RESUMO

Este trabalho aborda os resultados de um estudo sobre as potencialidades de utilização de um sistema que permite o acesso em língua natural (usando um subconjunto do Português) a uma base de dados.

O sistema testado. CALCC -- Consulta Automática de Legislação sobre a Construção Civil -- é sumariamente descrito. Os resultados da sua utilização são analisados sob os pontos de vista gramatical e de facilidade de utilização. As conclusões servirão como orientação futura do desenvolvimento de sistemas que intercalem coloquialmente, e em português, com os seus utilizadores.

1.INTRODUÇÃO

O sistema CALCC -- Consulta Automática de Legislação sobre a Construção Civil -- foi construído com o objectivo de cooperar com um utilizador fornecendo-lhe legislação aplicável a determinados domínios da construcão civil. A legislação conhecida pelo CALCC encontra-se compilada em (Neto, 1977).

O processo de cooperação entre o CALCC e os seus utilizadores assenta em dois aspectos fundamentais

1. A interacção entre o CALCC e os seus utilizadores é feita em língua natural. natural, num subconjunto do Português. A gramática é escrita em Prolos (Pereira et al,1978). uma linguagem de programação de muito alto nível e baseada na lógica de predicados. (esta gramática permite a análise sintáctica e semântica das frases; do utilizador.

ii) O sistema CALCC orienta o utilizador na procura de legislação que de facto se aplica a uma dada situação, através da focagem dos seus aspectos mais importantes.

O aspecto da interacção em língua natural permite que o sistema esteja ao alcance de utilizadores não informáticos. pois a utilização do CALCC não pressupõe nenhum tipo de conhecimentos de ordem informática, que ultrapassem o ligar e desligar de um terminal e o formato do protocolo de entrada em comunicação com o computador. Construiu-se assim um sistema orientado para o utilizador de forma a torná-lo o mais atractivo possível.

Todos os protocolos das conversas entre o sistema CALCC e os seus utilizadores foram gravados em ficheiro. A sua análise incidiu sobre

1)Ponto de vista gramatical -- Testar até que ponto o subconjunto do Português considerado se encontra adaptado quer ao vocabulário quer as construções gramaticais mais frequentemente utilizadas no mundo da consulta de legislação.

2)ponto de vista de facilidade de utilização -- Testar o grau de facilidade de utilização do sistema, analisando os motivos do não entendimento das frases dos utilizadores. Extrair conclusões para um futuro desenvolvimento de sistemas deste tiro. tendo em vista a construção de grandes sistemas informáticos cada vez mais orientados para o utilizador.

Antes de se apresentarem os resultados da análise dos protocolos. far-se-á uma descrição breve do sistema CALCC, de modo a introduzir o leitor no seu mundo de conhecimentos e capacidades.

2.BREVE DESCRIÇÃO DO SISTEMA CALCC

após ser lançado, o sistema CALCC encontra-se completamente à disposição do utilizador e expressa-o através da seguinte mensagem :

Olá

Bem vindo ao programa CALCC. Posso ajudá-lo fornecendo-lhe legislação aplicável a certos domínios da

construção civil. Possuo toda a legislação contida no Relatório de Actualização 5 da Relação das Disposições legais a Observar pelos Técnicos Responsáveis dos Projectos de Obras e sua Execução.

Publicada em Setembro de 1977.

Sempre que pretenda que seja eu a conduzir a conversa escreva 'fale.' se não escreva factos ou Perguntas.

Esta mensagem evidencia dois modos distintos de utilizar o

programa : um deles permite que o CALCC tome o controlo da conversar isto e'. o Programa pergunta e o utilizador responde: o outro, assume que o utilizador toma a iniciativa da conversa.

Estas duas possibilidades de utilizar o sistema assentam na definição de duas modalidades de diálogo. que chamaremos

"Dialogo 1" e Diálogo 2'. respectivamente.

A modalidade Diálogo 1 assenta no estabelecimento de um mapa de questões que o CALCC coloca ao utilizador permitindo-lhe, contudo. uma certa liberdade tanto na forma como na ordem pela qual as respostas são dadas

Salientamos que durante uma mesma sessão o utilizador pode usar qualquer das modalidades de diálogo ou até mesmo usar qualquer combinação destas. Assim una protocolo e' composto. regra geral. ror mais do que um diálogo.

