encadeamento de dependências que pode ser entendido como um tipo de recursão;

iii. as regras testam condições em diferentes níveis, lingüísticos ou não: gráfico, fonético, fonológico/fonêmico[42] e morfológico. E muito menos no nível fonético quanto seria esperado, o que é facilmente comprovado ao se considerar que simplesmente saber uma pronúncia válida de um vocábulo nem sempre implica ser capaz de grafá-lo corretamente, mesmo que se assuma bom conhecimento do sistema de escrita. Outra evidência cabal: foneticamente, em muitos dialetos, há uma inconsistência básica entre as grafias "parte" e "parto" quanto ao valor do grafema "t"[43], a qual deixa de existir numa perspectiva fonológica[44];

iv. as regras artificiais do sistema ortográfico interagem e se completam com as regras naturais da língua, numa dialética razoavelmente complexa. Por exemplo, a decodificação da tonicidade[45] (um dado virtualmente explícito) requer uma divisão silábica aproximada, que, por sua vez, está codificada de forma incompleta. É apenas considerando as reais

---

[42] Em geral, usaremos o adjetivo "fonológico" em oposição a "fonético", à maneira da escola lingüística de Praga. Para os não-iniciados, essa distinção será esclarecida em momento oportuno. Nessa acepção, temos um sinônimo de "fonológico" em "fonêmico", de acordo com a terminologia norte-americana, entretanto.

[43] Valendo ora [t͡ʃ], ora [t], respectivamente.

[44] Nas descrições da fonologia do português, os fones [t͡ʃ] e [t] não costumam se encontrar em oposição fonológica. Ambos são considerados simples manifestações superficiais alternativas (**alofones**) de um mesmo fonema /t/ que surgem em contextos mutuamente exclusivos. Essa hipótese só é desafiada pela existência de "tchau", que forma um par mínimo com "tau" (Houaiss, 01). Ambas as palavras são, entretanto, empréstimos de outras línguas (italiana e chinesa, respectivamente).

[45] Entenda-se aqui "decodificação da tonicidade" como a identificação da vogal tônica de um vocábulo a partir de dados presentes em sua grafia.

possibilidades fonéticas da língua que se torna possível preencher certas lacunas e proceder à divisão requerida. Como ilustração, confronte os vocábulos dos pares *(caindo, paina)* e *(cair, Cairo)* e responda:

a) por que as seqüências "c", "n", "nd", "ndo", "p" e "r" não figuram entre as possíveis sílabas (gráficas)?

b) por que há hiato em "c**a**|**i**n|do" e "c**a**|**i**r" e ditongo em "p**ai**|na" e "C**ai**|ro"? Como veremos, isso é regido por uma das regras mais complexas de todo o sistema, envolvendo um verdadeiro diálogo entre regras naturais e artificiais;

v. a minimização da representação de redundâncias leva até a exceções adicionais às regras mais gerais do sistema. Um exemplo de exceção resultante dessa preocupação é "rainha", que não aceita acento gráfico, "contrariando" uma regra bastante geral de codificação de hiatos, ou "quebra de ditongos". Isso ocorre pelo simples fato de, *fonologicamente* em português, não poder haver um "i" assilábico (formando ditongo decrescente com a vogal anterior, portanto) seguido do fonema comumente denotado por "**nh**". A grafia de um tal "i" seria bloqueada ou por impossível (dada a impossibilidade da própria entidade denotada) ou por redundante[46]. Portanto, sempre será silábico (constituindo, a rigor, hiato com uma eventual vogal anterior) um "i" que tiver sido legitimamente grafado antes de um "nh".

Caos aparente: leitura precária vs. expressividade

Talvez a raiz da complexidade do nosso sistema de escrita resida no que parece ser um requisito básico do seu projeto: **alta expressividade**[47]. O sistema permite e prescreve a codificação *em*

---

[46] Interpretamos o fone [ɲ] como mera conseqüência fonética de um "i" nasalizado, quer fonológico/silábico, quer não. Evidência disso é o próprio surgimento de um "i" silábico fonético em algumas variantes de pronúncia de "nhoque" e "nhambu", por exemplo, pronunciados "~~inhoque~~" e "inhambu". Esta última grafia chega inclusive a ser atestada (Aurélio, 96).

A "contração" de um "i" silábico com um [ɲ] subseqüente também é comum, veja, por exemplo, o surgimento de "nhô" a partir de "senhor" (Houaiss, 01), passando por "sinhô". Nossa interpretação é ainda corroborada pela origem de "iaiá" em "sinhá" (Houaiss, 01) e de "caminhonete/caminhão" em "camionete/camião" (Aurélio, 96).

[47] O termo "expressividade" entendido como taxa de conteúdo por "unidades" de forma.

paralelo de vários tipos de informação na grafia de um vocábulo — tais como divisão silábica, tonicidade, alguns dados fonológicos adicionais e outros tantos morfológicos — muitas vezes ignorados pelos sistemas ortográficos de outras línguas[48].

Como já adiantado, as regras de (de)codificação não são lá das mais fáceis, o que costuma levar a uma apreensão bastante parcial do sistema. Uma conseqüência natural e freqüente é o fenômeno que chamamos **leitura precária**: apenas uma fração da informação codificada costuma ser recuperada, ou seja, realmente *lida*, e não simplesmente *adivinhada* ou inferida a partir do contexto. Numa leitura precária, muitas das marcas prescritas pelo sistema não são sequer percebidas, a não ser talvez como meros caprichos sem justificativa, difíceis, portanto, de memorizar e inacessíveis à manipulação consciente (quando da inferência de alternativas válidas de grafia/pronúncia para um vocábulo jamais grafado/lido por um usuário específico).

Tomemos como exemplo o subsistema de acentuação gráfica. Erros como "~~melância~~", entre outros, parecem sugerir que muitos falantes acreditam que as palavras apresentem acento gráfico de forma arbitrária, quase como um capricho, de modo que fica inutilizado um mecanismo engenhoso que, em princípio, deveria ser fácil para o autor e útil para o leitor, que pode eventualmente desconhecer a pronúncia de uma palavra nova ou simplesmente precisar de algum tipo simples de desambigüização (p. ex. "atrib**u**i" *vs.* "atribu**í**"). Escapa-nos muitas vezes que a acentuação gráfica faz parte de um subsistema que complementa o de divisão silábica e permite identificar, sem margem de erro, a tonicidade de qualquer vocábulo bem-grafado.

Fechando o ciclo, a grafia de um "leitor precário" não codificará corretamente a totalidade das informações representadas pelo sistema e, como resultado, será mal-interpretada numa **leitura integral**. Grafias precárias, portanto, levam o "leitor integral" ao erro ou obrigam-no a "empobrecer" sua leitura. De fato, apenas uma leitura precária pode tirar "raízes" de "~~raizes~~", já que o que realmente está escrito nesse não-vocábulo é [ˈxaj.zis], algo tão diferente de [xa.ˈi.zis]

[48] Sistemas ideográficos, por exemplo, não representam nada disso, em princípio. Um exemplo menos radical seria o sistema ortográfico do inglês, que é muito mais "morfológico" (= "etimológico") que o do português e registra a pronúncia e a tonicidade de forma bastante incompleta.

quanto “pais” de “país” e que poderia ser interpretado como uma flexão de um nome “~~raize~~”/“raiz” ou até mesmo um verbo “~~raizer~~”.

Na verdade, acreditamos que boa parte da precariedade das grafias vigente no português seja devida ao fato de que uma porção da largura de banda permitida pelo sistema de escrita seja freqüentemente redundante e, assim, “desnecessária”. A exemplo de tantas outras, as frases “As ~~raizes~~ do velho ~~ipe~~ tinham ~~apodresido~~.” e “Ontem eu ~~cai~~ da ~~arvore~~.” dão dicas ortográficas suficientes para recuperar os vocábulos pretendidos, a despeito da grafia precária. Além disso, o acento gráfico em muitas palavras, como “raízes”, torna-se ainda mais “redundante” quando consideramos (i) ou que não existem palavras correspondentes de grafia idêntica salvo pela acentuação gráfica (“~~raizes~~” não existe, nem sequer [ˈxaj.zis]) ou (ii) que tais palavras existem; mas são bem menos freqüentes e, não raro, desconhecidas, quase como se não existissem (confronte “árvore” com “[que eu me] arvore”). Grosso modo, o acento gráfico não tem, por vezes, valor distintivo dentro das possibilidades oferecidas pela língua.

A interpretação exposta nesta seção da problemática que envolve o uso do nosso sistema ortográfico rende ainda alguns corolários interessantes, a saber:

- um leitor precário é insensível a muitas das regularidades do sistema e, exatamente por isso, percebe-o como uma entidade (ainda mais) caótica;
- no ensino, talvez mais valha enfatizar a leitura integral das grafias do que sua produção propriamente dita. Um leitor integral dispõe de si próprio como crítico competente de sua produção e, por exemplo, jamais grafará “parte” como “parti” ou vice-versa, *porque sabe ler os efeitos da alternância entre “e” e “i” naquelas grafias*;
- a reversão de erros *na fase ortográfica* pode ser entendida como uma leitura precária seguida de uma escrita integral. Vamos, portanto, modelar a leitura/escrita integral e identificar os pontos em que é normalmente empobrecida.

Esse último corolário poderia sugerir que um *spell-checker* deva realizar leituras sempre precárias para tornar o módulo de decodificação ortográfica mais simples e eficiente. Isso equivaleria, no entanto, a subestimar o usuário em seu conhecimento, mesmo que parcial, do sistema de escrita e ignorar as várias dicas que ele pode inserir na grafia de um não-vocábulo.

Sem dúvida, leituras precárias deverão ser freqüentemente realizadas, mas à maneira de um leitor integral, isto é, com plena "consciência" de quais pontos estarão sendo relaxados, como mais um subsídio à geração de sugestões de correção.

Divergências e inconsistências — esclarecimento e crítica

Na lexicografia consultada, existe alguma divergência quanto ao que seria o sistema ortográfico da língua. Não se trata, naturalmente, do registro de um vocábulo por um dicionário e não por outro, mas da divergência na grafia de um mesmo vocábulo *que só seria possível mediante o uso de sistemas ortográficos divergentes*, mesmo que apenas ligeiramente. Nesse sentido, o exemplo mais crítico de divergência é a grafia das flexões do verbo "delinqüir" no presente do indicativo, mais especificamente quando rizotônicas[49]. Tomemos a 3ª pessoa do singular do presente do indicativo: (Aurélio, 96) e (Houaiss, 01) grafam-na apenas como "delinqüe" e "delínqüe", respectivamente, enquanto (Guedes & Guedes, 94) registra "delínque" e ainda "delinqúi"[50].

Essas três obras acabam, por conseguinte, por supor sistemas ortográficos divergentes (i) no emprego do trema e (ii) nas regras de acentuação gráfica. Pequena que possa parecer e raro que se manifeste, a discordância entre (Aurélio, 96) e (Houaiss, 01) quanto ao subsistema de acentuação gráfica é significativa, como veremos. É que (Aurélio, 96) e seu "delinqüe" sem acento sugerem uma hipótese de acentuação muito mais elegante que a usual.

Para evitar equívocos, é importante notar que (Aurélio, 96) inclui uma reprodução do "Formulário Ortográfico" da Academia Brasileira de Letras, que, entre outros, prescreve as regras de acentuação gráfica adotadas na obra. É uma listagem bastante tradicional, em que "delinqüe" é apresentada, em observação relativa à regra de uso do trema (!), como exceção, injustificada, à

---

[49] São ditas **rizotônicas** as formas verbais acentuadas no radical, que, em português, só e sempre ocorrem no presente do indicativo, nas três pessoas do singular e na terceira do plural.

[50] Esta última obra, (Guedes & Guedes, 94), trata do português europeu, o que provavelmente explica tamanha disparidade. Note que "delinqúi" não constitui evidência de divergência quanto ao sistema ortográfico, mas apenas *quanto à própria* forma básica (pág. 70) *do verbo "delinqüir"*, que vem codificada de forma ambígua em sua grafia.

"regra de acentuação de paroxítonas terminadas em ditongo oral" (9ª regra de acentuação)[51]. Fique claro que não vamos tratar aqui daquelas regras, em que abundam exceções e casos especiais e segundo as quais "aniilo"[52] deveria ser acentuada (provavelmente porque os autores falharam em prever todas as possíveis exceções). Estamos falando de uma hipótese racional que (i) explica os dados observados de forma muito mais elegante, (ii) provavelmente norteou os idealizadores da acentuação gráfica (afinal de contas, por que fazer de "delinqüe" uma exceção e teimar em não acentuar "aniilo"?) e (iii) que tem sido descrita de forma desajeitada, talvez por condescendência[53], descrença e/ou carência de meios adequados de expressão. Talvez ainda por escolha infeliz do enfoque: se o "Formulário" tivesse tentado ensinar a ler, em vez de acentuar, como faremos aqui, o resultado poderia ter sido bem diferente.

Ensinando o computador a ler

Apesar de, a rigor, não ser fonético, nosso sistema ortográfico tem uma propriedade muito interessante: dada uma grafia correta qualquer, o usuário ideal (leitor integral) é sempre capaz de levantar algumas hipóteses de pronúncia bastante plausíveis para o vocábulo grafado, mesmo que o desconheça por completo. Além disso, tais pronúncias hipotéticas divergirão muito pouco, e raramente a ponto de impedir a compreensão por parte de um receptor que conheça o vocábulo em questão. O ponto mais crítico de divergência é, sem dúvida, o timbre (aberto/fechado) das vogais "e" e "o" em vogais tônicas orais que não formam ditongo[54].

---

[51] (Houaiss, 01) talvez tenha registrado "delínqüe" (i) ou por afronta ante a aparente arbitrariedade daquela observação ou (ii) por sua localização infeliz.

[52] Flexão de "aniilar" na 1ª pessoa do singular do presente do indicativo, pronunciado [a.niˈi.lO]. Por incrível que possa parecer, "aniilo" é a grafia inconteste dessa flexão, a despeito da 4ª regra de acentuação do "Formulário" e a ausência de disposições em contrário.

[53] As exceções injustificadas ("delinqüe") ou imprevistas ("aniilo") parecem sugerir que existe um círculo restrito de "iniciados" que conhece as verdadeiras regras de acentuação e suas justificativas, mas que as dissemina entre os "meros mortais" numa versão (quase) equivalente, mais fácil de ser seguida mas não justificada.

[54] Confronte, por exemplo, "pela" (prep.) com "cela" e "folha" com "molha" (verbo). De forma geral, essa alternância não é determinada pelo ambiente fonológico nem pela grafia. Isto é, essa ambigüidade não pode ser resolvida sem recurso ao vocábulo propriamente dito, sua morfologia e especificação fonológica plena. Alguns

De certa forma, a **componente fonético-fonológica** das informações representadas no sistema ortográfico do português serve para prover o leitor de algum material sonoro, *subespecificado*, com que consultar seu léxico interno. Tais consultas recuperarão, em geral, um conjunto de vocábulos, que será adicionalmente restringido a partir de dados contextuais ou ainda ortográficos, só que não propriamente concernentes à pronúncia, como o dado representado pela escolha da letra inicial em "sessão"/"cessão".

