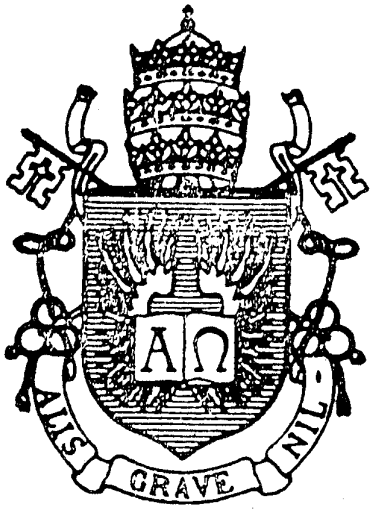


PUC

Série : Monografias em Ciência da Computação, 10/89

GRAMÁTICAS DE DETERMINAÇÃO: UMA ABORDAGEM METODOLÓGICA

Clarisse S. Souza

Departamento de Informática

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO

RUA MARQUÊS DE SÃO VICENTE, 225 – CEP 22453

RIO DE JANEIRO – BRASIL

PUC/RJ - DEPARTAMENTO DE INFORMÁTICA

Série : Monografias em Ciência da Computação,10/89

Editor: Paulo Augusto Silva Veloso May, 1989

GRAMÁTICAS DE DETERMINAÇÃO: UMA ABORDAGEM METODOLÓGICA

Clarisse S. Souza

Trabalho parcialmente patrocinado pela FINEP

Responsável por publicações:

Rosane Teles Lins Castilho
Assessoria de Biblioteca, Documentação e Informação
PUC RIO, Departamento de Informática
Rua Marquês de São Vicente, 225 - Gávea
22453 - Rio de Janeiro, RJ
BRASIL

Tel.:(021)529-9386 TELEX:31078 FAX:(021)274-4546
BITNET: userrtlc@lncc.bitnet

GRAMÁTICAS DE DETERMINAÇÃO: UMA PROPOSTA METODOLÓGICA

CLARISSE SIECKENIUS DE SOUZA

EMPRESA BRASILEIRA DE TELECOMUNICAÇÕES
AVENIDA PRESIDENTE VARGAS 1012
RIO DE JANEIRO - RJ / BRASIL

PONTIFÍCIA UNIVERSIDADE CATÓLICA
DEP. DE INFORMÁTICA
R. MARQUÊS DE SÃO VICENTE 225
RIO DE JANEIRO - RJ / BRASIL

RESUMO: O desenvolvimento de interfaces em **linguagem natural** para **bancos de dados** requer que se desenvolvam metodologias capazes de proporcionar ao usuário um máximo de **flexibilidade** na maneira de expressar-se, juntamente com toda a **precisão** esperada para a recuperação de informações em sistemas desta natureza. As **GRAMÁTICAS DE DETERMINAÇÃO** (GD's) apresentam uma proposta metodológica para o desenvolvimento da base de conhecimentos linguísticos em interfaces "naturais" para bases de dados.

1 - INTRODUÇÃO

O processamento de linguagem natural (PLN) comporta pesquisas com as mais diversas orientações, dentre as quais, para fins deste trabalho, interessa destacar três. São elas: (a) a investigação acerca do mecanismo processador da linguagem natural (de interesse eminentemente **psicolingüístico**); (b) a investigação acerca dos formalismos destinados a representar o conhecimento inconsciente que todo falante tem sobre a gramática de sua língua (de interesse eminentemente **sintático**); e (c) a investigação acerca da maneira mais eficiente de se conjugarem os dois temas anteriores, visando o **desenvolvimento de bases de conhecimentos lingüísticos** em sistemas destinados à compreensão/produção automática de textos em linguagem natural.

Existe uma afinidade entre a pesquisa do tipo (a) e a **psicologia**, entre a pesquisa do tipo (b) e a **matemática** e, finalmente, entre a pesquisa do tipo (c) e o que Coelho [1981] chamou de **engenharia da linguagem**, embora se imprima aqui alguma modificação. O termo **engenharia**, aqui, caracteriza a ênfase na construção de sistemas e de todos os aspectos envolvidos nesta atividade. O presente trabalho trata especificamente desta questão: é uma proposta de metodologia para o desenvolvimento de sistemas processadores de língua portuguesa, na aplicação de consultas a bases de dados.

Embora haja diversas propostas com objetivo semelhante, a metodologia das **GRAMÁTICAS DE DETERMINAÇÃO** (GD's) apresenta a particularidade de ter sido desenvolvida sob a ótica da lingüística aplicada, não no sentido de testar no campo computacional

hipóteses produzidas no âmbito da teoria lingüística, mas antes no sentido de utilizar insumos desta teoria para produzir uma alternativa metodológica **sem comprometimento teórico a priori**, motivada apenas pelas soluções exigidas pela aplicação de consulta a Bases de Dados (BD's).