Durante um diálogo com o utilizador o CALCC usa dois tipos de conhecimentos

Conhecimentos referente; 'as palavras da língua portuguesa. que se encontram agrupados num módulo que chamamos 'Dicionário'.

--Conhecimentos referentes à legislação aplicável aos domínios da construção civil. que se encontrem estruturados numa base de dados (representada também em lógica de predicados) a que chamamos "documentação Legal".

O CALCC e' constituído pelos seguintes módulos :

1>Módulo de Controlo -- Componente do sistema que toma a decisão sobre a modalidade de diálogo que o utilizador pretende seguir.

2>Módulo de Diálogo 1 -- Componente que executa a modalidade de diálogo em que o programa controla a conversa e interroga o utilizador de acordo com o mapa de questões.

3>Módulo de Diálogo 2 -- Componente sue executa a modalidade de diálogo em sue o utilizador controla a conversa.

4>Módulo Gramática -- Componente que cobre a sintaxe e a semântica do subconjunto do Português considerado. Este modulo e' autom aticamente invocado para analisar todas as contribuições dos utilizadores quer sejam perguntas ou respostas. A gramática utiliza informação armazenada no modulo Dicionário.

5>Módulo de Acesso -- Recebi o fluxo de controlo após a contribuição do utilizador ter sido analisada Pela Gramática e executa o acesso à base de dados Documentação Legal de acordo com o pedido de utilizador.

6>Módulo Tradutor -- Recebe a informação seleccionada pelo Acesso e fornece uma saída em língua natural 'utilizando informação armazenada no Dicionário.
representam acesso dos modulas respectivos a estruturas de informação. A organização modular do sistema CALCC tem reflexos na sequência dos estados da conversa. A figura 2 repres enta a

sequência de estados de conversa obtida com o CALCC.

Figura 1: Arquitectura do sistema CALCC

A figura 1 esquematiza a arquitectura do sistema CALCC e realça a articulação dos diversos módulos entre si. As setas largas indicam passagem do fluxo de controlo e as setas estreitas

Dentro de cada modalidade de dialogo a sequência de estados respectivamente. da conversa e representada pela Figura 3

Figura 3: Sequência de estados da conversa em. cada diálogo

Uma descrição mais detalhada. tanto do sistema CALCC como dos aspectos teóricos que orientaram a sua implementação. pode ser encontrada em (Cotta & Silva.1970).

3.RESULTADOS DA EXPERIÊNCIA E UTILIZAÇÃO

No parágrafo anterior descreveu-se sumariamente o sistema CALCC e a sua arquitectura modular. Este parágrafo aborda com algum detalhe os resultados da experiência de utilização do CALCC. os quais foram reunidos em 25 protocolos de diversos tipos de utilizadores, a maioria dos quais são informáticos.

Dado que o sistema CALCC permite ao utilizador recorrer a um numero qualquer de diálogos durante uma sessão. torna-se importante estudar o número de diálogos por protocolo na amostra recolhida. O primeiro histograma realça a distribuição do número de diálogos por protocolo. À esquerda podemos encontrar o histograma propriamente dito que permite medir o comprimento de um protocolo em número de diálogos. As três colunas da direita representam o número de diálogos com. esse comprimento, a respectiva frequência na amostra e as frequências acumuladas.

Figura 4: Histograma do número de diálogos por protocolo

Sob o ponto de vista de facilidade de utilização o valor médio do numero de diálogos por protocolo é satisfatório. apesar de a sua distribuição ser algo assimétrica.

Dado que o utilizador do CALCC Pode optar por duas modalidades distintas de diálogo é importante estudar a frequência de utilização do Diálogo 1 e do Diálogo 2. O diagrama da figura seguinte apresenta o número de utilizações dos dois diálogos e a respectiva percentagem na amostra recolhida.

Figura 5: Frequência de utilização das duas modalidades de diálogo

É importante realçar a grande percentagem de utilização do Diálogo 1 comparativamente ao Diálogo 2. 73% contra 27%. Este facto pode ser justificado pela maior facilidade de utilização do Diálogo 1 principalmente para os utilizadores iniciados. isto é para aqueles que utilizam pela primeira vez o sistema. Conclui-se que o sistema está vocacionado para a cooperação com utilizadores iniciados exigindo. portanto. um desenvolvimento mais acentuado desta modalidade de diálogo.