Nenhuma palavra é acentuada até que se prove o contrário

Vamos usar a codificação ortográfica do acento tônico como ponto de partida para nossa apresentação, apelando para a intuição do leitor e assumindo toda uma gama de conhecimentos freqüente no usuário letrado. Tratamos aqui, portanto, da componente do sistema que permite dizer que "médico" e "medico" lêem-se [ˈmɛ.di.cu] e [meˈdi.cu], respectivamente.

Apresentamos a seguir um núcleo preliminar de regras que praticamente dá conta da (de)codificação do acento tônico:

Regra 1: atribuição de acento tônico *default*

> Uma palavra sem acento gráfico assumirá o acento tônico *default*, que recairá sobre a primeira vogal gráfica (= "escrita") que o aceitar, da direita para esquerda. Não aceitará acento *default*:
>
> - a vogal das *terminações* (i) "a"/"e"/"o" seguido ou não de "s", (ii) "em"/"ens" ou (iii) "am"[55]; e
>
> - a vogal, silábica ou não, grafada como "ü"[56].

---

exemplos de casos extremos são os pares "molho" (subst., *ô*) *versus* "molho" (verbo, *ó*) e "colher" (subst., *ê*) *versus* "colher" (verbo, *ê*).

[55] Note que se trata de descrições gráficas completas: "ã", "ão" e "ais", por exemplo, não se incluem entre as terminações em questão.

[56] Há controvérsia acerca da existência de ápices silábicos grafados como "ü" no léxico ortográfico do português. Algumas palavras em que isso parece acontecer são o verbo "**argüir**" e algumas de suas flexões, tais como "(eu)

Todas as demais sílabas, em qualquer posição, aceitam o acento *default* de bom grado.

Regra 2: regra do último recurso ou da "saída estratégica"

Uma vogal será grafada com acento gráfico, talvez até *em substituição* a um trema, *se e somente se* (i) for tônica e (ii) a atribuição de acento tônico *default* aplicada a uma versão não-acentuada da grafia do vocábulo em que se insere fizer acentuar outra vogal. Ou seja, se e somente se o acento gráfico for estritamente necessário, corrigindo a leitura *default* da grafia, que não é adequada.

❖

Nosso núcleo é significativamente mais simples e sucinto que o prescrito pela didática tradicional (regras de acentuação das oxítonas, paroxítonas e proparoxítonas). Temos menos regras; e, o que é mais relevante, apenas a Regra 1 envolve memorização de listas arbitrárias, a outra constituindo algo que soa como puro "bom senso". Aliás e não por acaso, a lista arbitrária a ser memorizada na Regra 1 praticamente coincide com a constante na regra tradicional de acentuação das oxítonas.

Surpreendente que possa parecer, o núcleo tradicional pode ser derivado logicamente do nosso. Ao dizer que as terminações "a", "e", "o" e "em" seguidas ou não de "s" não recebem acento *default*, acabamos por fazer acentuar graficamente todas as oxítonas assim terminadas. Ao restringir o recebimento do acento *default* apenas por sílabas finais, acabamos por fazer acentuar graficamente todas as proparoxítonas, já que para toda antepenúltima sílaba sempre haverá uma sílaba mais à direita, a penúltima, apta a ser a tônica *default*. Por fim, de acordo com nosso

---

argüi" e "(ele) argüiu". Como não há consenso quanto à pronúncia dessas (poucas) palavras, interpretamos o **trema** simplesmente como marcando um "u" que não forma dígrafo com o "g"/"q" anterior, ou seja, que é pronunciado, seja como **vogal silábica** (vogal propriamente dita, ápice/centro de sílaba) *átona*, seja como **assilábica** (semivogal). Essa interpretação não interfere em absoluto na eficácia de nossas hipóteses nem é desmentida tampouco por nenhuma das obras consultadas, incluindo o "Formulário Ortográfico". Entretanto, talvez por mera abundância de evidência a favor, parece haver uma crença generalizada de que o trema marca um "u" não só pronunciado, mas também assilábico, ou seja, necessariamente formando ditongo. Mais de uma vez, chegamos a testemunhar pessoas que pronunciavam o verbo "argüir" como [aR.gu.ˈiR] assustarem-se com a grafia oficial dessa palavra, exclamando algo como "Mas aí está escrito [aR.ˈgw.iR]!".

núcleo, as diversas terminações [i(s), u(s), n, ã, etc.] listadas na regra tradicional de acentuação de paroxítonas gerarão sílabas finais receptivas ao acento *default*, de forma que os vocábulos paroxítonos com essas terminações deverão receber acento gráfico, ou serão lidos como oxítonos.

Além disso, nosso núcleo tem maior cobertura e é mais razoável (= apresenta menos exceções). Basta considerar que aqui já resolvemos o caso "delinqüe", antes excepcional, uma vez que, nessa palavra, o acento tônico *default* recai exatamente onde deveria. Outro caso já coberto por nossa hipótese e que tradicionalmente requer uma regra à parte é a acentuação gráfica do padrão *g/q-ú-e/i*, ou seja, "u" tônico precedido de "g"ou "q" e seguido de "e" ou "i", como em "(ele) argúi". Tomemos o exemplo em questão para demonstrar isso: em primeiro lugar, uma versão não-acentuada de "argúi" não poderia ser "argui", já que esta se lê [aRˈgi]; antes, deveria ser "argüi", em que o "u" é pronunciado mas não aceita o acento *default*. Dessa forma, como "argúi" sem acento gráfico será lido [aR.guˈi] ou [aRˈgwi], temos de substituir o trema pelo acento agudo para corrigir a leitura *default*.

Por fim, vale notar que, como recomendado anteriormente, nossas regras se diferenciam por enfocar a *leitura de grafias*, e não a mera acentuação gráfica de palavras já conhecidas. Temos de reconhecer, naturalmente, que as regras tradicionais também permitem derivar regras de leitura ou simplesmente justificar a leitura dos vocábulos sem acento gráfico, mas não sem uma boa dose de esforço adicional.

A conspiração das vogais: encontros vocálicos e tuíuíús

Tudo estaria resolvido não fosse um "pequeno detalhe", a saber: nossas regras e as tradicionais são expressas em termos como "sílaba", "número de sílabas", "rima" (parte de sílaba), "**ditongo**" e "**hiato**" (vogais numa mesma sílaba[57] ou em sílabas diferentes), etc., isto é, *direta ou*

[57] Um **ditongo** pode ser entendido como uma "vogal deslizante", que varia no tempo de uma qualidade vocálica para outra (Cagliari, 92). Acontece que, em qualquer ditongo, a variação é monotônica, mas *não-linear*, de forma que sempre uma das qualidades é claramente "dominante" (é mais acentuada, tem maior duração, etc.), a outra surgindo num "deslize" relativamente curto. Daí o termo inglês *glide* para essa qualidade "dominada".

indiretamente, todas assumem algum tipo de divisão silábica. À primeira vista, trata-se de um requisito trivial. É bem verdade que, em específico, a divisão silábica de palavras sem encontros vocálicos é realmente muito simples, cada vogal gráfica[58] correspondendo a um ápice silábico[59], devido à própria natureza fonética do português. No entanto, quando ao menos duas vogais (gráficas) se encontram, elas conspiram para dificultar um pouco as coisas.

Retomemos, agora, os pares-exemplo "c**a**|**i**n|do" *vs.* "p**ai**|na" e "c**a**|**i**r" *vs.* "C**ai**|ro". Considere ainda "c**ai**", "c**a**|**iu**", "c**ai**|**o**", "p**au**", "p**a**|**ul**", "constit**ui**" e "f**u**|**i**nha". Como se pode perceber, os grafemas "i" e "u" são bastante ambíguos, ora sendo receptivos ao acento *default* e freqüentemente constituindo hiato com a vogal precedente, ora passando o acento adiante e sendo preferencialmente proferidos como semivogais em ditongos decrescentes[60]. Logo, para que as regras de codificação do acento tônico possam ser aplicadas, essa ambigüidade tem que ser resolvida, o que se faz por meio da seguinte regra:

Regra 3: divisão silábica

> Os grafemas "i" e "u" corresponderão a ápices silábicos e, portanto, serão receptivos ao acento *default* se somente se:

Uma distinção clássica é feita entre **ditongos crescentes** e **decrescentes**, que se caracterizam respectivamente pela ocorrência do *glide* ou na cabeça (ex.: "iá", o grito do carateca, e...) ou na cauda (... "ai", o grito de sua vítima) do ditongo

[58] Às vezes, o termo "gráfico" se faz necessário para distinguir entre os grafemas "a", "e", "i", "o" e "u" ("vogais gráficas") e as vogais (fonéticas) propriamente ditas.

[59] A grande dificuldade da divisão silábica reside na identificação dos ápices silábicos, principalmente quando não estamos preocupados com a divisão em sílabas gráficas.

[60] Usamos "preferencialmente", porque há alguns poucos casos em que há bastante liberdade quanto à escolha entre ditongo ou hiato na pronúncia de um tal encontro vocálico. Confronte, por exemplo, "fui" com "possui" e "constitui". Não seria aceitável, e talvez até mais provável, pronunciar estes dois últimos com hiato, enquanto na pronúncia de "fui" tem-se uma preferência indiscutível pelo ditongo? Esse fenômeno será discutido em mais detalhe na seção sobre a *neutralização absoluta entre ditongos e hiatos* (pág. 75).

- precedidos de um elemento *gráfico* com que não possam *jamais* formar uma seqüência representativa de um ditongo[61]. Para facilitar o entendimento dos exemplos a seguir, diremos que, ao satisfazer essa condição, o "i" ou "u" em questão estará "fora de suspeita" ou que haverá "**suspeita**", caso contrário; ou

- seguidos de "nh" ou um "resto/término de sílaba" (**coda**) que não seja vazio ou grafado como "s"[62]; ou

- acentuados graficamente.

Os demais grafemas vocálicos ("a", "e", "o", "am", "ão", etc.) sempre representam ápices.

❖

Vale notar que, com a adição da Regra 3, temos uma hipótese que cobre toda a codificação do acento tônico[63] do português, bem como os maiores desafios da componente fonética do nosso sistema ortográfico.

---

[61] Costumeiramente, um tal elemento gráfico precedente é um "início de sílaba" (**onset**) composto de uma ou mais consoantes. Uma alternativa muito pouco freqüente é uma vogal que acabe por não formar um grupo representativo de um ditongo válido na língua, o que acontece, por exemplo, no verbo "aniilar" e suas flexões ("an**i**|**i**lo", "an**i**|**i**la", "an**i**|**i**le", etc.).

[62] É interessante notar que, à maneira da parte referente ao "nh" (já comentada à pág. 57), o resto dessa condição é quase que natural: nenhuma coda, exceto /Ø/ e /S/, seria fonologicamente aceitável numa sílaba que já contivesse um ditongo. Por exemplo, a sílaba fonológica /kajR/ simplesmente não ocorre em português, o que implica que não há necessidade de um acento gráfico para estabelecer que "cair" deva se ler [kaˈiR].

[63] Note bem o uso de "codificação do acento tônico" em vez de "acentuação gráfica". É que esta ainda se preocupa com a acentuação dos monossílabos, com convenções sem justificativa (acentuação da vogal tônica nos hiatos "ôo/êe") e com a codificação de timbre (aberto/fechado) tanto na escolha do diacrítico específico (agudo/circunflexo) das vogais tônicas cujo acento gráfico já é prescrito por nossas regras, quanto na prescrição de **acento (gráfico) diferencial**. Vale notar que o diacrítico diferencial sempre incide sobre vogais que já recebem acento tônico *default*, motivo por que incluímos a regra relativa aos ditongos tônicos abertos ("éu", "ói" e "éu") entre as de acentuação gráfica diferencial.

Não cobriremos a acentuação gráfica na íntegra, mas garantimos que as regras que faltam podem ser quase que *diretamente* incorporadas a nosso núcleo.

Agora podemos decodificar a presença de hiato em "r**a**|**i**z", "c**a**|**i**n|do" e "c**a**|**i**r" e de ditongo em "~~r~~**ai**|~~zes~~", "p**ai**|na" e "C**ai**|ro". Em cada uma das três primeiras, existe uma coda necessariamente fazendo do "i" um ápice, o que o torna um receptor em potencial do acento tônico *default*. Nas três últimas, por sua vez, nenhuma das condições para "destacar" o "i" é satisfeita, uma vez que os grafemas "z", "n" e "r", respectivamente, passam a representar um "início" de sílaba (onset). Aqui já temos uma amostra da complexidade da codificação silábica, uma vez que, por vezes, é necessário fazer uma divisão silábica parcial identificando algumas codas e onsets, por exemplo, para desbloquear pontos da própria divisão silábica. Para isso, o usuário equaciona os possíveis valores para os grafemas com as possibilidades fonéticas[64] da língua e chega, necessariamente, a algumas certezas, suficientes para resolver o "sistema de equações" embutido em cada grafia.

Melhor explicando e, para tanto, permitindo-nos uma analogia computacional, a complexidade reside no fato de o processamento da divisão silábica ser distribuído e dirigido pelo fluxo de dados (ou "**fluxo de certeza**"), a computação se propagando dos pontos de certeza (seqüências gráficas não-ambíguas) para os de incerteza. É como uma rede em que cada ponto de certeza contribui para a resolução dos pontos de incerteza que lhe são adjacentes, os quais, tão logo deixam de ser incertos, propagam certeza para outras regiões ainda incertas. Dois casos que ilustram bem essa visão de propagação ou recursividade são "caiu" e "tuiuiú". Em ambos os casos, a questão fundamental é como "distribuir" ápices e semivogais entre tantas vogais gráficas.

Analisemos primeiro "caiu", um caso mais fácil. Inicialmente, temos uma configuração do tipo "c a̲ i̯ u̯"[65], em que só é certo que o "a" é um ápice silábico. Nesse ponto, não é possível ainda resolver o *status* do "i", pois, embora parte das condições (acento gráfico, etc.) suficientes para torná-lo um ápice tenham falhado, ainda não sabemos se há uma coda que vá destacá-lo. Por outro lado, já é possível inferir a situação do "u": já que todas as condições falham[66] da disjunção

---

[64] Para uma descrição sistemática, veja Câmara Jr. (70).

[65] Esta é uma notação original e criada meramente para apoiar a compreensão dos exemplos subseqüentes. Legenda: $\underline{x}$ - ápice silábico; $\underline{x}$ - situação indefinida; $\underline{x}$ - vogal assilábica.

[66] Em específico, o "i" que antecede o "u", apesar de ainda estar em situação indefinida, é um predecessor com que este poderia formar um ditongo decrescente, o que descarta a primeira condição suficiente para fazer do "u" um ápice.

necessária para fazer dele um ápice, sabemos que se trata de uma semivogal. Temos então a configuração "c a̯ i̯ u̯", que permite resolver o resto de dúvida existente: a semivogal representada por "u" é exatamente a coda que destacará o "i", o que está expresso em "c a̯.i̯ u̯". Finalmente, como "i̯ u̯" é a sílaba mais à direita receptiva ao acento tônico *default*, é aí que ele recai, correspondendo à pronúncia que "caiu" pretende representar. A configuração final é, portanto, "c a̯ ˈi̯ u̯".