O desenvolvimento completo da proposta relativa às GD's encontra-se em Souza [1988] e as informações acerca do primeiro sistema **em operação** elaborado segundo esta metodologia podem ser encontradas em Bicharra,Carvalho & Fortuny [1986] ou Souza [1987]. Tal sistema - denominado FARAO - esteve disponível na EMBRATEL (empresa Brasileira de Telecomunicações), por um período superior a doze meses, para usuários de uma base de dados financeira.

2 - A CONSULTA A BASES DE DADOS

A consulta a bases de dados, em especial no momento de transição entre a utilização exclusiva de linguagens artificiais ou telas-menu e a possibilidade de expressão em uma língua natural, apresenta algumas características que devem ser levadas em conta no desenvolvimento de uma interface. Destacam-se (a) a reminiscência de formalismos próprios das consultas formuladas na **query language** anteriormente utilizada na aplicação, (b) a influência da estrutura da base de dados sobre a estrutura da consulta (marcando a recuperação de **dados**, e não tanto a recuperação de **informações**) e (c) o estilo "telegráfico", ou abreviado, revelador de uma sintaxe para-gramatical da língua. Exemplos de cada um destes fenômenos podem ser encontrados em Souza [1987,1988].

O desenvolvimento de interfaces em linguagem natural que não contemple as características expressivas dos usuários pode vir a ser tributado com um alto índice de rejeição de consultas formuladas, devido a uma excessiva rigidez gramatical na análise sintática. Trabalhos como o de Waltz [1978] ou o de Burton [1976] já davam conta do fenômeno. No primeiro caso, o autor propunha que, mediante fracasso na análise sintática, o processador tentasse recuperar o sentido da consulta através de uma análise semântica; no segundo, a proposta de análise pautava-se essencialmente em critérios semânticos, desde o primeiro passo, formulando-se uma **gramática semântica**. Nesta linha encontra-se também o trabalho de Hendrix e outros [1978].

O invariante dos três trabalhos citados, e de outros menos diretamente relacionados à aplicação de consulta a BD, é priorizar a informação semântica, mesmo em estágio de análise (teoricamente) sintática. Sob esta influência, as GRAMÁTICAS DE DETERMINAÇÃO funcionam com base em categorias sintáticas **semanticamente motivadas.** A denominação das GD's é inspirada na relação de **determinação** (subordinação) que se pode encontrar em constituintes complexos das gramáticas das línguas humanas, nos quais um elemento é o núcleo **(determinado)** do constituinte, enquanto outro(s) funciona(m) como modificador(es) **(determinante(s)).**

Para melhor esclarecer os pontos de vista defendidos pela proposta metodológica que aqui se apresenta, vejam-se os exemplos abaixo, onde se utiliza a linguagem de consulta **ADASCRIPT**, durante bastante tempo empregada no gerenciador **ADABAS**, e as formulações correspondentes à mesma consulta, em português.

(a) CONSULTA EM ADASCRIPT

```
FIND 59 WITH C4 = 3101 OR 3102
 IF RE NOT EQUAL SU REJECT
ACCUM A1 A2 A3 A4 A5 A6
CONTROL C4 EX
SUMMARY DISPLAY RE C4 EX A1 A2 A3 A4 A5 A6
```

(RE -> REGIÃO; SU -> SUL; C4 -> CONTA; EX ->EXECUTANTE)

(b) POSSÍVEL CONSULTA CORRESPONDENTE EM PORTUGUÊS

Qual é o orçado de janeiro a junho dos executantes da região sul, nas contas de mão de obra e material de consumo?

(c) POSSÍVEL CONSULTA CORRESPONDENTE EM PORTUGUÊS

Para os executantes da região sul, com material de consumo e mão de obra, dê o valor orçado para os meses de janeiro a junho.

Entretanto, o fato de a entrada da consulta dever ser digitada, bem como a transição entre uma forma de expressão pautada em códigos e formalismos e uma expressão "natural", aponta para a importância de se processarem consultas equivalentes a (a), (b) e (c), acima, com os seguintes perfis:

(d) CONSULTA "TELEGRAFADA"

Executantes da região sul, orçado janeiro a junho, mão de obra e material de consumo.

