Isabel Fraústo
ILTEC

Helena Soares
ILTEC

LINCE
um corrector ortográfico português

0. Introdução

O que aqui nos traz é o interesse pelo estudo e tratamento da língua, mais especificamente pelo tratamento informático da língua portuguesa. O que temos para apresentar já não é um projecto de investigação mas sim um pré-produto, o resultado de dois anos de trabalho sobre uma ideia do Professor Ernesto d'Andrade e sob a sua coordenação. O ILTEC apostou neste produto quando ele era apenas uma ideia, transformou-a em projecto e disponibilizou os meios técnicos, informáticos e humanos, que tornaram possível a sua concretização.

Depois foi tudo uma questão de persistência, a persistência própria de quem trabalha sem qualquer apoio externo, quer de instituições quer das empresas que na fase final acabam por interessar-se pela sua comercialização. Mas disto falaremos no ponto 1.

Nos pontos seguintes faremos uma breve descrição do corrector/hifenizador LINCE, do seu funcionamento e das futuras utilizações possíveis.

1. A Indústria e a Língua Portuguesa

A indústria portuguesa parece não estar atenta à necessidade de investir nesta área, o que acontece porque tem havido no estrangeiro quem se encarregue disso: "Se os americanos integram dicionários do português no software que utilizamos, porque é que nós portugueses havemos de investir na construção de dicionários informatizados?".

As respostas são de vária ordem e todas apontam para a mesma evidência: Porque os americanos estão a olhar para o mercado brasileiro, privilegiando essa variante do português: porque os dicionários construídos na América são informaticamente "fortes" e linguisticamente "fracos", o que se compreende dado o perfil das equipas de trabalho; porque

a competência linguística dessas equipas relativamente ao português é, de forma geral, a competência de falantes de uma L2 ou, simplesmente, porque não queremos que os americanos, os japoneses, os alemães ou... resolvam os nossos assuntos, e a língua que falamos é um assunto nosso.

Posta deste modo a questão parece ser política, e de facto é-o desde o momento em que se tornou uma questão económica cujas estratégias parecem suplantar os interesses culturais. Além do mais, é, no mínimo, estranho que os interesses comerciais imediatos abafem até ao silêncio total a nossa competência e as vantagens comerciais a médio e longo prazo da produção nacional.

Embora não caiba aqui esgotar este assunto, não podíamos deixar de referi-lo, sobretudo porque esta atitude da indústria portuguesa tem comprometido o desenvolvimento da investigação e dos trabalhos em curso nesta área.

De facto, talvez com um pouco de culpa nossa e pela forma como nos temos apresentado, não tem havido financiadores a apostar em grandes ideias para o estudo e tratamento do português. O melhor que se tem conseguido, salvo raríssimas excepções que convém referir mas que por alguma razão não conseguimos recordar, é convencer a indústria a comercializar protótipos ou produtos já desenvolvidos, o que à partida elimina a competitividade e não dá ao público qualquer garantia de qualidade.

2. Descrição

Após uma análise detalhada dos correctores existentes chegou-se à conclusão de que funcionavam com base em simples listas de palavras. Foi possível delectar que às palavras registadas nesses correctores não foi acrescentada informação morfológica suficiente para tratamento da flexão, nem regras fonológicas para divisão silábica. Isto significa que estes utilitários apenas incluem aquilo a que as empresas de informática chamam "dicionário" (e que consiste numa lista de palavras flexionadas), um algoritmo de reconhecimento para comparação de um texto com as palavras desse dicionário, e um algoritmo de sugestão que funciona quando no texto surge alguma palavra que não se encontra no dicionário.Quanto aos hifenizadores, tanto quanto nos foi dado observar, parecem ser meras adaptações dos hifenizadores de língua inglesa. Um corrector deste tipo torna-se insuficiente pois depende quase 100% de um "dicionário" e verificamos que por vezes foi dada entrada a todas as flexões de uma palavra mas que há muitas palavras em que não foram tratados aspectos tão essenciais como o género ou o número. Algo de estranho acontece ainda com a utilização de pronomes clíticos: num dos mais completos desses correctores verificámos que não foi introduzido o pronome "lhe", o que faz com que ao utilizar-se qualquer forma verbal com este pronome, o corrector identifique um "erro"; por outro lado, e ainda no mesmo produto,

e ainda no mesmo produto, não há qualquer restrição para a utilização de pronomes clíticos e o resultado é que, reconhecendo a unidade lexical, o pronome e o hífen, o corrector considerará correctas formas como, "lutas-me", "neva-se" etc..