A divisão silábica da palavra "**tuiuiú**", por sua vez, tem seu fluxo de certeza representado na Figura 3. É válido verificar o que aconteceria se "tuiuiú" não tivesse acento, não só para esclarecer de vez a aplicação da regra de divisão silábica, mas também para atestar a necessidade estrita de acentuação gráfica nessa palavra (Regra 2, "do último recurso"). Por isso, ainda na Figura 3, apresentamos o fluxo de certeza da divisão silábica de "~~tuiuiu~~", em que é recorrente o padrão encontrado em "caiu".

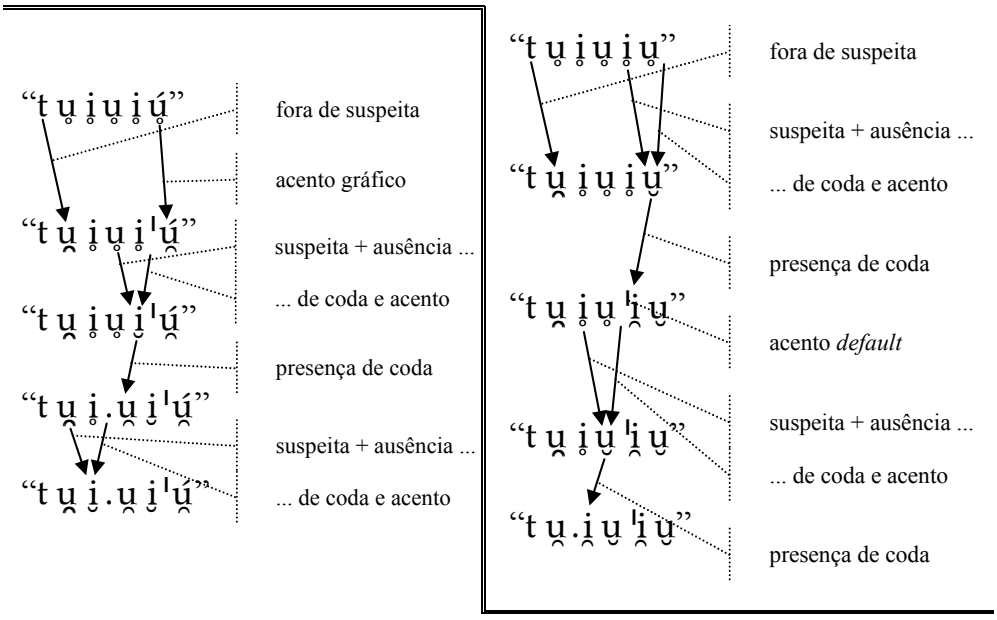


Figura 3: Fluxo de certeza na divisão silábica de "tuiuiú" e "~~tuiuiu~~".

O (mais que atestado) acento gráfico em "tuiuiú" é ainda mais digno de nota por se tratar de um caso que fica sem explicação em muitas formulações das regras de acentuação gráfica, como as apresentadas por Bechara (92), de André (94), Faraco & Moura (94), Rocha Lima (92) e Sacconi (92). São versões que incluem uma regra prescrevendo acento gráfico ao "'i' e 'u' tônicos que formam hiato com a vogal anterior", a qual seria a única cogitável para explicar "tuiuiú".

No entanto, tal regra não se aplica, já que não existe hiato nessa palavra[67]. O Formulário Ortográfico não comete esse deslize, incluindo uma regra alternativa que acentua o "'i' e 'u' tônicos *que não formam ditongo com a vogal anterior*". No entanto, é exatamente essa regra que peca por tentar acentuar "(eu) aniilo", outra grafia atestada. Nossas regras, por sua vez, explicam ambos os casos de forma natural, o que realmente sugere que a idéia de acento gráfico como último recurso não é apenas uma diretriz, mas uma *regra fundamental* na codificação do acento tônico no português.

Agora temos um conjunto de regras que explica o acento gráfico em "colégio", por exemplo, mas não pela tradicional "acentuam-se as paroxítonas terminadas em ditongo crescente oral", a começar pelo fato de que, de acordo com nossas regras de divisão silábica, *simplesmente não existem ditongos crescentes*, sejam orais ou nasais, o que é conseqüência direta da primeira condição imposta pela Regra 3. No caso de "colégio", o "i" é considerado representativo de um ápice silábico por ser precedido do onset grafado como "g"; sem o acento, portanto, leríamos [ko.le.ˈʒi.u], que *poderia*, com efeito, corresponder à pronúncia da flexão "(eu) colegio" de um suposto verbo "colegiar". O caso "história" é perfeitamente análogo, contando ainda com a flexão "(ele) historia", do verbo "historiar".

Generalizando: no esquema aqui proposto, ditongos crescentes não são, em princípio, representáveis, existindo apenas certo tipo de hiato que (i) *pode ser pronunciado como um ditongo crescente em pronúncia fluente* e (ii) cuja vogal mais à esquerda é sempre receptiva ao acento tônico *default*, como todo bom penúltimo ápice silábico. Como resultado, a maior parte das supostas paroxítonas terminadas em ditongo crescente jamais poderia ser grafada sem acento gráfico, uma vez que o acento *default* não seria capaz de ultrapassar o "ditongo" em questão[68].

---

[67] Como bem define Câmara Jr. (92), um hiato é um "efeito *acústico* produzido pela enunciação imediatamente seguida de duas vogais *silábicas*", algo que jamais ocorre em "tuiuiú". Nessa palavra, os grafemas "u" representam todas as vogais silábicas, as quais jamais são enunciadas em seguida, pois há sempre um "i" assilábico a entremeá-las.

[68] Defendemos "supostas" porque tais paroxítonas estariam mais para proparoxítonas, no esquema proposto. Como veremos logo a seguir, essa distinção (paroxítonas *versus* proparoxítonas) não é assim tão relevante ou mesmo

Mais uma vez, como na prescrição de um grafema único para representar os fones [t] ("par**t**o") e [t͡ʃ] ("par**t**e"), temos uma regra ortográfica que parece espelhar uma análise fonológica: a hipótese de ausência de (boa parte dos) ditongos crescentes fonológicos/fonêmicos em português não é nenhuma novidade; aliás, parece ser ponto pacífico entre os fonólogos (Câmara Jr., 70 e 77; Cagliari, 97b). Caso essa hipótese não pareça razoável, isso provavelmente se deve a pouca familiaridade com a distinção clássica (e não-trivial) entre os níveis *fonológico* e *fonético* da descrição lingüística, que tentaremos delinear a seguir e cujo domínio é simplesmente vital para o entendimento de todo o resto do trabalho.

Do fonológico e do fonético

A dicotomia "**fonêmico/fonológico**" *vs.* "**fonético**" foi uma boa solução encontrada pelos lingüistas para explicar, de forma elegante (= econômica), fenômenos diversos, tais como a instabilidade relativa da pronúncia dos vocábulos[69] ante sua estabilidade nos processos morfológicos e vice-versa, entre outros. Além disso, constitui um modelo de análise que faculta toda uma simplificação na descrição dos processos morfológicos, revelando uma regularidade subjacente muito maior que a aparente.

No **nível fonêmico/fonológico**, abstrato, cada vocábulo tem uma *representação única* — a **forma básica/fonêmica/fonológica** — (i) a partir da qual todas as suas variantes de pronúncia podem ser inferidas e justificadas por meio de regras gerais (processos fonológicos) e (ii) que serve de ponto de partida para as diversas operações morfológicas e a conseqüente geração de outras formas básicas. Por sua vez, o *nível fonético*, concreto, é o nível da enunciação, em que as potencialmente diversas variantes de pronúncia de uma forma básica se realizam.

A forma básica de um vocábulo representa, assim, sua face estável e permite evitar com simplicidade todo um conjunto de complicações teóricas. Tomemos, por exemplo, a palavra

---

clara para nós. "Maior parte" aqui se justifica pelo caso "delinqüe", que consideramos uma das poucas paroxítonas *realmente* terminadas em ditongo crescente *mas que, ironicamente, não é acentuada*.

[69] Referimo-nos aqui às variantes de pronúncia dos vocábulos, *em especial dentro de um mesmo dialeto*. Por exemplo, é muito comum que as variantes [ˈca.ʃɜ] e [ˈcaj.ʃɜ] do vocábulo "caixa" coexistam pacificamente e não causem nenhum problema de decodificação ou estranheza.

"**sal**". Em muitos dialetos do português, correta e exclusivamente[70] a pronunciamos **[**saw**]**, uma manifestação no nível fonético e, portanto, representada entre **colchetes**. Numa abordagem mono-nível, fica complicado explicar por que de [saw] derivamos [saˈlej.ɾu] e flexionamos [sajs], em vez de [saw.ˈlej.ɾu] e [saws] (cf. "mau" *vs.* "maus"). Levantando a hipótese de níveis subjacentes, entretanto, podemos supor uma forma básica única /sal/, grafada entre **barras** exatamente por habitar esse nível (fonêmico) e a que implicitamente corresponde a forma fonética [saw] se supusermos também duas **regras (processos) fonológicas(os)**: uma — intrínseca da língua — que cria uma articulação secundária velar para *fonemas* /l/ de final de sílaba, gerando o fone [ɬ] (que corresponde ao que se ouve, por exemplo, no português europeu, segundo Cagliari [97b]); e outra — dialetal — que desfaz, em seguida, a articulação primária original, lateral, gerando *fones* [w][71] (Cagliari, 97b). Assumindo a forma básica como ponto de partida para os processos morfológicos, fica fácil explicar, por exemplo, [saˈlej.ɾu]: a forma básica dessa palavra é formada adicionando-se as formas básicas dos morfemas envolvidos, a saber, /sal+ej.ɾ+U_{vt}+$Ø_{masc}$+$Ø_{sing}$/. Por regras diversas (de ajustamento silábico, combinação de morfemas, atribuição de acento, etc.), com prioridade sobre as que geram [w] de /l/ em final de sílaba, temos a nova configuração /saˈlej.ɾU/. Nessa nova posição silábica, o **segmento** (fone/fonema) /l/ não está mais no ambiente de aplicação da regra de velarização, ou seja, não mais se realiza como [w], o que impede o surgimento da forma [sawˈlej.ɾu].

Os dados fonéticos não são hipotéticos, antes têm manifestação física e podem ser comprovados empiricamente, por meio de equipamento de registro e análise adequado. Os dados

---

[70] Exceto quando estamos intencionalmente fazendo referência à sua ortografia, como na locução "Não se escreve [saw], mas [sal]!"

[71] Está aqui implícita a distinção entre "**fonema**" e "**fone**": o primeiro habita o nível fonêmico; e o último, o fonético. Não raro, um mesmo fonema (no exemplo, o /l/), pode acabar por se realizar foneticamente como diferentes fones ([l] ou [w] pós-vocálicos), ditos *alofones* desse fonema, quer livremente (como no exemplo), quer em função do contexto fonético. Um exemplo desse último caso, em português, ocorre com o fonema /t/, que, em muitos dialetos, terá por alofones ora [ʧ], ora [t], em função de ser seguido ou não por uma vogal palatal ([i]).

fonêmicos/fonológicos, por outro lado, *não podem ser ouvidos* (!) e não passam de uma hipótese para explicar os dados fonéticos e os padrões que insistem em aparecer nesses dados. Assim, a forma básica de um vocábulo não pode ser diretamente depreendida de uma única variante de pronúncia, ou seja, precisa ser inferida a partir de todo o conjunto de variantes de pronúncia de um dado vocábulo, bem como de diversas relações (morfológicas, contrastivas, etc.) que este estabelece com os demais vocábulos da língua. De brinde um belo insight final, devido a Lass (94): o **objetivo último** da forma básica é reter somente o que é arbitrário no significante, descartando *tudo* que pode, de alguma forma, ser considerado (foneticamente) motivado.

É importante notar que a explicação acima, não obstante tratar de uma dicotomia consensual em fonologia, é definitivamente partidarista, remontando a Leonard Bloomfield e à sua *morfofonêmica*[72] *processual* (Lass, 84). Entre os diversos compromissos teóricos que estamos assumindo, é oportuno atentar para os seguintes:

- temos na forma básica uma **estrutura analisada morfologicamente** e "impronunciável", por conter tanto marcas de fronteira de morfema ("+") quanto estruturas silábicas foneticamente impossíveis, porque incompletas, e fonemas subespecificados, incluindo arquifonemas[73]. Cabe notar aqui que todas as análises morfológicas e fonológicas presentes neste trabalho são (i) mais ilustrativas que definitivas e (ii) à base de Câmara Jr. (70, 77), Cagliari (97a, 97b) e Monteiro (86), bem como idéias próprias;

- temos, na verdade, um **número indeterminado de níveis**, cada processo (regra) fonológico "desaguando" num novo nível. Portanto, estritamente falando, o fonema habita apenas a forma básica, que jaz no nível de partida, por definição. Por vezes, entretanto, insistiremos em "fonema" ou no uso de barras (//) acima desse nível (como fizemos no exemplo e parece

---

[72] Termo proveniente da fusão de "morfologia" com "fonêmica", a que "fonológico/fonêmico" foi historicamente se igualando.

[73] Um **arquifonema** surge da **neutralização** de uma ou mais oposições fonológicas, representando, assim, uma classe de fones que podem se revezar numa dada posição, seja (i) livremente, como /U/, que se realiza como [o] ou [u] (ex.: "carr<u>o</u>") ; seja (ii) de forma dependente de contexto, como /S/, cujo vozeamento é determinado pelo segmento seguinte (ex.: compare "ci<u>s</u>ma" e "ci<u>s</u>co"). Como se pode observar nos exemplos, arquifonemas são usualmente transcritos em maiúsculas.

praxe), quando estaremos nos referindo a segmentos que ainda não sofreram nenhuma "mutação" desde a forma básica até o nível considerado;

- assume-se uma **ordenação de processos/regras**, definida pelo fonólogo, segundo a qual um processo só terá oportunidade de ser aplicado após exauridas as possibilidades de aplicação dos processos obrigatórios (isto é, não opcionais) de ordem inferior;

- ao longo de potencialmente muitos processos/níveis, perpetuam-se, em princípio, estruturas ainda não prontas para pronúncia. Trata-se de **processos/níveis pré-lexicais**;

- uma forma já pronta para pronúncia (fonética) ainda é passível de processos, muitas vezes opcionais. Trata-se de **processos/níveis pós-lexicais**. Um tal processo é a nasalização de vogais átonas pretônicas seguidas de fone nasal, a qual é opcional, como se pode perceber pela existência das variantes de pronúncia [maˈmẽw̃] e [mẽˈmẽw̃] da palavra "mamão";

- existe a possibilidade de **neutralização absoluta**, isto é, de oposições que só existem no nível fonêmico, jamais sendo percebidas no nível fonético. Por exemplo, apesar de haver homofonia entre "mau" e "mal" ([maw]) em muitos dialetos do português do Brasil, assumimos haver contraste entre suas formas básicas (/maw/ e /mal/, respectivamente) para toda a língua portuguesa, de modo a explicar os plurais /mawS/ e /malIS/.