(e) CONSULTA UTILIZANDO CÓDIGOS CONHECIDOS PARA O USUÁRIO

Executantes da SU, contas 3101 e 3102, orçado janeiro a junho

(f) CONSULTA "ARTIFICIAL" DE SINTAXE LIVRE

```
RE = SU,  C4 = 3101 E 3102  ORÇADO JANEIRO A JUNHO
CONTROLE EX E C4
```

Embora (d), (e) e (f) não tenham nada de "natural", as condições de contorno da atividade de consulta a bases de dados poderiam favorecer o aparecimento deste tipo de enunciado, considerando-se os fatores ergonômicos (menor esforço na digitação) e os resquícios de uma tradição de consultas em linguagem artificial.

Observe-se que o invariante entre todos os enunciados (de (a) a (f)) são as "chaves" de seleção, recuperação e (opcionalmente) publicação de dados. Isto justifica a opção por se trabalhar com categorias sintáticas **semanticamente motivadas.** Na realidade, os exemplos dão mostras de "palavras-chave", ou "pontos cardeais" para a consulta. Estes são os elementeos de um **padrão mínimo** na sintaxe da mesma. Assim como, de maneira geral, as linguagens artificiais apresentam um padrão

C -> S R P

onde C é a consulta, S é o constituinte correspondente aos critérios de seleção, R o correspondente à recuperação e P o correspondente às instruções para a publicação de dados em tela ou impressora, também em linguagem "natural" (ou pseudo-natural, face às considerações acima) há um padrão mínimo invariante de consulta

B -> F P

onde B é um binômio (consulta) formado por uma parte F "fornecida" (critérios de seleção) e uma parte P "pedida" (solicitação de recuperação). Note-se que a publicação, nas consultas "naturais", é parte do bom senso: se não for especificada, deve apresentar a informação inferida a partir de F e P; caso contrário, obedece a uma sintaxe

B -> F P N

onde N é o constituinte que contém as informações relativas ao modo de publicação de dados.

Deve-se notar que o princípio da **determinação** fica bastante claro na sintaxe obtida. Um binômio B compõe-se de duas partes: uma parte **determinada** F e uma parte **determinante** P. Cada uma, por sua vez, possui uma parte determinada e uma parte determinante. No caso de F, o determinado são as regiões e/ou executantes e o determinante as contas. No caso de P, o determinado é o valor orçado ou realizado, enquanto o determinante são os meses. A atribuição de função determinada ou determinante aos elementos da base de dados **emana da modelagem da base de dados e da modelagem do usuário** (veja-se a este respeito Souza [1988]).

Por **convenção** das GD's, as classes determinadas ganham índice 1 e as classes determinantes ganham índice 3. O índice 2, também utilizado, é reservado a classes híbridas.

A título de exemplo, veja-se a consulta (g), a seguir, onde só as palavras em negrito são significativas para o sistema.

(g) Qual o **realizado** no mês de **junho** de toda a região **norte** nas contas de **mão de obra** . Dê o resultado por **executante** e por **conta.**

PALAVRA RECONHECIDA	CATEGORIA SINTÁTICA CORRESPONDENTE
REALIZADO	P 1
JUNHO	P 3
NORTE	F 1
MÃO DE OBRA	F 3
EXECUTANTE	N 1
CONTA	N 1

(h) GRAMÁTICA DE DETERMINAÇÃO (NUCLEAR)

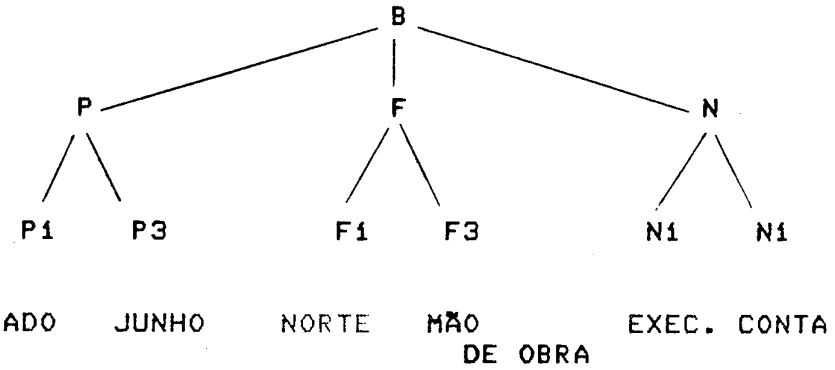
B -> F P N

F -> F1 F3

P -> P1 P3

N -> N1
N -> N N1

(i) ÁRVORE CORRESPONDENTE À ANÁLISE BÁSICA DE (G)



A observação detalhada da árvore (i) poderia levar a supor que as **GRAMÁTICAS DE DETERMINAÇÃO** não passam de uma sofisticação sintática para o polêmico processo de análise por **palavra-chave**. Para afastar tal suposição, vejam-se as consultas abaixo, corretamente interpretadas pelo sistema **FARAO**, mencionado anteriormente, nas quais fica clara a maior potencialidade das GD's no que concerne os fenômenos da língua por ela tratados. Antes de cada exemplo será mencionado o caso a que ele se refere, e as palavras que foram necessárias para o processamento estão sublinhadas.