Estes são alguns dos problemas que o corrector LINCE conseguiu ultrapassar graças à sua natureza e aos seus princípios de funcionamento.

O corrector/hifenizador LINCE inclui uma lista de 40.000 palavras classificadas segundo um modelo de geração predominantemente ortográfico, com algumas especificações morfológicas e sem qualquer especificação semântica. Partiu-se do princípio de que é possível fornecer ao computador uma série de regras que lhe permitem flexionar correctamente as palavras, evitando-se esquecimentos e garantindo uma verificação automática de formas com alguma complexidade (como é o caso dos verbos com alternância vocálica e dos verbos com conjugação pronominal). Por outro lado, a existência de um algoritmo de geração permite fazer a verificação de cerca de 600.000 palavras do português (número francamente superior ao que é comum encontrar) sem afectar a rapidez necessária a um programa deste tipo. A não introdução de especificações semânticas permite uma significativa redução do número de entradas e evita situações de duplicação quer de entradas a reconhecer quer de formas a sugerir.

É ainda de salientar que mediante um mínimo de especificações adicionais poderá ser facilmente feita a actualização ortográfica deste analisador em 1994, data em que se prevê a entrada em vigor do Acordo Ortográfico de 1990.

3. Funcionamento

O corrector hifenizador LINCE funciona recorrendo aos elementos que a seguir se descrevem:

3.1. Lista de 40.000 palavras do português europeu

Estas palavras, seleccionadas a partir de um corpus inicial de 80.000 palavras, foram devidamente classificadas segundo o modelo definido no algoritmo de geração com separação de entradas verbais e não verbais. Foram incluídos nomes próprios para tratamento de maiúsculas, abreviaturas gerais e interjeições.

3.2. Algoritmo de geração

Foi definido um conjunto de regras de geração que permite ao computador flexionar as 40.000 entradas classificadas expandindo-as a cerca de 600.000 palavras. Além dos aspectos referidos em 2., estas regras visam optimizar o desempenho do corrector, sobretudo no que diz respeito à rapidez de execução e à ocupação em disco. Por este motivo foram adoptadas estratégias que privilegiam o aspecto ortográfico em detrimento dos traços semânticos ou mesmo da análise morfológica. Por exemplo, poderá parecer aberrante que a palavra "bota" seja formada a partir da entrada "bolo" segundo as mesmas regras que são utilizadas para o tratamento do género, no entanto, e porque o que está em causa é apenas a ortografia, este tipo de classificação permite-nos, com apenas uma entrada, abranger oito formas (as duas que referimos, os diminutivos "bolinho" e bolinha" e os respectivos plurais).

Através do algoritmo de flexão são tratados os seguintes aspectos:

3.2.1. Não-verbos:
- Género
- Número
- Flexão de palavras compostas
- Adverbialização
- Superlativização
- Diminutivos

3.2.2. Verbos
- V. regulares
- V. irregulares
- V. com alternância vocálica
- V. unipessoais
- V. impessoais
- V. defectivos
- Nominalização
- Adverbialização
- Prefixação
- Conjugação pronominal
- Conjugação pronominal reflexa

3.3. Algoritmo de hifenização

Foi concebido um algoritmo para aplicação de regras fonológicas com vista à hifenização das palavras do português. Para além do objectivo de divisão silábica foram tomadas em

consideração algumas exigências gráficas para processamento de texto e edição uma vez que inicialmente este corrector será integrado num programa de design e paginação.