  A licença à neutralização absoluta, enquanto pressuposto teórico, é objeto de crítica de várias hipóteses fonológicas alternativas. Reclamam os críticos que, assim, torna-se teoricamente possível derivar diversas formas básicas alternativas para uma mesma forma fonética. No entanto, como essa parece ser exatamente a causa primeira de uma vasta e, o que é melhor, aparentemente heterogênea coleção de erros ortográficos, não hesitamos em adotá-la neste trabalho. Como veremos, os erros em "(eu) ~~rapito~~" ("rapto"), "~~degrais~~" ("degraus"), "~~anães~~" ("anões/anãos"), "(ele) ~~róba~~" ("rouba"), "(ele) ~~salda~~" ("saúda"), "(eles se) ~~imíscuem~~" ("imiscuem") podem ser reconstituídos de forma (i) uniforme, (ii) verossímil e desafiante para a maioria dos perfis de usuário e (iii) otimista, numa hipótese de erro único (de *inferência de forma básica*).

Made in Taiwan
ou Os ditongos e hiatos de "Paraguai"

Pois bem, após essa não tão breve digressão, estamos em posição de explicar claramente nossa posição acerca de muitas questões. Em primeiro lugar, corroborar definitivamente a hipótese da inexistência de ditongos crescentes fonológicos. Na verdade, o único ***onglide*** (do inglês "*onset* + *glide*", ou seja, uma semivogal em início de sílaba) fonológico que normalmente se aceita é o fonema /w/ (grafado "u/ü") precedido de /g/ ou /k/. De fato, existe um excelente argumento para se admitirem ditongos crescentes na forma básica de alguns verbos como "delin**qüi**r" e "a**gua**r", o mesmo, aliás, para os banir das palavras "obli**qu**|**a**r" e "ar**gü**|**i**r". Já aqui se pode pressentir algo de grande: se tal argumento for válido, o par "delinqüir" e "obliquar" terminará por atestar a **neutralização absoluta de uma oposição hiato-ditongo**, uma reivindicação que acreditamos meio polêmica.

É o seguinte: certas flexões verbais do presente do indicativo — a saber, as pessoas do singular e a terceira do plural — têm a peculiaridade de incluírem um deslocamento do acento tônico para *a vogal silábica mais à direita da forma básica do radical verbal*, resultando nas então chamadas formas rizotônicas. Essa é uma das marcas registradas da língua, presente em todas as suas variedades dialetais, e um dos motivos para que comumente atribuam a "raptar", pronunciado no Brasil [xa.piˈtaR], a forma básica /xa.p.t + a_{vt} + R_{dmt}/ e a "apitar", pronunciado [a.piˈtaR], a forma /a.pi.t + a_{vt} + R_{dmt}/. Temos aí um caso de neutralização absoluta de um fonema, /i/, com sua própria ausência, bem como a explicação de por que devemos flexionar "(eu) rapto/apito" e não "~~rapito~~/~~apto~~".

Nesse cenário, vale a pena estudar a situação de "obliquar", flexionado "(eu) obliq**u**o". Se supusermos, numa análise mais tradicional, que sua forma básica seja /o.bli.kw + a_{vt} + R_{dmt}/, seremos obrigados a considerá-lo um verbo (bastante) irregular, ou senão teríamos que aceitar a flexão "(eu) ~~oblíquo~~" (cf. "delinqüir" — /de.liN.kw + i_{vt} + R_{dmt}/ — e "(ele) del**i**nqüe"). A irregularidade, no caso, reside na necessidade de assumirmos um alomorfe /o.bli.k**u**/ do radical, específico para as formas rizotônicas. Estaríamos, então, colocando "obliquar" em "pé de irregularidade" com "odiar", por exemplo.

Há uma alternativa, entretanto, bem mais elegante e atraente — em especial para nossos propósitos — que permite considerar "obliquar" um verbo perfeitamente regular. Consiste em assumir a forma básica /o.bli.k**u** + a_{vt} + R_{dmt}/, *às custas de aceitar a neutralização absoluta da oposição ditongo-hiato nesse contexto*. Porque, segundo as regras usuais de adaptação de morfemas na flexão verbal, chegaríamos à forma pré-lexical /o.bliˈku.U/ ("obliquo"), mas também a /o.bli.kuˈej/ ("obliqüei"), em lugar do tradicional /o.bliˈkwej/. Ou seja, teríamos que derivar a pronúncia mais que atestada [o.bliˈkwej], com ditongo, de /o.bli.kuˈej/ sem jamais passar por [o.bli.kuˈej]. Será isso tão sério?

Argumentamos que não e vamos além: defendemos (i) a neutralização absoluta *de toda e qualquer oposição fonológica ditongo-hiato* e (ii) que o que sobra de hiatos e ditongos na forma básica é tão-somente a tonicidade dos elementos. Pense em várias palavras que contenham ditongo tônico, como "p**o**uso", "c**o**isa", "pap**é**is", "f**o**i", "obliqu**a**r", "obliqü**e**i", etc., e experimente *pronunciá-las com hiato, mantendo, no entanto, a tonicidade e um mínimo de naturalidade* (por exemplo, em vez de [ˈpow.zu], pronunciar [ˈpo.u.zu]). A impressão não é mais ou menos a mesma que pronunciar [ˈka.ɣo] ("carro") em vez de [ˈka.ɣu]? Ou seja, as palavras soam meio estranhas, mas nunca acabam por ser confundidas — sintoma certo de neutralização. Supomos, nesse caso, que ocorra algo típico da neutralização (absoluta ou não): a preferência inconteste por uma das formas neutralizadas, como a pronúncia de "rapto" como [ˈxa.pi.tu], e não [ˈxap.tu], ou a de "carro" como [ˈka.ɣu]. No caso em questão, a forma preferida é o ditongo, o que já ocorria quando da pronúncia de hiatos fonológicos como ditongos crescentes. E que atire a primeira pedra quem preferir [ˈmaR.si.u] a [ˈmaR.sju] ("Márcio") e [ma.goˈaR] a [maˈgwaR] ("magoar")!

Um bom contra-exemplo seria o par "(eu) **vôo/vou**" em que a oposição ditongo-hiato não parece neutralizada. Contra-atacamos com "**possui/fui**" (!), em que a tradição gramatical só faz ouvir ditongos. No entanto, em pronúncia relativamente lenta, como a que evidencia a oposição entre "v**ô**|**o**" e "v**ou**", haverá (a mesma) oposição entre "poss**u**|**i**" e "f**ui**", o que acreditamos ser um dos motivos do tão freqüente "~~possue~~". Atribuímos essa oposição ao fato — mais psico que lingüístico — de o falante manipular meio que conscientemente os morfemas flexionais realizados como o segundo "o" de "vôo" e o "i" de "possui" e estar mais acostumado a usá-

los como ápice sílabico, já que são respectivamente os mesmos morfemas encontrados em "cant**o**" e "toss**e**" e *tantos* outros verbos sem hiato. Na corrente da fala fluente, entretanto, essa oposição é certamente minimizada, se é que não desaparece de todo.

Por fim, um contra-exemplo desesperado seria "boa", com vistas a dar prova da existência de hiatos estáveis. Com efeito, não podemos refutar. Nem queremos, lembrando que jamais cogitamos que estes não existissem. Admitimos que os há aos montes — "baú", "jaú", "(eu) caí/roí", etc. — com a devida nota de que é inconcebível, no português do Brasil, pronunciar uma versão ditongada de quaisquer dessas palavras, simplesmente porquanto não haja um glide correspondente ao "a". Ou seja, o ditongo só não está lá porque não consegue.

A situação fica ainda mais tranqüila quando consideramos ditongos átonos, que nos rendem exemplos ainda mais convincentes. Tente pronunciar "arruinar", "ajuizar" e "cuidar" com hiato: as pronúncias alternativas nem mais soam estranhas, se é que soam alternativas! Flexionemos agora ambas as palavras — "(eu) arru**í**no/aju**í**zo/c**u**ido" — e perceberemos uma notável diferença de comportamento. Considerar irregulares quaisquer desses verbos parece exagero, principalmente quando nos lembramos de que "arruinar" e "ajuizar" são derivados de "ruína" e "juízo", respectivamente, ambos com hiato na forma básica. Nada mais natural que aceitar a "herança" do hiato pelos verbos derivados, o qual, uma vez privado de sua tonicidade, passa a ser pronunciado não só como ditongo, mas, em ditongo, tanto crescente quanto decrescente[74]!

Em conclusão, postulamos hiatos nas formas básicas de "obliquar", "argüir", "arruinar", "ajuizar" e "saudar"[75] e ditongos naquelas de "delinqüir", "aguar", "cuidar", "curto-circuitar", "pausar", de forma que reivindicamos a neutralização absoluta da oposição fonológica ditongo-hiato. Como veremos no próximo capítulo, isso nos permite, entre outros, corrigir "(eu) ~~saldo~~/~~agôo~~" ("saúdo/águo") de forma lingüisticamente embasada e, por isso mesmo, sem acarretar hipóteses absurdas de correção em outros casos, tais como "~~paú~~/~~ênjuo~~" para "~~pal~~/~~enjuo~~" ("pau/(eu) enjôo").

---

[74] Essa neutralização ocorre por se tratar de um ditongo átono cujos ingredientes são as vogais /i/ e /u/, ambas passíveis de assumir o papel de glide.

[75] Contrariamente à nossa crença original, "saudar", de acordo com o Aurélio, não é derivado de "saúde". No entanto, ambos têm, em última análise, origem no latim "*salute*" e certamente passaram por processos idênticos de evolução fonológica para chegarem às formas atuais.

Além disso, aqui terminamos de reunir os elementos para os importantes resultados da seção seguinte.

Lançar âncoras!

Ao longo deste capítulo, fomos fazendo apontamentos quanto à natureza do nosso sistema ortográfico, bem como apontando/sugerindo alguns dos desafios que ele impõe a seus usuários. Vamos agora apresentar uma síntese dos principais resultados que decorrem de nossos apontamentos, ressaltando alguns pontos (de erro) geralmente mal-apreendidos.

Âncora fonológica

Em primeiro lugar, notamos que nosso sistema tem uma **vocação** muito mais **fonológica** que fonética. Isso já começa a ser evidenciado quando ele se furta a representar muitos detalhes fonéticos redundantes, isto é, que podem ser deduzidos pelo contexto, como, por exemplo, a nasalidade do "a" tônico de "cama"[76]("~~cãma~~"), a articulação secundária desenvolvida pelo fonema /d/ em "tar**d**e" ("~~tardji~~"?), um casual ditongo nasal em "ganho" ("~~gainho~~" ou "~~gaio~~", ambos com ou sem til) ou outro, freqüentíssimo e oral, em "arroz" ("~~arrois~~")[77].

No entanto, uma afinidade com a forma básica (ou com formas pré-lexicais em níveis avançados, já que a forma básica sofre, muitas vezes, duras mutações[78]) fica realmente patente quando atentamos para a maneira com que nosso sistema ortográfico lida com muitos casos de neutralização absoluta e variação livre. Muitas vezes, representam-se oposições que foneticamente se desfazem, mas que são relevantes em processos morfológicos. Constituem exemplos claros disso os seguintes pares: "mal/mau" ("males/maus"), "fogaréu/anel" ("fogaréus/anéis"), "estourar/escorar" ("[eu] est**ou**ro/esc**o**ro"), "peneirar/esperar" ("[eu] pen**ei**ro/esp**e**ro"), "raptar/apitar" ("[eu] r**a**pto/ap**i**to"), "dignar/enguiçar" ("[eu] d**i**gno/engu**i**ço"),

[76] Toda vogal tônica seguida de segmento nasal se torna necessariamente nasal. Por conseguinte, a vogal tônica de "cama" não costuma ter sua nasalidade representada em sua forma básica.

[77] Note que "arrozal" não herda o ditongo de "arroz", o que evidencia a natureza meramente fonética deste ditongo.

[78] Isso se deve à vocação/âncora fonética do nosso sistema, resumida a seguir.

“imiscuir/delinqüir” (“[eles] imisc**u**em/del**i**nqüem”), “recuar/apropinquar” (“[ele] rec**u**a/aprop**í**nqua”) e “magoar/aguar” (“[eu] magôo/águo”).

Como também já vimos, a fidelidade à forma básica é parcial, ou seja, nem toda neutralização absoluta é ortograficamente desfeita. A oposição fonológica ditongo-hiato, por exemplo, não é representada quando temos glides grafados como “i” ou “u” (“saudar/arruinar” *vs.* “pausar/cuidar”). Adicionalmente, enquanto o grafema “u” precedido de “c” jamais representa um glide fonológico (“imiscuir”, “recuar”, etc.), o oposto não é certo quando é “g” ou “q” o predecessor, o que se deve tão-somente às exceções “obliquar”[79] e “argüir”.

Esses e quaisquer outros pontos de conexão entre ortografia e fonologia podem ser ignorados pelo usuário; e, quanto mais numerosos, tanto mais o usuário tende à **transcrição fonética**, que corresponde a ancorar a ortografia não na forma básica ou algum nível pré-lexical, mas em níveis pós-lexicais. Um tipo mais difícil de transcrição fonética surge quando o usuário elege conjuntos alternativos de aspectos (fonéticos) a serem representados, chegando a escrever “~~cãma~~” (“cama”), “~~patio~~” (“patinho”), “~~gaia~~” (“ganha”) e assim por diante.

Âncora étimo-morfológica

Provavelmente o maior determinante da vocação fonológica de nosso sistema ortográfico seja sua **vocação étimo-morfológica** : na medida do possível e com exceções, nossa **unidade de grafia** é o morfema, em vez do vocábulo. Ou seja, tenta-se maximizar a consistência da grafia de um mesmo morfema dentro do conjunto de todos os vocábulos dos quais ele participa. Assim, se grafamos “cruel” com “l” por motivo fonológico, consistentemente grafaremos “crueldade”, mesmo que, neste derivado, a dica fonológica não seja mais relevante para a flexão. Na seqüência, já que grafamos a vogal temática de “crueldad**e**” com “e” (uma terminação convenientemente átona até segunda ordem, assim como a vogal temática), incentivaremos sua grafia consistente nos vocábulos em que aparece. Daí “pel**e**”, “dent**e**”, “verm**e**” e assim por diante.

---

[79] Só como curiosidade: pode-se até argumentar que “obliquar” seja um “não-vocábulo oficial”, já que devia (?) ser grafado “oblicuar”. O caso de“argüir”, por sua vez, não tem solução semelhante, já que, segundo nossas normas ortográficas, não há outra forma de grafar esse vocábulo.

Assim se explicam muitas das decisões grafêmicas no nível do vocábulo, enquanto, no nível do morfema, o caso é bem outro. Muitas não têm motivação fonológica, como a decisão por "l" em "cruel"; antes, remontam a questões históricas ou etimológicas, o que, na sincronia da ortografia e para o usuário comum, é dizer que são simplesmente *arbitrárias*. Essa arbitrariedade acaba por dar lugar a um jogo de oposições (grafêmicas) que só se estabelecem na língua escrita. Tal é o que ocorre, por exemplo, entre os vocábulos "seção", "sessão" e "cessão" e entre "balde" e "~~baude~~". Em conseqüência, qualquer usuário tenderá a fazer algum tipo de "confusão grafêmica" ao grafar um determinado morfema pela primeira vez.