(j) COORDENAÇÃO SENTENCIAL E SUB-SENTENCIAL

Quero saber o orçado e o realizado anual do DED e do DIE e do DIE quero saber apenas o realizado em junho.

(k) ANÁFORAS COM PRONOMES PESSOAIS RETOS

Diga qual é o realizado em junho da região sul e do Distrito de São Paulo com a receita operacional de telex e diga depois qual era o orçado deles na conta de facilidades.

(l) RECUPERAÇÃO DE LISTAS ORDENADAS/QUANTIFICADAS POR PRONOMES INDEFINIDOS E/OU NUMERAIS ORDINAIS/CARDINAIS

Diga o realizado dos distritos de Manaus, Belém, Boa Vista e Rio Branco nos meses de abril e maio e depois dê o orçado dos três últimos nos mesmos meses.

(m) NEGAÇÃO E EXCLUSÃO

Quero saber o valor orçado de todos os executantes da área nacional, exceto o DRB, para a receita operacional de junho.

(n) ELIPSES

Qual o realizado dos distritos da região centro-leste por mês e qual o orçado por conta?

Vários outros casos, de origem gramatical ou não (i.e. particularidades da base de dados consultada) são corretamente processados. Entretanto, os exemplos acima já são suficientes para demonstrar que as GD's não se restringem a um mero tratamento por palavra-chave. Na seção a seguir serão enumeradas as características fundamentais desta metodologia, que garantem a devida interpretação de um grande volume de consultas em linguagem natural ou pseudo-natural a uma base de dados.

3 - GRAMÁTICAS DE DETERMINAÇÃO

Para obter os resultados indicados na seção anterior, as **gramáticas de determinação** incorporam uma série de dispositivos e estratégias, que serão descritos a seguir. Dentre eles figuram a **leitura seletiva da entrada** e uma interpretação semântica próxima de uma **tradução para comando de consulta**, que não são propriamente parte do formalismo sintático, mas que são pressupostas por ele.

A **leitura seletiva da entrada** (ou consulta) é o que assegura a flexibilidade da expressão do usuário. Nem todas as palavras por ele empregadas são **reconhecidas** pelo sistema interpretador. Apenas aquelas que correspondam a uma informação semântica relevante ou a um vocábulo da língua portuguesa que sinalize processos sintáticos tratados (e.g. pronomes, conjunções e partículas diversas) serão reconhecidas à primeira varredura da consulta pelo módulo de análise lexical. Exemplos da seção 2 mostram, em consultas possíveis, quais palavras são "vistas".

O interpretador semântico orientado para uma tradução da consulta em linguagem (pseudo)natural para uma **query language** é fortemente inspirado pela proposta dos compiladores de linguagens de programação. Ressalta-se aqui, igualmente, a distinção entre a consulta a uma **base de dados** (estática) e a consulta a uma **base de conhecimentos** (dinâmica). Neste último caso, visto se tratar de **sistemas inteligentes**, a base é ampliada pela informação contida nas consultas (através de inferências). Assim, o nível de análise fina do interpretador **precisa ser** mais sofisticado do que na consulta a bases de dados sem pretensões inferenciais complexas.

Concentrando-se no aparato **sintático**, as GD's estão aliadas a um processo de parsing onde figuram os seguintes componentes: (a) **regras** e (b) **meta-regras.**

REGRAS

As regras de uma gramática de determinação podem ser **declarativas ou programáticas.** Quando declarativas, são **regras de reescritura**, que refletem uma análise **bottom-up** da entrada, no padrão:

(j) A B C -> X Y Z

onde A, B e C, X, Y e Z são constituintes sintáticos, **terminais ou não-terminais**, restringidos por condições específicas que subcategorizam as regras em (1) **livres de contexto**, (2) **sensíveis a contexto** e (3) **transformativas.** Quando programáticas, as regras efetuam modificações drásticas na estrutura em análise (e.g. **cópia de constituintes**) ou decidem ambiguidades entre elementos não-contínuos da mesma.