O estudo da experiência de utilização do CALCC sob o ponto de vista de facilidade de utilização passa necessariamente pela análise do 'comprimento' dos diálogos em número de contribuições (perguntas ou respostas) dos utilizadores O histograma seguinte exibe a distribuição das contribuições dos utilizadores por dialogo.

Figura 6: Histograma do número de contribuições por diálogo

A maior percentagem de diálogos com 5 contribuições é justificada pelo facto de um diálogo bem

sucedido na modalidade Diálogo 1 ser composto por contribuições do utilizador e pelo facto de ser esta a modalidade de diálogo mais frequentemente utilizada. A existência de diálogos com um grande comprimento em termos de contribuições dos utilizadores sugere dificuldades na interacção com o sistema. Este ponto. intimamente relacionados com o número de contribuições não entendidas pelo sistema será abordado adiante.

O diagrama seguinte representa a distribuição das 309 contribuições dos utilizadores Pelas duas modalidades de dialogo e os respectivos valores médios.

Figura 7:Repartição das contribuições pelas duas modalidades de diálogo

A existência de um valor médio 6.5 do número de contribuições no Diálogo 1 significa que existe uma dificuldade embora pequena. na interacção com o CALCC se atendermos ao facto de que um diálogo bem sucedido nesta modalidade ter comprimento 5. Quanto ao Diálogo 2 nada se pode concluir visto que assumindo o utilizador o controlo da conversa nesta modalidade de diálogo. o respectivo comprimento em termos de contribuições é absolutamente aleatório.

Um factor muito importante tanto no que respeita ao ponto de vista de facilidade de utilização como ao ponto de análise gramatical, é o comprimento das contribuições do utilizador em número de palavras. O estudo deste factor vai permitir ao leitor aperceber-se melhor das situações de não entendimento entre o sistema e o utilizador quando estas forem utilizadas.

Figura 8: Histograma do número de palavras por contribuição

O valor médio 2.6 obtido para a distribuição do número de palavras por contribuição é extremamente baixo devido à grande percentagem 54.4% correspondente a 168 contribuições com apenas uma palavra. Este número pode ser explicado por dois motivos: em primeiro lugar a modalidade de diálogo mais usada na amostra que estamos a analisar coloca o CALCC a interrogar e o utilizador a responder em segundo lugar o CALCC permite que o utilizador lhe responda apenas com uma palavra não lhe exigindo a construção de uma frase completa. Assim, por exemplo, quando o programa pergunta: "Que tipo de construção pretende? "o utilizador tem a liberdade de responder "supermercado" interpretando o CALCC, a resposta como equivalente a ."pretendo construir um supermercado". O grande número de contribuições com uma só palavra reflecte também as limitações linguísticas actuais do CALCC. Em alguns casos estas contribuições são antecedidas por outras de maior comprimento que não foram compreendidas o que contribui para o seu condicionalismo.

A amostra recolhida possui um número total de 114 contribuições dos utilizadores que não foram entendidas pelo programa. Isto corresponde a uma frequência de 37%, percentagem elevada para as características do sistema.
Foram seleccionadas 3 situações típicas em que a contribuição do utilizador não é compreendida pelo sistema:

--utilização de uma construção gramatical desconhecida (não considerada na Gramática)

--utilização de uma palavra desconhecida (não pertencente ao Dicionário)

--erros de escrita

e foram agrupadas todas as outras situações possíveis. A distribuição das contribuições não compreendidas encontra-se representada no quadro da figura 9

Figura 9 Motivos das contribuições não correspondidas

Esta distribuição permite tirar as seguintes conclusões :

--A percentagem de erros de escrita é bastante pequena contrariamente ao que se previa. Este facto demonstra a existência de um certo cuidado por parte dos utilizadores na interacção com o sistema.

--A grande percentagem' de utilização de palavras desconhecidas sugere uma ampliação do dicionário

do sistema tanto no aspecto sintáctico como no aspecto semântico.

--A percentagem de construções gramaticais desconhecidas sugere também a criação de novas regras gramaticais.

Estes dois últimos aspectos constituirão linhas de desenvolvimento do sistema, com vista à melhor adaptação do subconjunto do Português ao mundo da consulta de legislação.