3.4. Algoritmo de reconhecimento

Sendo sobretudo da responsabilidade da "software house" PRIBERAM, com quem estamos a colaborar para a implementação informática do corrector, este algoritmo permitirá comparar as palavras de um dado texto com as formas correctas flexionadas pelo LINCE, através de operações como a ponderação de semelhança das palavras do texto com as entradas do corrector, a descompactação dessas entradas e a sua flexão e hifenização. No caso de existir incompatibilidade entre uma palavra do texto e as do corrector, este assinalará uma mensagem de erro e porá em funcionamento o algoritmo de sugestão.

3.5. Algoritmo de sugestão

Além das especificações informáticas da responsabilidade da PRIBERAM, este algoritmo incluirá ainda algumas especificações determinadas pelo ILTEC no que diz respeito à análise e previsão de erros, de forma a facilitar o reconhecimento do tipo de erro e reduzir e hierarquizar as hipóteses de sugestão de formas alternativas. O utilizador poderá escolher entre as várias formas sugeridas ou manter a forma que inicialmente digitou, criando assim um dicionário pessoal.

4. Futuras utilizações

4.1. Ensino - Verbos

As dificuldades que a flexão verbal em português apresenta para a aprendizagem são por demais conhecidas e cremos que o gerador de formas verbais deste corrector poderá constituir um precioso auxiliar para o ensino do português.

Dada a sua especificidade, o ficheiro de verbos do português foi tratado individualmente e possui cerca de 6.000 entradas no infinitivo. A classificação a que estas entradas foram sujeitas e a aplicação do respectivo algoritmo de flexão constituem um gerador de formas verbais passível de múltiplas utilizações de que damos alguns exemplos.

4.1.1. Reconhecimento do infinitivo de um dado verbo e

4.1.1.1. apresentação de toda a flexão devidamente identificada com TEMPO e MODO

4.1.1.2. indicação da existência de palavras homófonas, apresentação de sinónimos e flexão do verbo pedido

4.1.2. reconhecimento de qualquer forma verbal e

4.1.2.1. apresentação do infinitivo do verbo e de toda a flexão verbal

4.1.2.2. apresentação do infinitivo do verbo com identificação do TEMPO, MODO, pessoa e número da forma em questão

4.1.2.3. apresentação do infinitivo do verbo e de toda a flexão no TEMPO e MODO apresentados

4.1.2.4. identificação de verbo inexistente ou forma incorrecta e apresentação de alternativas

4.1.3. flexão de um dado TEMPO e MODO a pedido do utilizador com indicação do infinitivo

4.1.4. Jogos didácticos

A partir do gerador de formas verbais poder-se-ia ainda considerar a hipótese da construção de jogos didácticos para o ensino da língua portuguesa no que diz respeito à flexão verbal. Vemos como possível o desenvolvimento a curto prazo de utilitários para o ensino de língua tanto a nacionais como a estrangeiros.

4.2. Thesauri

Como ficou descrito nos pontos 2. e 3., o LINCE apresenta casos de ambiguidade de classe morfológica, sem que isso prejudique o seu funcionamento. No entanto, e pensando nas futuras extensões deste corrector, está já em estudo a integração de uma componente de análise morfológica necessária para que possa ser integrado noutros produtos. A utilização do LINCE como base para a construção de thesauri implicará ainda a integração de especificações semânticas para apresentação de relações de sinonímia, antonímia, etc.

4.3. Terminologias

Paralelamente, a inclusão de vocabulário especializado permitirá a extensão da aplicabilidade deste dicionário, facilitando a sua utilização em textos técnico-científicos.

4.4 Dicionários Bilingues

Através da atribuição de equivalentes às entradas do corrector, o LINCE poderá servir de base para a construção de dicionários bilingues.

4.5. Corrector Sintáctico / Corrector estilístico

A construção de um utilitário deste tipo foi já formalizada em projecto pelo ILTEC. Trata-se do projecto GRAMÁTICO cuja finalidade é desenvolver um módulo de correcção sintática para processadores de texto, no qual o LINCE será utilizado como base lexical para a construção do analizador morfológico. Numa fase mais avançada prevê-se que esse mesmo analizador morfológico seja integrado num módulo de correcção estilística.