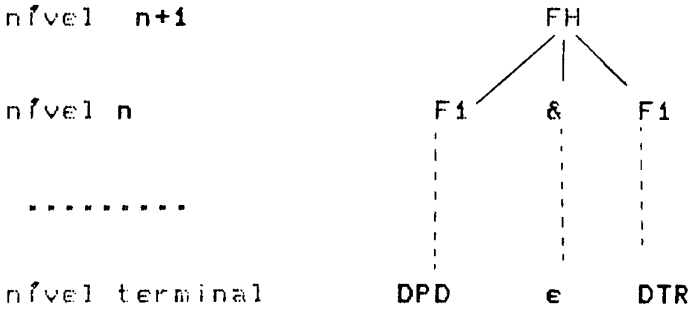
REGRAS LIVRES DE CONTEXTO

A nomenclatura adotada pela GD é mais alusória do que definitória, no sentido de que expressões como **livre de contexto** ou **sensível a contexto** não refletem exatamente os preceitos a elas associados na teoria das linguagens formais. Por exemplo, o **sentido** da leitura de uma regra declarativa da GD é **o inverso** do sentido normalmente utilizado (gerativo). Como a análise da entrada é **bottom-up**, as regras se enunciam **dos terminais para a raiz**, e não da raiz para os terminais.

O padrão de regra declarativa enunciado em (j), quando apresenta a restrição de que **à direita da seta só se pode ter um símbolo e de hierarquia superior aos símbolos da esquerda**, especifica uma **regra livre de contexto.** Em outras palavras, as regras livres de contexto **agregam** constituintes de nível n sob um único constituinte de nível **n+1.** A operação destas regras reflete um movimento vertical na árvore representativa da análise.

EXEMPLO DE REGRA LIVRE DE CONTEXTO

F1 & F1 -> FH



REGRAS SENSÍVEIS A CONTEXTO

Quando o padrão das regras declarativas (j) apresenta a restrição de o **módulo da cadeia à esquerda da seta ter de ser igual ao módulo da cadeia à direita** (i.e. o mesmo número de constituintes tem de aparecer de um e de outro lado da seta) **e apenas um dos constituintes à esquerda é reescrito à direita,** trata-se de uma regra sensível a contexto. Estas regras **não efetuam agregação,** e sim uma mudança de classe, o que significa que os constituintes de nível **n,** à esquerda da seta, permanecem no mesmo nível **n,** após a reescritura de um deles, no contexto especificado pelos demais. O movimento relativo a estas regras, na árvore representativa da análise da entrada, é horizontal.

Para ilustrar o efeito de uma regra sensível a contexto, seja o seguinte:

- os **numerais** 4 e 5 têm classe sintática Q3, significando **quantificação atribuída.**

- as palavras **divisão** ou **divisões** têm classe N2, arbitrariamente correspondendo a um **nome de campo.**

Claramente, as construções sintáticas (k) e (l), abaixo, representam um caso em que o **contexto** desambigua a classificação dos numerais, resolvendo se eles são (1) a **quantidade de divisões** ou (2) a **designação das divisões.**

(k) ... 4 ou 5 divisões têm orçamento controlado ...

(l) ... a divisão 4 ou 5, do DPD, tem orçamento controlado ...

As regras sensíveis a contexto capazes de solucionar este problema são (m) e (n) abaixo.

(m) ND Q3 -> ND Fd (porque **4** é um **fornecido** transformado)

(n) Fd **ou** Q3 -> Fd **ou** Fd (para **propagar** a transformação para todos os Q3 ambíguos que houver)

Este exemplo ilustra também um dispositivo muito útil das GD's que são as **classes transientes.** Estas classes são manipulações simbólicas **em contexto,** nas quais uma determinada classe assume **temporariamente** uma outra denominação, a fim de que determinadas reescrituras possam ser feitas. No caso, quando **4** e **5** são designações de divisões, a classe final adequada é F1, mas reescreve-se temporariamente como Fd para que se possa propagar a transformação de Q3 em F1 a todos os elementos coordenados de uma cadeia. Uma regra **sensível a contexto** final transforma todos os Fd em F1, para normalizar a análise.

REGRAS TRANSFORMATIVAS

As regras transformativas das GD's operam, como o nome já o diz, transformações estruturais. O contexto geral de reescritura (j) é restringido de maneira especial, conforme se trate de uma ou outra transformação.

Para as transformações de **apagamento,** a restrição de (j) é que **o lado esquerdo da seta tenha módulo maior ou igual a 2** e que **o lado direito da seta seja menor que o módulo do lado esquerdo.** Para esclarecer o motivo e o funcionamento das regras transformativas de apagamento, seja a consulta (o), abaixo.