Na memorização do arbitrário, a repetição é a melhor aliada; logo, quanto mais produtivo/freqüente um morfema, tanto menos provável errar sua grafia. As desinências e morfemas derivacionais menos esdrúxulos têm, portanto, grafias *extremamente* estáveis para o usuário que já ancorou a ortografia na morfologia. A ponto de chegar a erros como "(ele) ~~possue~~" (cf. "[ele] part**e**/vend**e**/morr**e**/beb**e**/sent**e**/etc."), um caso de **hiper-regularização**, ou de "regra sem cabimento"[80]. Esse mesmo usuário dará dicas valiosas em seus não-vocábulos, como as presentes em "amação" e "amassão", ambos dificilmente lexicalizados, mas facilmente interpretáveis como derivados de "amar" e "amasso", respectivamente.

Por fim, vale notar que nem tudo são rosas e que inconsistências existem aos montes, entre as quais figura como expoente máximo, em nossa opinião, a presente no par "extensão" e "estender".

Âncora fonética

Apesar de a termos, de certa forma, menosprezado, nosso sistema ortográfico tem uma **vocação fonética** indiscutível, a saber: dados os possíveis valores dos grafemas e um conjunto de regras de leitura de seqüências grafêmicas (parte significativa das quais já enunciamos), é estritamente requerido que, para toda grafia, haja uma leitura correspondente a uma variante de pronúncia do vocábulo grafado, *mesmo que seja à revelia da forma básica ou do étimo*. Prova disso é a grafia "via**j**ar" do verbo derivado de "via**g**em".

---

[80] A hiper-regularização se refere ao ato ou tendência mais ou menos universal de ignorar regras de exceção a regras mais gerais, em especial se aquelas não parecem justificadas.

Apesar de nunca ser içada por completo, a âncora fonética, muitas vezes, fica meio fora de lugar. Pontos de erro comuns incidem nas seguintes áreas:

- **grafemização**, ou seja, a identificação e interpretação de grafemas. Nesse caso, os erros mais comuns parecem ser quanto ao uso do trema e confusão no par "ão/am". Especificamente para usuários em processo de alfabetização, que muitas vezes precisam sussurrar as palavras para conseguir grafá-las, Cagliari (92) notou confusão grafêmica devida à desativação das cordas vocais (por exemplo, "p" e "b", nessa situação, passam a ter praticamente o mesmo valor).

  Este pesquisador relata, ainda quanto a erros na grafemização, um outro fenômeno, muito interessante e de ampla aplicação, chamado **hipercorreção**. Consiste em o usuário, ao tomar conhecimento da correção para um dado erro seu, rever suas hipóteses de forma a (i) efetivamente eliminar uma classe de erros, mas (ii) incorrer em novos erros. Por exemplo, um usuário que grafe "carro" como "~~caru~~" e seja corrigido poderá, num ataque de hipercorreção, grafar "rua" como "~~rroa~~";

- **regras de codificação da tonicidade**: erros de acentuação gráfica são extremamente comuns, mas não são todos da mesma natureza. Identificamos três classes notáveis e disjuntas, a saber: (i) classe *de menos*, que parece ser a mais freqüente e se caracteriza pela ausência de acento gráfico quando este está prescrito; (ii) classe *demais*, caracterizada pela presença de acento gráfico quando este é redundante; e (iii) classe *melância*, caracterizada pela presença de acento gráfico em vogal átona, ou seja, pela completa ignorância do significado do acento. Supor acento gráfico *demais* tem uma vantagem sobre supor qualquer uma das outras classes: o sistema não fica a conjecturar qual, afinal de contas, deve ser a sílaba tônica da palavra em questão;

- **regra de divisão silábica:** "z" (coda) ou "nh" como sucessores não parecem convencer como justificativa para tornar silábico um "i" ou "u" sob suspeita, gerando "~~raíz~~" e "~~campaínha~~". Vale lembrar que, então, hipercorreção pode levar a "~~raizes~~".

Capítulo V Alguns Erros naturais

Depois de tanto discutir sobre nosso sistema ortográfico, pode ter ficado a impressão de que ele seja o origem de todos os males. Na verdade, ainda há toda uma horda de erros que prescindem da própria existência da escrita, sendo cometidos, em maior ou menor grau, por todos os falantes, inclusive analfabetos. Chamamo-los "naturais" porque surgem naturalmente da convivência do falante com o sistema lingüístico. Não obstante, esses erros, cujos efeitos já se fazem sentir na fala, propagar-se-ão para a grafia, sempre que possível, por meio da âncora fonética de nossa ortografia.

Neste capítulo, vamos analisar duas grandes classes de erro costumeiramente de alto desafio, a saber: a deturpação fonológica, que ocorre nos domínios fonologia, e o erro de classificação, nos da morfologia.

V.1 Deturpação fonológica

Chamamos **deturpação fonológica** a qualquer malformação na forma básica dos vocábulos/morfemas. *Quando evidente*, corresponde a uma pronúncia impossível ou estigmatizada para o vocábulo pretendido. Por exemplo, uma simples — e mais que normal — pronúncia sem ditongo do verbo "roubar" — [x**o̲**ˈbax] — pode esconder a deturpação que se *evidencia* em todas as formas rizotônicas do verbo: [ˈxɔ̲.bu] ("~~róbu~~"), [ˈxɔ̲.bas], [ˈxɔ̲.ba], [ˈxɔ̲.bãw̃]. Trata-se, portanto, de um caso de **deturpação (foneticamente) neutralizável**, ou seja, que ora se evidencia, ora fica latente, em função do ambiente (fonético) de aplicação do morfema afetado. O mesmo não ocorre, por exemplo, com as deturpações fonológicas presentes em "~~estrupo~~", "~~mortandela~~" e "~~salchicha~~". Qualquer que seja o ambiente de aplicação dos morfemas deturpados em questão, sua malformação ficará patente, o que se pode apreciar em "~~estrupar~~", "~~mortandelinha~~" e "~~salchichão~~". Trata-se de casos de **deturpação (foneticamente) estável**.

Postulamos essa dicotomia (neutralizável *vs*. estável) por acreditarmos que os dois tipos de

deturpação sejam fundamentalmente distintos, em pelo menos três aspectos: (i) motivação (lingüística *vs.* cognitiva), (ii) efeitos sobre a produção de não-vocábulos e (iii) forma de correção. Porquanto nossas hipóteses sobre a deturpação fonológica estável sejam ainda prototípicas, de um lado, mas complexas, do outro, preferimos nos abster de apresentá-las.

Deturpação fonológica neutralizável

A **motivação da deturpação neutralizável** é definitivamente lingüística. Formas básicas seguras para os morfemas são adquiridas/inferidas a partir de uma amostragem representativa de como estes se realizam em diferentes contextos fonéticos. Os falantes, entretanto, não são fonólogos e estão longe da sistemática paciência científica; logo, contentam-se com amostras não tão representativas. Tal é a origem da deturpação neutralizável: insuficiência de dados (ou uma certa preguiça).

Para entendermos melhor o problema, tentemos inferir a forma básica de "roubar", cuja grafia fingiremos não conhecer. Basta tomarmos a amostra *{*[ˈxow.bu]*}* para inferirmos a forma básica (simplificada) /xowˈbax/, que gera trivialmente [xoˈbax] ou [xoˈba] ("~~robá~~") por meio de processos fonológicos pós-lexicais consensuais. Perfeito, mas só porque fizemos uma amostragem excelente. Tomemos agora uma amostra menos feliz, *{*[xoˈbax]*}*. Nesse caso, ficamos em situação difícil quanto à forma subjacente do fone [o], já que este se encontra em sílaba átona, ou seja, posição de neutralização absoluta entre os fonemas /**o**/ e /ɔ/ (cf. "cólera" *vs.* "colérico"). Uma agravante ainda é a possibilidade de este fone ter sido derivado pós-lexicalmente de um /ow/ subjacente. Logo, estamos em dúvida entre três candidatos à vaga. Se insistirmos em nossa amostra insuficiente (o que acontece com freqüência, devido à maior freqüência de certos padrões fonêmicos), teremos simplesmente que adivinhar; e muito provavelmente arriscaremos um /ɔ/ básico.

Revisemos agora a origem da deturpação neutralizável: insuficiência de dados ante (i) neutralização absoluta, (ii) processos fonológicos pós-lexicais e (iii) variabilidade de condicionamento fonético (deslocamento acentual), a qual desfaz ambientes de neutralização. Remonta, no fundo, a um **erro de inferência de forma básica**, que, em última análise, explica não-vocábulos como "(eu) ~~pen~~[ɛ]~~ro~~/~~cav~~[ɔ]~~co~~/~~saldo~~" ("peneiro/cavuco/saúdo") e "~~degrais~~"

(“degraus”).

Na Seção V.3 (página 93), fazemos breves considerações acerca da reversão desse tipo de erro, usando o processamento morfológico descrito na próxima seção.

V.2 Erros de classificação — Um estudo de caso em morfologia

Toda uma gama de não-vocábulos resulta obviamente de uma operação equivocada no nível da morfologia, tais como “~~cidadões~~” (“cidadãos”), “~~reaveu~~” (“reouve”), “~~transporam~~” (“transpuseram”), “~~di~~” (“dei”), “~~constrangiu~~” (“constrangeu”), “~~diminói~~” (“diminui”) e “~~vareia~~” (“varia”), quando da flexão, e “~~planejação~~” (“planejamento”), “~~incortês~~” (“descortês”) e “~~pré-câmara~~” (“antecâmara”), quando da derivação. Apesar de se tratar de erros relativos a operações diversas (ora flexão, ora derivação; flexão ora verbal, ora nominal; derivação ora por prefixação, ora por sufixação; etc.), um resultado interessante é que todos podem ser devidamente cobertos, de maneira uniforme e elegante, ao serem entendidos não propriamente como erros morfológicos (todos os exemplos acima “fazem morfologicamente sentido”), mas como *erros de classificação* por parte do usuário.

Esse salto de abstração nos parece muito relevante e valioso, visto que deixamos de lidar com questões puramente lingüísticas para também considerar outras de interesse para muitos domínios. O objetivo desta seção é exatamente relatar essa experiência, primeiro (re)construindo um modelo teórico da classificação *como uma operação passível de erros reversíveis* e, em seguida, demonstrando como esse modelo é instanciado na reversão dos erros ortográficos em questão, como um estudo de caso.

Em oportunidade única neste trabalho, apresentamos considerações não só acerca da gramática de reconstituição, mas também sobre aspectos de estimativa de utilidade.

Classificação: uma operação potencialmente confusa

> **Termo de (ir)responsabilidade:** a discussão a seguir não tem absolutamente a pretensão de formalizar, de forma definitiva ou completa, a problemática da classificação; antes, finge-se de formalização para tratar dos conceitos envolvidos de forma mais clara e, por que não, apresentar um esboço razoável de um tal formalismo.

Um primeiro engano que se pode cometer na tentativa de modelar a operação de classificação é reduzi-la ao mero estabelecimento do valor-verdade do predicado *instância/2* a seguir:

$$instância(O, C) \leftrightarrow \text{objeto } O \text{ é instância da classe } C$$

ou ainda defini-la como algo do tipo:

$$classes(O) = \{c \in \mathbf{U} | \ instância(O, c)\}$$

A primeira acepção ingenuamente destitui a classificação de seu caráter de "escolha dentro de um conjunto de possíveis classes". A função *classes(x)*, por sua vez, respeita esse caráter, mas peca por não restringir o universo de opções, ou seja, por não contextualizar a escolha. Vejamos uma definição mais adequada:

Def. 1: Dados um objeto *O* e um conjunto de classes *Contexto* quaisquer, a (operação de) **classificação** de *O* em *Contexto*, denotada por ***classes(O, Contexto)***, é o conjunto

$$classes(O, Contexto) = \{c \in Contexto | \ instância(O, c)\}$$

Ou seja, o conjunto de todas as classes em *Contexto* que têm *O* por instância. Note que a classificação pode ser vazia e que vale a seguinte equivalência lógica:

$$instância(O, C) \equiv [classes(O, \{C\}) = \{C\}]$$

Qualquer sombra de preciosismo desaparece ao se compararem as operações a seguir quanto ao potencial de confusão envolvido:

classes(escorpião, {número, letra, cor})
classes(escorpião, {aracnídeo, mamífero})
classes(escorpião, {inseto, aracnídeo, crustáceo})

Além disso, os seguintes erros:

✸*classes(escorpião, {carro, bicicleta}) = {bicicleta}*
✸*classes(escorpião, {inseto, aracnídeo, crustáceo}) = {inseto}*

variam muito quanto à verossimilhança.

Estabelecido o papel do contexto numa classificação, é natural que passemos a considerar protótipos formais para *verossimilhança de um erro de classificação* e *grau de confusão*. Pode parecer, a princípio, que a verossimilhança está diretamente relacionada com o grau de confusão da operação de classificação propriamente dita. Essa idéia é desmascarada em confrontos do tipo:

- ✸*classes(escorpião, {inseto, aracnídeo, crustáceo, bicicleta}) = {bicicleta}*
- ✸*classes(escorpião, {inseto, aracnídeo, crustáceo, bicicleta}) = {crustáceo}*

O grande contraste entre os exemplos acima é que um escorpião é em muitos aspectos *similar* a um crustáceo, enquanto praticamente não pode ser sequer *comparado* com uma bicicleta. Naturalmente entra em questão uma ferramenta cognitiva básica subjacente à classificação: a comparação entre objetos, envolvendo tanto a fatoração de propriedades comuns quanto a identificação de diferenças. Em nossa modelagem, o *front-end* dessa ferramenta é alguma função ***confusão***: $P(U_{classes}) \rightarrow [0, 1]$, tal que

confusão(x) = grau de confusão/similaridade/uniformidade entre as classes pertencentes a x.

Dispondo-se de uma boa *confusão*, um fator de influência na verossimilhança de um erro de classificação ✸*classes(Classificando, Contexto) = Classes* é o seguinte:

$$confusão(Classes \cup \{<Classificando>^{81}\}).$$

Falta apenas contextualizar esse fator. Para tanto, optamos por contrastá-lo com o somatório dos fatores para todas as possíveis respostas (uma alternativa seria considerar apenas o fator máximo), exceto a resposta vazia, que é tratada à parte. Isso é implementado na definição a seguir.