A repartição deste tipo de situações pelas duas modalidades de diálogo pode ser encontrada nos quadros das figuras 10 e 11

ESTAS PERCENTAGENS TÊM COMO BASE O NÚMERO DE CONTRIBUIÇÕES EM CADA UMA DAS MODALIDADES DE DIALOGO. são PERCENTAGENS RELATIVAS.

Figura 10: Repartição das contribuições não compreendidas pelas duas modalidades de diálogo

Figura 11: Repartição dos motivos de não compreensão pelas duas modalidades de diálogo

A observação dos quadros das figuras anteriores revela uma Pequena percentagem de situações de não compreensão no Dialogo 1, enquanto essa Percentagem é bastante elevada no Diálogo 2.

Por outro lado, os quadros anteriores permitem também concluir que o motivo mais importante de não compreensão no Diálogo 1, consiste na utilização de palavras desconhecidas (52%) e no Diálogo 2. consiste na utilização de construções gramaticais desconhecidas. Resultado óbvio. se atendermos à natureza das duas modalidades de dialogo, Um resultado interessante consiste no facto de a percentagem de erros de escrita ser superior no Dialogo 2 do que no diálogo 1.

Os 63% de frequência de situações de não entendimento no Diálogo 2 são provenientes em 39% da utilização de construções gramaticais desconhecidas. 13% de erros de escrita e 11% da utilização de palavras desconhecidas. Este facto sugere que as regras gramaticais. constantes da gramática do sistema. necessitam ser modificadas para que o subconjunto do Português se encontre adaptado de facto ao mundo da consulta de legislação.

4 ORIENTAÇÃO FUTURA NO DESENVOLVIMENTO DO CALCC

Os resultados analisados no parágrafo anterior sugerem algumas linhas de desenvolvimento dos sistemas que interactuam coloquialmente em Português, e em particular do sistema CALCC.

As principais linhas de desenvolvimento do CALCC visam três objectivos

1) alteração da arquitectura

2)ampliação do seu poder conversacional e

3) actualização do conhecimento

Este parágrafo é dedicado 'à argumentação da validade destas linhas de desenvolvimento. tendo em consideração por um lado, os resultados da experiência de utilização e. por outro lado, a necessidade de construir sistemas cada vez mais orientados para os utilizadores não informáticos.

alteração da arquitectura

Os resultados da experiência de utilização sugerem que um utilizador do sistema CALCC pretende encontrar uma orientação que lhe não pode ser dada pela consulta directa de (Neto.1977). Estes resultados mostram uma maioria de utilizações da modalidade de Diálogo 1 (Que executa esta orientação). respectivamente 73%, contra 23% de utilizações do Diálogo 2 (Que inclusivamente lhe

permite uma maior liberdade) Estas percentagens sugerem a evolução da arquitectura do sistema no sentido da fusão das duas modalidades de diálogo. Uma só modalidade de diálogo orientadora do utilizador (equivalente em alguns aspectos ao Diálogo 1. nomeadamente na existência de um mapa de questões) permitiria ao utilizador interrogar o sistema quando este espera uma resposta. Assim, as perguntas do utilizador dariam origem a sub-diálogos organizados em cenários. através dos quais o sistema modelaria a sua conversa.

Ainda sobre a alteração da arquitectura do sistema, outro factor merece ser desenvolvido: a base de dados do sistema e o seu modulo de acesso. tendo em vista permitir o acesso directo a disco.

Esta limitação actualmente existente na linguagem de programação Prolos obriga ao carregamento para memória central da base de dados do sistema. o que o torna bastante pesado em termos de espaço : 65 k após o carregamento e. em media. 88 k após a execução. Pensamos que estes valores podem ser consideravelmente diminuídos se a linguagem Proloq for dotada de um algoritmo de acesso directo a disco.

Os tempos de resposta, actualmente conseguidos, são em média 2.18 s para o Dialogo 1 e 3.80 s para o Dialogo 2 (considerando tempo de análise da contribuição e tempo de pesquisa na base de dados) Estes tempos estão intimamente relacionados com a não existência de acesso directo a disco. No entanto, esta dependência não é muito clara. For um lado, a base de dados em memória central deveria diminuir o tenro de acesso à informação. Por outro lado o seu tamanho implica a existência de frequentes acções de swapping, o que aumenta significativamente o tempo de resposta.