(o) Diga o orçamento da **região norte** para o **mês de agosto.**

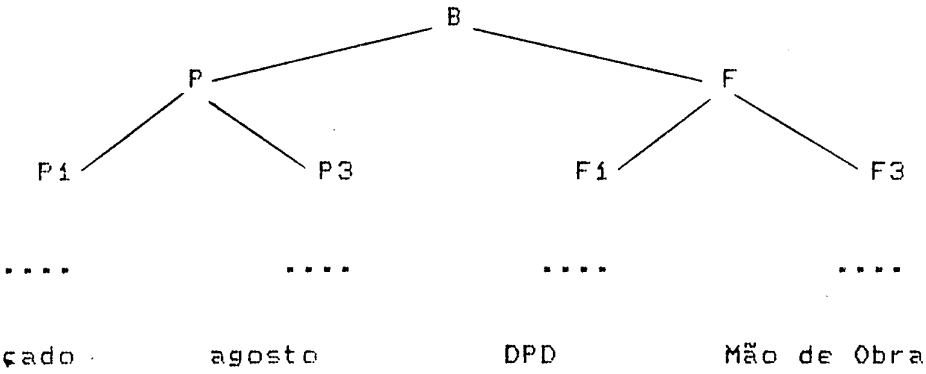
Observe-se que **norte** só pode ser região e que **agosto** só pode ser mês. Isto revela uma **redundância** entre os dois componentes das expressões em negrito de (o). Uma vez detectada esta duplicação, pode-se perfeitamente dispensar as palavras **região** e **mês**. Isto se faz através da regra **transformativa de apagamento** (p), a seguir, que é um caso particular de regra sensível a contexto, isto é: não opera na vertical, mas na horizontal, relativamente à árvore representativa da estrutura de (o).

(p) Nr Fr -> Fr (pois a **semântica** de **norte** já presume **região**)
Nm Fm -> Fm (por motivo semelhante ao caso anterior)

Portanto, todo **N? F? -> F?**, seja ? o símbolo que for.

A outra regra transformativa prevista para uma GD é a de **permutação**, em que a ordem dos elementos constantes do lado esquerdo da seta é alterada do lado direito. O contexto geral (j) fica restringido da seguinte forma: **para a regra de permuta, o lado esquerdo da seta tem de ter módulo maior ou igual a 3 e o módulo do lado direito da seta tem de ser igual ao do lado esquerdo (e constem as mesmas classes em amboso os lados).** Trata-se, novamente, de uma subcategorização de regra sensível a contexto. As **permutas** têm a função de **normalizar** estruturas para reescrituras em estágio avançado de análise. Se se lembrar que o objetivo final da estruturação da consulta é atingir algo como (q), abaixo,

(q)



pode acontecer de estes elementos terem aparecido na consulta em **ordem** diferente daquela que consta em (q). Veja-se (q'):

(q') Quero o **orçado** do **DPD** para **mão de obra** em **agosto.**

A certa altura da análise, o que se terá é algo como:

- P1 (Orçado) F (F1 (DPD) F3 (Mão de Obra)) P3(Agosto)

Para que se possa agregar P1 e P3 sob P, pela regra livre de contexto P1 P3 -> P, é preciso que se **permutem** os elementos da consulta (q'), o que é feito através da regra (r) abaixo.

(r) P1 F P3 -> P1 P3 F (note-se que, sendo sensível a contexto, esta regra **não agrega elementos**, isto é, opera na **horizontal** da árvore)

As regras **programáticas** da GD são **procedimentos** especiais para mudanças muito drásticas na estrutura em análise. Pode-se exemplificá-las através de dois casos típicos: **cópias** e **apagamentos incondicionais.**

Os casos de **cópia** estão envolvidos em todas as **anáforas**, onde um pronome está ocupando o lugar de um elemento já mencionado na sentença (ou no contexto geral do diálogo). Contudo, a estrutura final deve apresentar não o pronome, mas o elemento em questão, reproduzido. Por exemplo:

(s) Quero o realizado da **região sul** e do **distrito de São Paulo** e, depois, o orçado **deste último** para o **telex nacional.**

A expressão **deste último**, no processo de resolução da anáfora refere-se ao **distrito de São Paulo**, que deverá constar da árvore final, **duplicado** (e não pronominalizado). Ora, para isto é preciso **copiar** o constituinte **distrito de São Paulo**, o que se faz através de uma **regra programática** (ou procedure).

Outra utilização importante das regras programáticas é o **apagamento incondicional.** Para exemplificá-lo, considere-se o emprego das vírgulas (,) na língua portuguesa. De maneira geral, ao escrever, muitas pessoas empregam erroneamente as **vírgulas.** Devido a este problema, elas não podem ser **confiáveis**, e, consequentemente, não precisam ser "vistas" pelo processador na maior parte do processo de **parsing.** Entretanto, em construções coordenadas (entre outras), elas desempenham um papel importante. Veja-se o grupo (t) de regras referentes à coordenação (recursiva) de F1's sob um nódulo FH.