---

[81] A operação <*X*> denota a conversão da entidade *X* em classe, ou seja, <*X*> é uma classe calculada que tenha *X* como instância e *seja o mais específica possível* Essa operação poderá não ser trivial em função de como o conceito de classe estiver sendo "implementado".

$$verossimil\ hança(Cndo, Cntxt, Rslt) = \frac{fator(Cndo, Cntxt, Rslt)}{\sum_{qq \in P(Cntxt)} fator(Cndo, Cntxt, qq)}$$

Def. 2: onde:

$$fator(Cndo, Cntxt, Rslt) = \begin{cases} 1 - \underset{Cs \in P(Cntxt)}{Máx} \{confusão(Cs \cup \{<Cndo>\})\}, & Rslt = \varnothing \\ confusão(Cs \cup \{<Cndo>\}), & \forall Rslt \neq \varnothing. \end{cases}$$

Basta agora definir uma boa função *confusão*. Como qualquer tal função analisa um conjunto de classes, é necessário agora que decidamos por algum tipo de "implementação" para o conceito de *classe*. Consideremos classes como conjuntos de predicados, valendo as seguintes definições:

Def. 3: $instância(O, C) \leftrightarrow \forall p\ [p \notin C \vee p(O)]$

ou seja, um objeto O será considerado instância de uma classe C sse todo predicado especificado em C for válido para O.

Sejam *A* e *B* classes:

Def. 4: $A \dot{-} B = \{p \in A | \sim\exists q\ (q \in B \wedge q \equiv p)\}$.

Def. 5: $A\ .\Lambda.\ B = \{p \in A | \exists q\ (q \in B \wedge q \equiv p)\}$.

Def. 6: $A\ .\mathrm{U}.\ B = (A \dot{-} B) \cup (B \dot{-} A) \cup\ (A\ .\Lambda.\ B)$.

As três operações entre classes definidas acima são respectivamente análogas às operações de diferença, interseção e união entre conjuntos, devidamente adaptadas para levar em conta a equivalência entre predicados. A definição da operação *.U.*, menos direta, visa apenas evitar o surgimento de predicados equivalentes na classe resultante, o que equivaleria a permitir elementos duplicados. Vale notar que o resultado dessas operações, por sua vez, sempre será uma classe.

Segue um primeiro protótipo razoável de confusão:

$$confusão(\{c_i\}) = \frac{\#\left(.\Lambda._{i=1}^{n} c_i\right)}{\#\left(.\mathrm{U}._{i=1}^{n} c_i\right)}$$

A versão acima, apesar de ingênua, é essencialmente perfeita, crescendo conforme o número de propriedades em comum a todas as classes aumenta, mas sem perder a noção de proporção com o número total de propriedades envolvido. Entretanto, não é realista ao dar um mesmo peso a todas as propriedades. É fato que certas propriedades de uma classe são "sentidas" como mais características do que outras. Acreditamos, por exemplo, que "produzir leite" seja normalmente considerado bem mais característico da classe dos mamíferos do que "ter sangue quente", apesar de, a rigor, a presença de ambas as propriedades ser condição necessária para classificar um animal como mamífero. Para lidar com esse problema, podemos assumir a existência de uma função

$$e: (U_{predicados} \times P(U_{classes})) \rightarrow [0, 1]$$

que calcula um **grau de pertinência** de propriedades (predicados) a contextos (conjuntos de classes). Uma sugestão para realizar o cálculo dessa função para contextos não-unitários seria:

$$e(p, \{c_1, c_2, \ldots, c_n\}) = \sum_{i=1}^{n} \frac{e(p, \{c_i\})}{n}$$

que não passa de uma simples média aritmética e reduz o problema ao cálculo para contextos unitários. Este último, no entanto, não é absolutamente trivial, variando com o domínio de aplicação e constituindo, em última análise, um parâmetro pessoal do usuário[82].

Agora podemos lançar uma versão razoável de confusão, primeiro apenas definindo, como mero apoio notacional, a função *eTotal*, que realiza o somatório dos graus de pertinência observados em um conjunto de predicados (classe).

$$eTotal(Classe, Contexto) = \sum_{p \in Classe} e(p, Contexto)$$

[82] Suspeitamos, não obstante, que se possam derivar bons estimadores para esse parâmetro por meio da análise estatística e estrutural da hierarquia de classes envolvida. Dados sobre erros comuns são também úteis.

$$confusão(Cs) = \frac{eTotal(\bigwedge_{c \in Cs} c, Cs)}{eTotal(\bigcup_{c \in Cs} c, Cs)}$$

Pode-se observar que o que diferencia a **nova versão de confusão** é basicamente o fato de que, agora, as propriedades são contadas com pesos próprios, dados pela função de grau de pertinência.

Aplicação à morfologia

A idéia de classificação não é de forma alguma estranha aos processos morfológicos de uma língua. Basta lembrar que os termos **paradigma** e *modelo* são de uso corrente na literatura sobre morfologia (Monteiro, 86), não sendo senão sinônimos de *classe*. Entre os exemplos comuns de classes "morfológicas" em português, contam-se "adjetivos que fazem o superlativo em *-íssimo/-érrimo*", "verbos que se conjugam como *cantar/vender/partir/pôr/passear/odiar/construir/etc.*", "temas verbais que fazem substantivos abstratos em *-ção/-mento*" e assim por diante. Naturalmente, erros de classificação são esperados dentro de cada um dos três blocos de classes (contextos) apresentados. Não-vocábulos como "~~conjugamento~~", "~~vareia~~" e "~~diminói~~" resultam respectivamente de erros de classificação como:

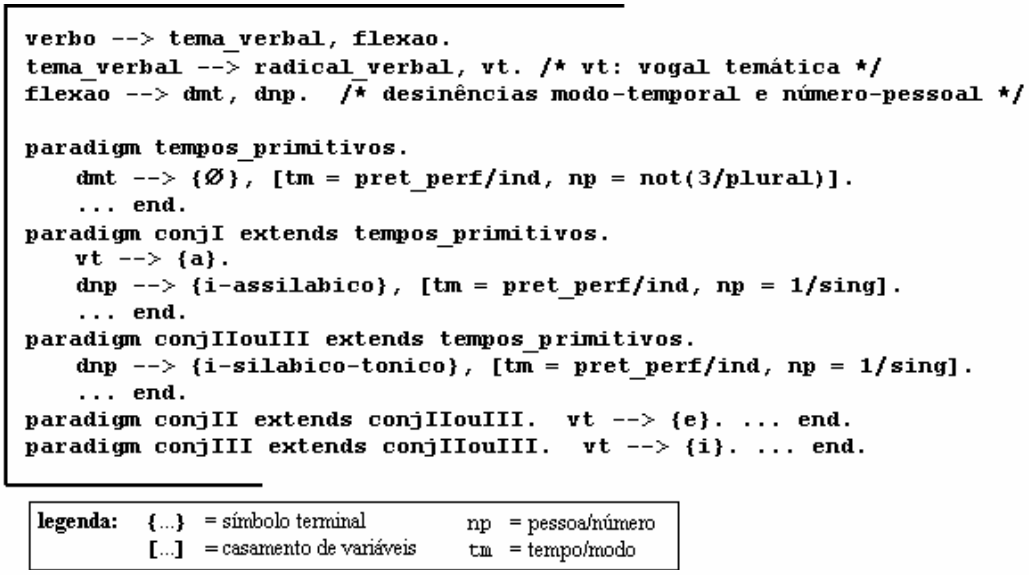
- *classes(conjugar, {<-ção>, <-mento>}) = {<-mento>}*
- *classes(variar, {<cantar>, <odiar>}) = {<odiar>}*
- *classes(diminuir, {<partir>, <construir> }) = {<construir>}*

Para levantar hipóteses de formação de palavras contendo esse tipo de erro e, em seguida, proceder a uma reversão adequada, optamos por representar o conhecimento necessário por meio de uma **gramática de palavras**[83] (Agirre et al., 92; Sengupta & Chaudhuri, 96) baseada em unificação (Shieber, 86), segundo um modelo inspirado nas *GLFs*[84], mas bastante simplificado e estendido, inclusive de forma a incorporar o conceito de paradigmas. A referida simplificação consiste na existência de um nível único de unificação, ou seja, não existem variáveis locais. A

[83]Do inglês *"word grammar"*.

[84]Gramáticas Léxico-Funcionais.

Figura 4 apresenta uma amostra de código no formalismo utilizado para que se tenha uma idéia de como essas características são realizadas.



```
verbo --> tema_verbal, flexao.
tema_verbal --> radical_verbal, vt. /* vt: vogal temática */
flexao --> dmt, dnp.  /* desinências modo-temporal e número-pessoal */

paradigm tempos_primitivos.
    dmt --> {Ø}, [tm = pret_perf/ind, np = not(3/plural)].
    ... end.
paradigm conjI extends tempos_primitivos.
    vt --> {a}.
    dnp --> {i-assilabico}, [tm = pret_perf/ind, np = 1/sing].
    ... end.
paradigm conjIIouIII extends tempos_primitivos.
    dnp --> {i-silabico-tonico}, [tm = pret_perf/ind, np = 1/sing].
    ... end.
paradigm conjII extends conjIIouIII.  vt --> {e}. ... end.
paradigm conjIII extends conjIIouIII.  vt --> {i}. ... end.
```

legenda: {...} = símbolo terminal; [...] = casamento de variáveis; np = pessoa/número; tm = tempo/modo

Figura 4: Amostra do formalismo gramatical utilizado.

A conveniência de um modelo gramatical baseado em unificação, nessa aplicação, é a expressão natural do fenômeno da concordância, que também ocorre no nível da morfologia. Dessa forma, nas regras de produção, as diversas noções gramaticais (gênero, número, grau, tempo, modo e pessoa) vão sendo "montadas" pela unificação de certas variáveis, auxiliando a reversão de erros de flexão. Além disso, algumas noções semânticas associadas aos morfemas derivacionais (prefixos e sufixos lexicais) são anotadas da mesma forma, enfocando erros de derivação, principalmente de bloqueio.

Como se pode observar na Figura 4, a operação de generalização/especialização pode ser expressa com naturalidade (a palavra reservada **`extends`** introduz uma lista de superclasses). Classes, nesse contexto, podem ser entendidas simplesmente como blocos hierarquizados de regras de produção. Em conformidade com essa visão, existe uma importante restrição: subclasses sempre herdam integralmente o comportamento (não só a interface) de suas superclasses, ou seja, não há redefinição (*overriding*), apenas extensão.

A semântica do conceito de classe no nosso formalismo gramatical é a seguinte: se, na geração/*análise* de uma palavra, uma regra de produção definida em uma dada classe *C* é usada,

então passam a ser consideradas inaplicáveis, nos próximos passos, as regras definidas nas demais classes da hierarquia[85] de *C*, com exceção de suas superclasses e subclasses, diretas ou não. Dessa forma, os símbolos usados no lado direito das regras de uma classe qualquer fazem referência a entidades já ou ainda por serem definidas (i) em seus "ancestrais" e "descendentes", (ii) globalmente ou (iii) em classes de outras hierarquias.

De "~~dí~~" para "deí"

Para elucidar como o formalismo acima, em conjunto com as idéias gerais sobre classificação, pode ser usado na reversão de erros na formação de palavras, rastrearemos um único exemplo (interessante) de erro de conjugação verbal, uma vez que o processamento nos demais casos é análogo (e geralmente mais simples). O não-vocábulo em questão é "~~di~~" ("[eu] dei", flexão do verbo "dar"), pouco comum na forma escrita, mas muito reveladora.

O processo de reversão é disparado no momento em que se constata que ✷~~di~~ não é uma cadeia pertencente ao léxico[86]. Como conseqüência, uma série de hipóteses de reversão deve ser considerada, inclusive supondo erros em outros níveis que o morfológico, irrelevantes na presente discussão (por exemplo, a sugestão de correção "de" — preposição — suporia um erro não-morfológico, nesse caso). Dispondo-se do conhecimento apresentado de forma simplificada na Figura 4, bem como algumas regras de adaptação morfofonêmica e um *parser bottom-up* adequado, são levantadas duas hipóteses de formação para ✷~~di~~, apresentadas de forma simplificada a seguir:

$$\boldsymbol{H_{conjII}} \quad \textbf{:}\ d_{radical\_verbal} + \textbf{<}\boldsymbol{\text{✷}conjII}\textbf{>} + e_{vt} + \varnothing_{dmt} + \text{i-silábico-tônico}_{dnp}$$

$$\boldsymbol{H_{conjIII}} \quad \textbf{:}\ d_{radical\_verbal} + \textbf{<}\boldsymbol{\text{✷}conjIII}\textbf{>} + i_{vt} + \varnothing_{dmt} + \text{i-silábico-tônico}_{dnp}$$

O *parser* pára exatamente nesse ponto, não continuando a montagem da árvore de derivação porque está instruído a parar em regras que contenham um ou mais **pontos de decisão de classe**,

---

[85] A *hierarquia* de *C* é o conjunto de todas as classes que guardam algum "parentesco" com *C*, ou seja, todas aquelas que têm alguma superclasse em comum com *C*.

[86] Assumimos, nesta discussão, um ferramental à altura da biblioteca KLS.

marcados com <✸> nas hipóteses acima[87]. Ambas supõem que a decisão de classe foi errada; mas a *intenção* original do autor, não, a qual está anotada nas variáveis *np* e *tm*, unificadas na análise (parcial) da cadeia. Em ambos os casos, essas variáveis informam que a intenção original do autor seria a flexão da 3ª pessoa do singular do pretérito perfeito do indicativo de um suposto verbo de radical "d".

Nesse ponto, encontramos uma situação singular, distinta das analisadas na apresentação do nosso modelo de classificação, visto que agora nada se pode dizer acerca do suposto verbo — o *classificando*. Ou seja, torna-se impossível calcular *a priori* a verossimilhança do erro de classificação considerado. O *framework*, entretanto, não fica invalidado: o sistema corretor supõe uma alta verossimilhança que será verificada *a posteriori*. Essa suposição fornece um resultado valioso: qualquer que seja o verbo correto que o usuário devesse ter usado, *este terá propriedades de alto peso em comum com o suposto verbo conjugado na(s) classe(s) errada(s)*.

As propriedades em questão são um conjunto fixo e restrito de flexões mais características ou mais freqüentemente usadas. Poderíamos ter a definição para o grau de pertinência *"e"* apresentada na Tabela IV, onde *Cx* é o contexto da operação de classificação, que fica disponível de forma trivial e consiste na hierarquia de *conjII*/*conjIII*, ou seja, *temposPrimitivos* e todas as suas subclasses.

[87] Um *ponto de decisão de classe* é um momento, na *geração* de uma palavra, em que o gerador/autor tem que se decidir por uma classe para continuar a geração. No exemplo analisado e supondo uma ordem de geração da esquerda para direita (visto tratar-se de sufixação), o *único* tal ponto ocorre logo antes da expansão do símbolo não-terminal *vt*.

Tabela IV: Definição de e(x, Cx) e flexões correlatas segundo cada hipótese.

x ***(configuração de variáveis)***	***e(x, Cx)***	**flexionando** ***d+<✷ConjII>***	**flexionando** ***d+<✷ConjIII>***
[tm = inf_impessoal, np = 0]	1	✷der	✷dir
[tm = pres/ind, np = 1/sing]	1	✷do	✷do
[tm = pret_perf/ind, np = 1/sing]	1	✷di	✷di
[tm = pret_perf/ind, np = 3/sing]	1	✓**deu**	✷diu
[tm = participio, gn = masc/sing]	1	✷dido	✷dido
outros	0	–	–

Consideremos primeiramente H_{conjII}. Na avaliação dessa hipótese, o corretor verifica se alguma propriedade pertinente da classe *ConjII* também vale para algum verbo lexicalizado. Isso envolve, para cada configuração de variáveis pertinente *X*, (i) retomar a geração a partir de onde o *parsing* parou, assumindo *X* como parte indispensável da configuração final, (ii) consultar a cadeia resultante (terceira coluna da Tabela IV) no léxico e, em caso de sucesso, (iii) verificar se as propriedades gramaticais dessa cadeia, no léxico, correspondem às representadas pela configuração *X* e pelos símbolos não-terminais da própria gramática de palavras (no exemplo, não se pode esquecer que estamos tratando do desenvolvimento de **`verbo`**s).