Resumidamente, no que respeita a alteração da arquitectura do sistema, os desenvolvimentos mais urgentes são os seguintes

- Estruturação dos diálogos em cenários gerados a partir de uma única modalidade de diálogo orientadora dos utilizadores.

- Reformulação da base de dados e do módulo de acesso de acordo com um algoritmo de acesso directo a disco.

Ampliação do poder conversacional

A não adaptação do subconjunto do Português ao mundo da consulta de- legislação é evidenciada pelas grandes percentagens de frases não entendidas por utilização de palavras ou construções gramaticais desconhecidas, respectivamente 40% e 26%. Isto exige um investimento na ampliação do dicionário do CALCC e na ampliação da sua gramática. Por outro lado, no que respeita ao dicionário, este desenvolvimento tem que ser dependente da reformulação da base de dados visto o seu tamanho aconselhar o armazenamento em disco.

A percentagem de ocorrência de erros de escrita de 19% justifica um investimento no sentido de dotar o CALCC com um mecanismo que lhe permita não só detectar os erros do utilizador como também corrigi-los de uma forma acordada com este ultimo.

Pensamos num mecanismo que, ao detectar um erro de escrita, avise o utilizador propondo-lhe uma série de palavras alternativas de resolução da situação de erro. Prosseguindo de seguida a análise da frase com a nova palavra escolhida pelo utilizador.

A construção de sistemas cuja forma de interacção se aproxima cada vez mais do diálogo humano usual implica a construção de mecanismos que Permitam executar tarefas inerentes à compreensão da língua natural, como Por exemplo a resolução de elisões e de referencias Pronominais.

Em resumo, Propomos que sejam considerados os seguintes aspectos na ampliação do Poder conversacional do CALCC

--Ampliação do dicionário do sistema

--Ampliação das construções sintácticas da gramática

--Detecção dos motivos de não compreensão das frases e construção

de mecanismos para o tratamento dos erros de escrita

--Resolução de elisões

- Resolução da referencia Pronominal

Actualização do conhecimento

A questão da actualização do conhecimento é de fundamental importância num sistema com as características do CALCC orientado para facilitar o mais possível a sua utilização. O processo usual, que consiste em recorrer a programas de edita gem e manipulação de ficheiros para actualizar os componentes da base de dados do sistema, é bastante custoso e envolve o trabalho de técnicos de legislação (detectar as alterações a fazer) e de informática (proceder a essas alterações).

Uma linha de desenvolvimento do CALCC consiste em construir mecanismos que permitam uma alteração dinâmica da sua base de dados. Concretamente propõe-se a concepção de uma modalidade de dialogo, convenientemente protegida e somente ao alcance de técnicos de legislação (por razoes óbvias relacionadas com a segurança dos próprios dados) que permita ao sistema alterar a sua base de dados de acordo com as directivas do utilizador (expressas em língua natural) criando assim um sistema dinâmico de actualização do conhecimento.

Esta modalidade de dialogo poderia mesmo incidir sobre uma extensão do dicionário, não sendo permitidas, no entanto, quaisquer alterações às palavras já existentes.

Este método de actualização da base de dados de um sistema é frequentemente utilizado em programas de Inteligência Artificial, orientados para utilizadores não informáticos como é o caso, por exemplo de MYCIN (Shortliffe,1976), o qual apoia os médicos no diagnóstico e posologia de determinados tipos de infecções sanguíneas.

5. BIBLIOGRAFIA

COELHO, H. (1979)
A program conversing in Portuguese and providing a library service
Univ. of Edinburqh, Ph. D. Thesis

COTTA,J.C.;SILVA, A P (1978)
Interacção com bases de dados
LNEC

KONOLIGE, K. (197?)
A framework. for a Portable natural languase interface to
large data bases
SRi Artificial Inteligence Center
Technical Note 197

LEHMANN,H.;OTT,N.;ZOEPPRITZ, M.(1978)
User experiments with natural language for data base access
IBM Germany
Proceedings 7th International Conference on Computacional
Linguistics

NETO.F.s. (1977)
Relação das disposições legais a observar pelos técnicos responsáveis dos projectos de obras e sua execução

LNEC

PEREIRA, L.M. PEREIRA,F.C.;WARREN.D.H.D.(1978)
User's Guide to Decsystem -1O Prolos
LNEC