(t) F1 e F1 -> FH

F1 , FH -> FH

O grupo (t) processa estruturas como: "DPD, DTR, DTF e DCD", onde cada elemento tem classe F1.

Passado este ponto, as **vírgulas** da consulta não têm mais utilidade para a análise sintática proposta. Neste caso, uma regra **programática** apaga todas as vírgulas da sentença e a estrutura de trabalho do analisador fica descarregada de elementos espúrios para as etapas seguintes.

META-REGRAS

As meta-regras das GD's são o **controle** do processo de análise. As regras citadas até este momento são divididas em **ciclos** ou **blocos**, que se aplicam **sequencialmente** em um mesmo estágio. Cada bloco contém regras que se encontram **ordenadas** entre si, para que a aplicação de uma não impeça a aplicação de outra, a seguir. As meta-regras são regras que **chamam** os blocos de regras declarativas e programáticas, possibilitando - inclusive - que determinado ciclo seja re-aplicado, por exemplo, depois de uma operação de cópia de constituinte (o que é indispensável).

Esta estratégia tem a vantagem de impedir que todas as regras da base de conhecimento do sistema sejam tentadas a cada passo da análise. Assim, as regras em que só constam elementos altamente agregados (i.e. perto da raiz da árvore) são carregadas apenas em estágio compatível com esta exigência. Se os elementos no estágio corrente são terminais, a regra final P F -> B não é chamada por qualquer meta-regra.

O efeito de controle das meta-regras pode ser graficamente explicado pela figura a seguir.

META-REGRA 1	META-REGRA 2	META-REGRA 3
A1 B1 C1	B2 C2 A2 C2	B3 A3 B3 C3
CICLO A1 regra1 regra2 ... regran	CICLO A2 regra1 regra2 ... regran	CICLO A3 regra1 regra2 ... regran
CICLO B1 regra1 regra2 ... regran	CICLO B2 regra1 regra2 ... regran	CICLO B3 regra1 regra2 ... regran
CICLO C1 regra1 regra2 ... regran	CICLO C2 regra1 regra2 ... regran	CICLO C3 regra1 regra2 ... regran

Observe-se que uma vez chamados os ciclos, as regras que deles constam são lidas sequencialmente. Também, as meta-regras são chamadas sequencialmente, ou seja: primeiro a meta-regra1, que é toda processada, depois a 2 e a 3.

4 - POTENCIALIDADE DAS GRAMÁTICAS DE DETERMINAÇÃO

As Gramáticas de Determinação apresentam algumas vantagens bastante interessantes para a aplicação de interface a bases de dados. Dentre elas destacam-se as que se discutem a seguir.

FLEXIBILIDADE EXPRESSIVA

O principal atrativo das GD's é a **flexibilidade** que o **usuário** tem na expressão de suas consultas. A leitura seletiva da entrada, aliada ao fenômeno de sintatização de categorias semânticas, permite que consultas **em linguagem corrente, em linguagem telegrafada, em linguagem semi-formal** e, por conseguinte, em boa parcela de casos de **agramaticalidade** sejam devidamente interpretadas, sem que se obrigue o usuário a uma expressão **correta** ou **treinada.**

EFICIÊNCIA ANALÍTICA

O segundo grande atrativo das GD's é a utilização de categorias sintáticas **que são e enquanto são** relevantes em termos interpretativos. Isto pode ser percebido em pelo menos duas etapas distintas: inicialmente na leitura seletiva da entrada e, em seguida, nas regras de apagamento. Estes recursos contribuem para que a estrutura final que será entregue ao interpretador seja **limpa e desbastada.** Certamente esta característica contribui para aumentar a eficiência **representativa** da estrutura sintática de trabalho e a eficiência do **processo de interpretação,** visto que praticamente não há elementos de semântica vazia chegando ao processador do significado da consulta (o que pode ocorrer em outras abordagens alternativas, sobretudo naquelas que seguem propostas tradicionais para a gramática do português).

FERRAMENTA METODOLÓGICA

Outro ponto interessante das Gramáticas de Determinação é o fato de que - considerando que elas foram desenvolvidas para a aplicação de consultas a bases de dados - suas categorias sintáticas apresentam uma correspondência com a modelagem da base de dados utilizada. Por exemplo, os **valores** assumidos pelos campos da base têm de apresentar uma classe global **F,** ao passo que os nomes de campos tenderiam a ter uma classe global **P.** Os atributos tenderiam a ter função determinante **3,** ao passo que as entidades deveriam ter função determinada **1,** em seus respectivos constituintes. Esta possibilidade tem motivado o desenvolvimento de pesquisas sobre as GD's, no sentido de se conseguir automatizar o mapeamento da modelagem de dados para uma **proto-gramática** GD, na qual estejam constando as macro-relações de determinação para as consultas a serem processadas.