Na avaliação de H_{conjII}, como se pode observar na Tabela IV, esse procedimento obteve uma única propriedade de ligação — a flexão "deu" — entre ✷der e um verbo correto qualquer. A alta pertinência dessa propriedade basta para validar a suposição de verossimilhança do erro feita anteriormente. Por fim, o último passo da correção consiste em pedir ao léxico que flexione "deu" segundo a intenção original do autor, gerando "dei", uma boa sugestão de correção.

Dessa forma, o que poderia talvez ser considerado uma hipótese absurda, a saber, conjugar-se "dar" como um verbo da 2ª conjugação, revela-se realista. De fato, o erro em "~~di~~", bem como a maior parte dos erros morfológicos, resulta de um ato de inteligência: a analogia "*[ele] vendeu* está para *[ele] deu* assim como *[eu] vendi* está para *[eu] di*" é perfeita e revela que a semelhança

entre infinitivos é apenas uma dentre as muitas causas de confusão na conjugação verbal.

A avaliação de $H_{conjIII}$ segue o mesmo procedimento que H_{conjII}, sem gerar, no entanto, cadeias lexicalizadas, o que invalida *a posteriori* a suposição da verossimilhança do erro de $H_{conjIII}$. Vale observar, ainda, que o erro foi corrigido sem a necessidade da montagem de uma árvore de derivação completa, um dos motivos por que optamos por análise *bottom-up*.

Alguns exemplos interessantes que podem ser rastreados da mesma forma são "~~constrangiu~~" ([ele] constrangeu) e "~~reaveu~~" ([ele] reouve), casos em que as propriedades de ligação seriam respectivamente "constrangido" e "reaver". Erros como o presente em "~~vareia~~" ([ele] varia) podem ser revertidos da mesma forma, apenas exigindo a extensão da gramática com a adição de outros paradigmas (o de "odiar", no caso).

V.3 De "~~saldo~~" para "saúdo"

O que não enfatizamos na seção precedente é que nosso *parser* morfológico hipotético opera entre os níveis fonético e fonêmico dos (não-)vocábulos, em vez de sobre sua grafia, como é costume. Isso o torna especialmente útil na reversão de erros de inferência da forma básica.

Como vimos, o *parser* não só levanta hipóteses de análise morfológica para não-vocábulos de entrada ("~~di~~" = flexão de "~~der~~"), como também reconstitui formas fortemente correlacionadas com as hipóteses levantadas ("~~di~~" → "~~der~~" → "deu" → "dei"). Nesse processo, formas fonêmicas são construídas; mudanças de contexto fonético (deslocamento acentual), operadas; e, portanto, situações de neutralização absoluta e variação livre, explicitadas ou ampliadas, podendo servir de gatilho para hipóteses de erro de inferência da forma básica.

Tomemos o exemplo de "~~saldo~~". Seu processamento morfológico, de acordo com o procedimento descrito anteriormente, acabará por relacioná-lo com "~~saldar~~". Nesse ponto, como vimos, o reversor de erros de classificação supõe que "~~saldar~~" seria elo de "~~saldo~~" com seu verdadeiro paradigma. Por que não supor também que a forma básica de "~~saldar~~" constituiria o resultado de um erro de inferência da forma básica, já que a neutralização nesse novo padrão acentual abrange ainda um possível hiato? Dessa hipótese, chegaríamos com facilidade a "saudar" e daí a "saúdo", por meio do léxico e das pretensões flexionais supostas para "~~saldo~~" no início do processo.

Outro caso que se resolve de maneira semelhante é o de "~~degrais~~", que o *parser* logo suporá

como flexão de “~~degral~~”. Eis mais uma oportunidade de supor um erro de inferência da forma básica, gerar “degrau” e daí chegar a “degraus” via léxico.

Capítulo VI Conclusões e Trabalhos futuros

Termina aqui e assim nosso desabafo. Esperamos ter demonstrado definitivamente que ainda há muito a ser desenvolvido no aconselhamento ortográfico, como um todo e especialmente para o português do Brasil. Que há ainda espaço para soluções engenhosas nos mais variados níveis: especificação de requisitos (utilidade), projeto (arquitetura genérica) e implementação (reversão de erros de classificação e de inferência da forma básica). Que a lingüística é realmente uma ferramenta poderosa na resolução do problema. E que ainda há muito a ser estudado e explorado nesse sentido.

Quanto a trabalhos futuros, é bela, aguda e lisonjeira a visão de toda essa obra como uma imensa e tentadora seção de muitos trabalhos futuros. No entanto, para não quebrar o protocolo, listamos as seguintes frentes:

- complementar nosso modelo informal/semiformal de gramática de reconstituição, em especial quanto a muitos tópicos de morfologia (processos de formação de palavras, por exemplo);
- desenvolver um modelo formal unificado de gramática de reconstituição para a reversão de erros ortográficos;
- desenvolver um modelo razoável de reversão e previsão da deturpação fonológica estável;
- investigar a inferência semi-automática de pontos de erro, incluindo a formulação e aplicação de meta-hipóteses de hipercorreção e hiper-regularização;
- implementar um conselheiro ortográfico segundo nossa arquitetura;
- desenvolver um projeto sério de avaliação de conselheiros ortográficos centrados em utilidade, estritamente necessário para comprovar as hipóteses levantadas neste trabalho.

Referências bibliográficas

(Agirre et al., 92) Agirre, E., Alegria, I., Arregi, X., Artola, X., Díaz De Ilarraza, A., Maritxalar, M., Sarasola, K., Urkia, M. XUXEN: A Spelling Checker/Corrector for Basque Based on Two-Level Morphology. *In 3rd Conf. of Applied NLP*, 1992, 119-125.

(Almeida & Pinto, 95) Almeida, J. J., Pinto, U. Jspell — um módulo para análise léxica genérica de linguagem natural. *In Actas do Congresso da Associação Portuguesa de Lingüística*, Évora, 1995.

(Angel et al., 83) Angel, R. C., Freund, G. E., Willet, P. Automatic spelling correction using a trigram similarity measure. *Information Processing & Management, 19* (1983), 255-261.

(Aurélio, 96) *Dicionário Aurélio Eletrônico.* Versão 2.0. Copyright © 1996 Editora Nova Fronteira.

(Basílio, 80) Basílio, M. *Estruturas Lexicais do Português: Uma Abordagem Gerativa.* Editora Vozes Ltda, 1980.

(Bechara, 92) Bechara, E. *Moderna Gramática Portuguesa.* Companhia Editora Nacional, 34ª edição, 1992.

(Boivie, 81) Boivie, R. H. *Directory assistance revisited.* Memorando, AT & T Bell Labs Technology, 12 de junho de 1981.

(Braga et al., 00) Braga, P. A., Carvalho, A. C. P. L. F., Ludermir, T. B. *Redes Neurais Artificiais: Teoria e Aplicações.* Editora LTC (Livros Técnicos e Científicos), 2000.

(Burr, 87) Burr, D. J. Experiments with a conectionist test reader. *In IEEE International Conference on Neural Networks*. IEEE, Nova Iorque, IV: 717-724, 1987.

(Cagliari, 92) Cagliari, L. C. *Alfabetização & Lingüística.* Editora Scipione, 5a edição, 1992.

(Cagliari, 97a) Cagliari, L. C. *Análise Fonológica*. Edição do Autor, 1997.

(Cagliari, 97b) Cagliari, L. C. *Fonologia do Português — Análise pela Geometria de Traços.* Edição do Autor, 1997.

(Câmara Jr., 70) Câmara Jr., J. M. *Estrutura da Língua Portuguesa*. Editora Vozes, 21ª edição, 1970.

(Câmara Jr., 77) Câmara Jr., J. M. *Dicionário de Lingüística e Gramática.* Editora Vozes, 16ª edição, 1977.

(Cherkassky & Vassilas, 89a) Cherkassky, V., Vassilas, N. Backpropagation networks for spelling correction. *Neural Networks, 1, 3* (julho/1989), 166-173.

(Cherkassky & Vassilas, 89b) Cherkassky, V., Vassilas, N. Performance of backpropagation networks for associative database retrieval. *Int. J. Comput. Neural Net*, 1989.

(Church & Gale, 91) Church, K. W., Gale, W. A. Enhanced Good-Turing and cat-cal: Two new methods for estimating probabilities of English bigrams. *Comput. Speech Lang*, 1991.

(Contant & Brunelle, 92) Contant, C., Brunelle, E. Exploratexte: Un analyseur a l'affut des erreurs grammaticales. *In Actes du colloque lexiques-grammaires compares*, Universite du Quebec a Montreal, 1992.

(Corbin, 94) Corbin, D. Méthodes em morphologie dérivationelle. *Cahiers de Léxicologie*, 44, Besançon, 1984.

(Daelemans et al., 84) Daelemans, W., Bakker, D., Schotel, H. Automatische detectie en correctie van spelfouten. *Informatie, 26* (1984), 949-1024.

(Damerau, 64) Damerau, F. J. A Technique for computer detection and correction of spelling errors. *Communications of the ACM, 7, 3* (mar./1964), 171-176.

(de Almeida, 92) de Almeida, N. M. *Gramática Metódica da Língua Portuguesa.* Editora Saraiva, 38ª edição, 1992.

(de André, 94) de André, H. A. *Gramática Ilustrada.* Editora Moderna, 4ª edição, 1994.

(De Heer, 82) De Heer, T. The application of the concept of homeosemy to natural language information retrieval. *Information Processing & Management, 18* (1982), 229-236.

(Deffner et al., 90) Deffner, R., Eder, K., Geiger, H. Word recognition as a first step towards natural language processing with artificial neural nets. *In Proceedings of KONNAI-90*, 1990.

(Demasco & McCoy, 92) Demasco, P. W., McCoy, K. F. generating text from compressed input: An intelligent interface for people with severe motor impairments. *Communications of the ACM, 35, 5* (maio/1992), 68-78.

(Desmarais, 98) Desmarais, L. Learning how to spell using a spell checker. *Canadian Modern Language Review-Revue Canadienne Des Langues Vivantes, 55, 1* (1998), 76-96.

(DTS, 98) *Revisor Gramatical DTS.* Versão 3.0. Copyright ©1998 DTS Software.

(Dunlavey, 81) Dunlavey, M. R. On spelling correction an beyond. *Communications of the ACM, 24, 9* (set/1981), 608.

(Durham et al., 83) Durham, I., Lamb, D. A., Saxe, J. B. Spelling correction in user interfaces. *Communications of the ACM, 26, 10* (out./1983), 764-773.

(Faraco & Moura, 94) Faraco, C. E., Moura, F. M. *Gramática.* 13ª ed., Editora Ática, 1994.

(Foley, 78) Foley, J. Quatre principes de l'analyse morphologique. *Langages.* 85, Larousse, 1978.

(Forney, 73) Forney, G. D. Jr. The Viterbi algorithm. *Proc. IEEE, 61, 3* (1973), 268-278.

(Gentner et al., 83) Gentner, D. R., Grudin, J., Larochelle, S., Norman, D. A., Rumelhart, D. E. Studies of typing from the LNR typing research group. *In Cognitive Aspects of Skilled Typewriting*, W. E. Cooper, Editora Springer-Verlag, Nova Iorque, 1983.

(Goshtasby & Ehrich, 88) Goshtasby, A., Ehrich, R. W. Contextual word recognition using

probabilistic relaxation labeling. *Pattern Recognition, 21, 5* (1988), 455-462.

(Grudin, 83) Grudin, J. Error Patterns in Skilled and Novice Transcription Typewriting. *In Cognitive Aspects of Skilled Typewriting*, W. E. Cooper, Editora Springer-Verlag, Nova Iorque, 1983.

(Gupta, 98) Gupta, R. Can spelling checkers help the novice writer? *British Journal Of Educational Technology. 29, 3* (1998), 255-266.

(Guedes & Guedes, 94) Guedes, A. M., Guedes, R. *Dicionário Prático de Conjugação dos Verbos da Língua Portuguesa*. Bertrand Editora, 1994.

(Hawley, 82) Hawley, M. J. *Interactive spelling correction in Unix: The METRIC Library.* Memorando, AT & T Bell Labs Tech., 31 de agosto de 1982.

(Ho et al., 91) Ho, T. K., Hull, J. J., Srihari, S. N. Word recognition with multi-level contextual knowledge. *In Proceedings of IDCAR-91*, 905-915, 1991.

(Houaiss, 01) *Dicionário Houaiss da língua portuguesa.* Editora Objetiva, 2001.

(Itautec, 99) *Redação Língua Portuguesa.* Versão 7.1. Copyright ©1995-1999 Itautec-Philco.

(Jones et al., 91) Jones, M. A., Story, G. A., Ballard, B. W. Integrating multiple level sources in a Bayesian OCR post-processor. *In Proceedings of IDCAR-91,* 925-933, 1991.

(Michaelis, 98) *Michaelis: moderno dicionário da língua portuguesa.* Companhia Melhoramentos de São Paulo, 1998.

(Monteiro, 86) Monteiro, J. L. *Morfologia Portuguesa*. Editora da Universidade Federal do Ceará (EUFC), 1986.

(Kahan et al., 87) Kahan, S., Pavlidis, T., Baird, H. S. On the recognition of characters of any font size. *IEEE Transactions on Pattern Analysis & Machine Intelligence* (PAMI-9)*, 9* (1987), 274-287.

(Kernigham et al., 90) Kernigham, M. D., Church, K. W., Gale, W. A. A spelling correction program based on a noisy channel model. *In Proceedings of COLING-90, The !3th*

International Conference on Computaional Linguistics, vol. 2 (Helsinki), Editora Hans Karlgren, 205-210, 1990.

(Klavans & Chodorow, 91) Klavans, J., Chodorow, M. Using a Morphological Analyzer to Teach Theoretical Morphology. *Computers And The Humanities, 25, 5* (1991), 281-287.

(Knuth, 73) Knuth, D. E. *The Art of Programming*. Vol. 3: *Sorting and Searching*. Addison-Wesley, 1973.

(Koskenniemi, 83) Koskenniemi, K. *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production.* Publicação n° 11, Departamento de Lingüística Geral, Universidade de Helsinki, 1983.

(Kukich, 88) Kukich, K. variations on a back-progation name recognition net. *In Proceedings of the Advance Technology Conference*, vol. 2, 722-735, 1988.

(Kukich, 90) Kukich. K., A comparison of some novel and traditional lexical distance metrics for spelling correction. *In Proceedings of INCC-90-Paris*, 309-313, 1990.