ELASTICIDADE

Por elasticidade entende-se, aqui, a capacidade de a GD expandir-se em sentido vertical - isto é, tornando-se mais profunda e perspícua, para perceber fenômenos da língua de maior sutileza - e em sentido horizontal - isto é, tornando-se aplicável a várias bases de dados, o que remete ao fator transportabilidade. No que tange à perspicuidade, uma Gramática de Determinação pode ser tão profunda e detalhada quanto se queira, não havendo - nos mecanismos utilizados - um indício evidente de que determinadas construções esperadas na consulta em linguagem natural a bases de dados não possam ser compreendidas. A manipulação simbólica facultada pelas regras de reescritura não parece restringir **a priori** o tratamento de fenômenos de pressuposição, por exemplo, marcados por quantificadores ou outras categorias. Já no que tange à transportabilidade, a dependência de domínio das interfaces em linguagem natural é uma imposição de ordem prática. Entretanto, a separação das regras em ciclos pode perfeitamente isolar fenômenos linguísticos supra-ambientais (coordenação, cópias, permutas e outros) que estes são transportáveis (porque são essencialmente sintáticos).

A desvantagem que apresentam as GD's em seu atual estágio de desenvolvimento é a sua complexidade. O número de regras é grande e a novidade das categorias sintáticas com que elas trabalham pode dificultar bastante a manutenção de uma base de conhecimentos linguísticos. O próprio processo de manipulação simbólica por reescritura não é amplamente dominado e pode representar um obstáculo para o desenvolvimento de interfaces para BD's através desta metodologia. Entretanto, como já se disse, há pesquisas sendo realizadas para minorar este problema, possibilitando que se tirem todas as vantagens das GD's naquilo que elas apresentam de positivo e atraente.

5 - CONCLUSÕES

O campo de processamento de linguagem natural, no Brasil, embora interdisciplinar por definição, não tem contado com o desenvolvimento de metodologias efetivamente **interdisciplinares**. A maioria dos trabalhos se desenvolve apenas no campo da informática, sem a participação de lingüistas. Se é verdade que todo cientista computacional brasileiro fala português, não é verdade que todo lingüista possa desenvolver sistemas computacionais. Isto encoraja pesquisas unilateralmente realizadas e - obviamente - o lado produtivo não é o da linguística.

As **GRAMÁTICAS DE DETERMINAÇÃO** foram desenvolvidas por um especialista em lingüística, há alguns anos participando ativamente em pesquisas na área de inteligência artificial. A leitura detalhada de sua bibliografia básica [Souza,1988] mostra claramente a tendência do trabalho. Contudo, o objetivo de seu

desenvolvimento é o de sensibilizar os pesquisadores do processamento de linguagem natural aplicada ao português para a necessidade da criação de um novo espaço de investigações conjuntas, onde se possam contemplar as especificidades do processador (computador) e da linguagem (semi-natural, uma vez que atualmente digitada), obtendo-se interpretadores mais eficientes em termos computacionais e flexíveis em termos expressivos para a atividade de consulta a bases de dados.

REFERÊNCIAS BIBLIOGRÁFICAS

BURTON,R. **Semantic Grammar: an Engineering Technique for Constructing Natural Language Understanding Systems**, S.L., S.P.,1976.

COELHO,H. **Sobre a Engenharia da Linguagem**, in Anais do XIV Congresso Nacional de Informática, SUCESU, São Paulo, 1981.

GARCIA,A.,CARVALHO,E.,FORTUNY,J. **Aspectos da Implementação de um Processador de Linguagem Natural em PASCAL**, in Anais do III Simpósio Brasileiro de Inteligência Artificial, SBC, Rio de Janeiro, Novembro de 1986.

HENDRIX,G.,SACERDOTTI,E. et alii **Developing a Natural Language Interface to Complex Data**, in Transactions of the ACM on Data Base Systems, 3: 105-147, Junho,1978.

SOUZA,C. **Consulta a Bases de Dados em Português Corrente**, in Anais da X Reunião Anual do ADAGRUPO, Águas de Lindóia, Abril, 1987.

_____ **Gramáticas de Determinação: uma Ferramenta para o Processamento da Língua Portuguesa**, Tese de Doutorado, Depto. de Letras, PUC/RJ, Janeiro, 1988. (Versão Final)

WALTZ,D. **English Language Question-Answering System for a Large Relational Data Base**, in Communications of the ACM 21:526-536, Julho, 1978.