(Kukich, 92) Kukich, K. Techniques for Automatically Correcting Words in Text. *ACM Computing Surveys, 24, 4* (1992), 377-439.

(Lass, 84) Lass, R. *Phonology - An introduction to basic concepts.* Cambridge University Press, 1984.

(Lee et al., 90) Lee, Y.-H., Evens, M., Micheal, J. A., Rovick, A. *A. Spelling correction for an intelligent tutoring system*. Technical report, Dept. of Computer Science, Illinois Institute of Technology, Chicago, 1990.

(Levenshtein, 66) Levenshtein, V. I. Binary codes capable of correcting deletions, insertions and reversals. *Sov. Phys. Dokl, 10* (1966), 707-710.

(Lexikon, 97) *Gramática Eletrônica.* Versão 1.0. Copyright ©1997 Lexikon Informática.

(Lins et al., 99) Lins, R. D., Carnelo, H. ª L., Moura, R. S. Um SOS para a Língua Portuguesa. *In Actas do IV Encontro para o Processamento Computacional da Língua Portuguesa Escrita*

e Falada (PROPOR'99), 1999, 129-138.

(Lucchesi & Kowaltowski, 93) Lucchesi, C. L., Kowaltowski, T. Applications of Finite Automata Representing Large Vocabularies. *Software — Practice and Experience, 23, 1* (1993), 15-30.

(McClurg & Kasakow, 98) Mcclurg, P., Kasakow, N. Word-Processors, Spelling Checkers, and Drill-and-Practice Programs – Effective Tools for Spelling Instruction. *Journal Of Educational Computing Research, 5, 2* (1989), 187-198.

(Means, 1988) Means, L. G. Cn yur cmputr raed ths. In Proceedings of the 2nd Applied Natural language Processing Conference, 93-100, 1988.

(Microsoft, 97) *Microsoft® Word 97* (Editor de textos que embute um corretor ortográfico). Copyright ©1983-1997 Microsoft Corporation.

(Microsoft, 99) *Microsoft® Word 2000 (9.0.2812)* (Editor de textos que embute um corretor ortográfico) Copyright ©1983-1999 Microsoft Corporation.

(Milne et al., 96) Milne S., Shiu E., Cook J. Development of a Model of User Attributes and Its Implementation within an Adaptive Tutoring System. *User Modeling And User-Adapted Interaction, 6*, *4* (1996), 303-335.

(Mitton, 87) Mitton, R. Spelling checkers, spelling correctors, and the misspelings of poor spellers. *Information Processing & Management, 23, 5* (1987), 495-505.

(Monteiro, 86) Monteiro, J. L. *Morfologia Portuguesa*. Editora da Universidade Federal do Ceará (EUFC), 1986.

(Mor & Fraenkel, 82) Mor, M., Fraenkel, A. S. A hash code method for detecting end correcting spelling errors. *Communications of the ACM, 25, 12* (1982), 935-938.

(Nemhauser, 66) Nemhauser, G. L. *Introduction to Dynamic Programming.* Wiley, Nova Iorque, 1966.

(Odell & Russel, 18) Odell, M. K., Russel, R. C. *U.S. Patent Numbers 1,261,167 (1918) and*

1,435,663 (1922). U.S. Patent Office, Washington, D. C.

(Oshika et al., 88) Oshika, T., Machi, F., Evans, B., Tom, J. Computational techniques for improved name search. *In Proceedings of the 2nd Annual Applied Natural Language Conference*, 203-210, 1988.

(Pacheco, 96) Pacheco, H. C. F. *Uma Ferramenta de Auxílio à Redação.* Dissertação de Mestrado, Departamento de Ciência da Computação, Instituto de Ciências Exatas, UFMG, 1996.

(Pijls et al., 87) Pijls, F., Daelemans, W., Kempen, G. Artificial-Intelligence Tools for Grammar and Spelling Instruction. *Instructional Science, 16, 4* (1987), 319-336.

(Pollock & Zamora, 83) Pollock, J. J., Zamora, A. Collection and characterization of spelling errors in scientific and scholarly text. *J. Amer. soc. Inf. Sci., 34, 1* (1983), 51-58.

(Pollock & Zamora, 84) Pollock, J. J., Zamora, A. Automatic spelling correction in scientific and scholarly text. *Communications of the ACM, 27, 4* (1984), 358-368.

(Rhyne & Wolf, 91) Rhyne, J. R., Wolf, C. G. Paperlike user interfaces. RC 17271 (#76097), IBM Research Division, T. J. Watson Research Center, Yorktown Heights, Nova Iorque, 1991.

(Robertson & Willet, 92) Robertson, A. M., Willet, P. Searching for historical word-forms in a database of 17th-century English text using spelling corrector methods. *In Proceedings of the 15th Annual International SIGIR Meeting* (SIGIR'92, Dinamarca), 256-265. ACM, Nova Iorque, 1992.

(Rocha Lima, 92) Rocha Lima, C. H. *Gramática Normativa da Língua Portuguesa.* Livraria José Olympio Editora, 31ª edição, 1992.

(Russel & Norvig, 95) Russel, S. J., Norvig, P. *Artificial Intelligence: A Modern Approach.* Prentice-Hall International, Inc., 1995**.**

(Sacconi, 92) Sacconi, L. A. *Gramática Essencial da Língua Portuguesa.* 9ª ed., Atual Editora,

1992.

(Sandmann, 89) Sandmann, A. J. *Formação de Palavras no Português Brasileiro Contemporâneo.* Scientia et Labor (Editora da UFPR)/Ícone Editora, 1989.

(Sengupta & Chaudhuri, 96) Sengupta, P., Chaudhuri, B. Morphological Processing of Indian Languages for Lexical Interaction with Application to Spelling Error Correction. *In Sadhana-Academy Proceedings In Engineering Sciences, 21,* Part 3, Jun. 1996, 363-380.

(Shieber, 86) Shieber, S. M. *An Introduction to Unification-based Approaches to Grammar.* CSLI Lecture Notes Series, Chicago: University of Chicago Press.

(Shinghal & Toussaint, 79a) Shinghal, R., Toussaint, G. T. Experiments in Text recognition with the modified Viterbi algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (PAMI-1)*, 4* (1979), 184-193.

(Shinghal & Toussaint, 79b) Shinghal, R., Toussaint, G. T. A bottom-up and top-down approach to using context in text recognition. *Int. J. Man-Machine Studies, 11* (1979), 201-212.

(Sidorov, 79) Sidorov, A. A. Analysis of word similarity on spelling correction systems. *Program. Comput. Softw., 5* (1979), 274-277.

(Silva & Koch, 94) Silva, M. C. P. S., Koch, I. V. *Lingüística Aplicada ao Português: Morfologia.* Cortez Editora, 7ª edição, 1994.

(Spenke et al., 84) Spenke, M., Beilken, C., Mattern, F., Mevenkamp, M., H. M. A language independent error recovery method for LL(1) parsers. *Software—Practice & Experience, 14, 11* (1984).

(Srihari et al., 83) Srihari, S., Hull, J. J., Choudhari, R. Integrating diverse knowledge sources in text recognition. *ACM Transactions on Office Information Systems, 1, 1* (1983), 68-87.

(Srihari, 84) Srihari, S. *Computer Text recognition and Error Correction.* IEEE Computer Society Press, 1984.

(van Berkel & De Smedt, 88) van Berbel, B., De Smedt, K. Triphone Analysis: a Combined Method for the Correction of Orthographical and Typographical Errors. *In Proceedings of the 2nd Applied Natural Language Processing Conference.* Association for Computational Linguistics, 1988.

(Veronis, 88a) Veronis, J. Computerized correction of phonographic errors. *Computer Hum, 22* (1988), 43-56.

(Veronis, 88b) Veronis, J. Morphosyntactic correction in natural language interfaces. *In Proceedings of the 12th International Conference on Computational Linguistics*, 708-713, 1988.

(Wagner, 74) Wagner, R. A. Order-n correction for regular languages. *Communications of the ACM, 17, 5* (1974), 265-268.

(Webster, 83) *Webster's New World Misspeller's Dictionary*. Editora Simon and Schuster, Nova Iorque, 1983.

(Wright & Newell, 91) Wright, A. G., Newal, A. F. Computer help for poor spellers. *British Journal of Educational Technology, 22, 2* (1991), 146-148.

(Yannakoudakis & Fawthorp, 83a) Yannakoudakis, E. J., Fawthorp D. An intelligent spelling corrector. *Information Processing & Management, 19, 2* (1983), 101-108.

(Yannakoudakis & Fawthorp, 83b) Yannakoudakis, E. J., Fawthorp D. The rules of spelling errors. *Information Processing & Management, 19, 2* (1983), 87-99.

(Zamora et al., 81) Zamora, E. M., Pollock, J. J., Zamora, A. The Use of Trigram Analysis for Spelling Error Detection. *Information Processing & Management, 17, 6* (1981), 305-316.

(Zhao & Truemper, 99) Zhao Y, Truemper K. Effective Spell Checking by Learning User Behavior. *Applied Artificial Intelligence, 13, 8* (1999), 725-742.

Índice Remissivo

A

acento diferencial, **66**
acentuação gráfica, **62**–**70**
  acento default, **62**
  como último recurso, **63**
  regra 1 - acento default, **62**
  regra 2 - do último recurso, **63**
  regra 3 - divisão silábica, **65**
  regras tradicionais, 63–64
aconselhamento ortográfico, **1**
alofones, **56**
análise de n-gramas. *consulte* técnicas
análise reversa. *consulte* utilidade
aprendizado
  de novas palavras em tempo de execução, 11
  por parte do usuário. *consulte* valor educacional
argüir, **62**
argumento de utilidade, 38–41
  atributos, **40**
  componentes, **39**
  critério de utilidade, **43**
  e perfil do usuário, 38
  força, **43**
  função-atributo, 42
  função-reconstituição, 43
arquifonema, **72**
assilábico. *consulte* vogal assilábica

B

barras, **71**, 72
biblioteca KLS-GT, **45**
bigrama. *consulte* n-grama (n=2)
bloqueio, **9**
  formas bloqueantes, **10**

Ch

chaves de similaridade. *consulte* técnicas

C

coda, **66**, *consulte* sílaba
colchetes, **71**
componente fonético-fonológica, **62**
confusão, **85**
  versão final, **88**
conjunto das partes, **43**
conselheiro ortográfico, **2**, *consulte* valor educacional
  acadêmicos para o português, 11, **27**–**28**
  comerciais para o português, 5–11
conselheiro qualquer, 50
corretor
  interativo qualquer, 50
  ortográfico. *consulte* conselheiro ortográfico
critério de utilidade, **43**

D

deturpação estável, **81**
deturpação fonológica, **81**
deturpação neutralizável, **81**
  motivação, 82
dissimilaridade. *consulte* similaridade
ditongo, **64**, *consulte* encontro vocálico
  crescente, **65**
  decrescente, **65**
divisão silábica
  ditongos e hiatos, **65**

E

e (grau de pertinência), **87**

educação. *consulte* valor educacional
encontro vocálico
  grafia, 65
erro
  de classificação, **83**–**93**
  de inferência de forma básica, **82**
  morfológico, **83**–**93**
  versus item malformado, 46
expressividade, **57**
`extends`, **89**

F

fluxo de certeza, 67
fone, **71**
fonema, **71**
fonemas/fones específicos
  [ɲ], 57
  [ʧ] e [t], 56
fonêmico, **70**
fonético, **70**
fonológico, **70**
força de um argumento de utilidade, **43**
forma
  básica, **70**
forma básica
  objetivo último, 72
fui *vs.* possui, **75**
função-atributo, 42
função-reconstituição, 43

G

gato, **46**
grafema, **23**
grafemização, **80**
gramática de palavras, **88**
gramática de reconstituição, **51**
grau de pertinência, **87**

H

hiato, **64**, *consulte* encontro vocálico
hipercorreção, **80**
hiper-regularização, **79**

K

KLS-GT (biblioteca), **45**

L

leitura
  integral, **58**
  precária, **58**

M

medida de utilidade. *consulte* utilidade
microondas, **46**
mínima distância de edição. *consulte* técnicas
morfologia, **83**–**93**
motor de reversão, **51**

N

não-vocábulo, **15**
neologia. *consulte* bloqueio
  tentativa frustrada, 10
  teste de hipótese, 10
neologismo. *o resultado da* neologia
netralização absoluta
  ditongos e hiatos, **74**–**77**
neutralização, **72**
  absoluta, **73**
n-grama, **32**, *consulte* técnicas
nh, 57
nível
  fonêmico/fonológico, **70**
  pós-lexical, 73
  pré-lexical, 73

O

onglide, **74**
onset, **66**, *consulte* sílaba
ordenação processual, **73**
ortografia, **49**, *consulte* sistema ortográfico
otimismo, **46**

e argumento de utilidade, 41

P

P(X) (conjunto das partes), **43**
paradigma, **88**
paradigma reverso. *consulte* reversão
paradigmas de correção, 25
ponto de erro, **48**
português
conselheiros ortográficos acadêmicos, 11, **27**–**28**
conselheiros ortográficos comerciais, 5–11
possui *vs.* fui, **75**
precisão
da primeira sugestão, **5**
das n primeiras sugestões, **5**
processo
pós-lexical, **73**
pré-lexical, **73**
processo fonológico, **71**
produção de vocábulos, **49**
profundidade, **46**

R

realismo, **41**
reconstituibilidade, **41**
reconstituição, **39**
redes neurais. *consulte* técnicas
regra fonológica, **71**
regras. *consulte* técnicas baseadas em regras
reversão, **48**, *consulte* técnicas reversas
de erros ortográficos, **48**–**49**
genérica, **46**–**48**
justificativas, **44**
motor de, **51**
passo a passo, 90

S

sal, **71**
segmento, **71**
sílaba
coda, **66**
onset, **66**
silábico. *consulte* vogal silábica
sistema de escrita. *consulte* sistema ortográfico
sistema ortográfico, **55**
divergências e incosistências, 60–61
expressividade, **57**
falácias, 55–57
vocação étimo-morfológica, **78**
vocação fonética, **62**, **79**
vocação fonológica, **77**
spell-checker. *consulte* conselheiro ortográfico
suspeita, **66**

T

técnicas (de correção ortográfica)
absolutas, **24**
análise de n-gramas, **32**
baseadas em regras, **31**
chaves de similaridade, **29**
estatísticas, **24**
lingüísticas, **24**
mínima distância de edição, **26**
redes neurais, **34**
relativas, **24**, 32
reversas, **25**
taxonomia, **24**–**26**
transcrição fonética, **78**
trema, **63**
trigrama. *consulte* n-grama (n=3)
tuiuiú, **68**

U

unidade de grafia, **78**
unigrama. *consulte* n-grama (n=1)
utilidade, **5**, *consulte* argumento de utilidade
análise reversa, **5**, 38, 39, 40
contra flexões demais, 8
medida, 42–43
por hipótese, **41**
versus precisão, 16

V

valor educacional, 3
verossimilhança, **40**
vocação. *consulte* sistema ortográfico
vogal
    assilábica, **63**
    silábica, **63**
vôo *vs.* vou, **75**

~

/ / - barras, **71**, 72

□

[ ] - colchetes, **71**