MINISTÉRIO DA CIÊNCIA E TECNOLOGIA
INSTITUTO DE PESQUISAS ESPACIAIS

AUTORIZAÇÃO PARA PUBLICAÇÃO
AUTHORIZATION FOR PUBLICATION

PALAVRAS CHAVES/KEY WORDS (AUTORES/AUTHORS)

TEXTO - RESUMO - ÍNDICE - LINGÜÍSTICA
INTELIGÊNCIA ARTIFICIAL - LINGUAGEM
NATURAL

AUTORIZADA POR/AUTHORIZED BY

Luiz Gylvan M. Filho

AUTOR RESPONSAVEL
RESPONSIBLE AUTHOR

Carlos A. Oliveira

DISTRIBUIÇÃO/DISTRIBUTION

- [ ] INTERNA / INTERNAL
- [x] EXTERNA / EXTERNAL
- [ ] RESTRITA / RESTRICTED

REVISADA POR / REVISED BY

Marciana L. Ribeiro

CDU/UDC

681.3.019

DATA / DATE

Novembro 1990

PUBLICAÇÃO Nº
PUBLICATION NO

INPE-5183-PRE/1651

ORIGEM
ORIGIN

NCO/OBT

PROJETO
PROJECT

BWIPS

Nº DE PAG. NO OF PAGES	ULTIMA PAG. LAST PAGE
10	10

VERSÃO VERSION	Nº DE MAPAS NO OF MAPS

TÍTULO/TITLE

PROPOSTA PARA UM GERADOR DE (PRÉ) ÍNDICE E DE (PRÉ) RESUMOS, A PARTIR DE UMA INTERPRETAÇÃO LINGÜÍSTICO - PRAGMÁTICA

AUTORES/AUTHORSHIP

Carlos Alberto de Oliveira

RESUMO - NOTAS/ABSTRACT - NOTES

Propõe-se, neste trabalho, as bases para a elaboração de um gerador automático de (pré) índices e de (pré) resumos, a partir de uma abordagem lingüístico-pragmática. Sucintamente, o critério utilizado para a discriminação das "palavras" e/ou expressões que comporão índices e resumos será o do número maior de ocorrências dessas no texto-fonte, aliado a uma heurística relacionada ao nível de profundidade da ocorrência dentro de uma representação de conhecimento lingüística. Para tanto, utilizar-se-á um protótipo de processamento de linguagem natural já desenvolvido e baseado em conhecimento lingüístico textual.

OBSERVAÇÕES/REMARKS

Aceito para ser publicado nos Anais do IV Seminário de Automação em Bibliotecas e Centros de Documentação, de 03 a 06 de dezembro, 1990, na USP, São Paulo, SP.

PROPOSTA PARA UM GERADOR DE (PRÉ)ÍNDICES E DE (PRÉ)RESUMOS, A PARTIR DE UMA INTERPRETAÇÃO LINGÜÍSTICO-PRAGMÁTICA

Carlos Alberto de Oliveira

Instituto de Pesquisas Espaciais - INPE
Secretaria de Ciência e Tecnologia
Caixa Postal 515 - 12201 - São José dos Campos - SP, Brasil

RESUMO

Propõe-se, neste trabalho, as bases para a elaboração de um gerador automático de (pré)índices e de (pré)resumos, a partir de uma abordagem lingüístico-pragmática. Sucintamente, o critério utilizado para a discriminação das "palavras" e/ou expressões que comporão índices e resumos será o do número maior de ocorrências dessas no texto-fonte, aliado a uma heurística relacionada ao nível de profundidade da ocorrência dentro de uma representação de conhecimento lingüística. Para tanto, utilizar-se-á um protótipo de processamento de linguagem natural ja desenvolvido e baseado em conhecimento lingüístico textual.

ABSTRACT

The basis of building a preliminar index and resume automatic generator through a linguistic-pragmatic approach is proposed. Words and phrases will be selected by using the number of events on source text criterion. Also, this number will be related with the deep level of the event on a linguistic knowledge representation. A developed text linguistic knowledge-based prototype of natural language processor will be used to realize it.

1. INTRODUÇÃO

Considerando a grande velocidade com que informações são geradas e a quantidade exponencial de informações em documentos dos mais variados formatos e especificações, excedendo em muito a capacidade humana de manipulação eficiente e rápida dos mesmos, é um truísmo observar a necessidade atual de um tratamento automático ou automatizado como braço auxiliar em centros de documentação.

Some-se a isso o fato de que, ao se editar um texto escrito, é desejável que se possa fazer uso de algum processo automatizado: 1) para correção de "erros" (sejam eles, de formatação, ortográficos ou lingüísticos propriamente ditos); 2) para gerar índices; 3) para gerar resumos. Os três ítens citados, além de servirem como apoio ao usuário-editor, são de grande valia para o serviço bibliotecário, no que concerne, entre outros fatores, à economia e à uniformidade.

Nesse contexto, a área de Processamento de Linguagem Natural (PLN), inscrita na de Inteligência Artificial, pode oferecer subsídos importantes para a execução de tal tipo de tarefa.

Deixando de lado a correção de "erros", pois já existem no mercado aplicativos e/ou editores de texto que, de alguma forma, executam tal tarefa, realçam-se os ítens "geração de índices" e "geração de resumos" como os que podem se valer dos métodos e técnicas da área de PLN, no que concerne ao processo de automação em bibliotecas e centros de documentação.

No Instituto de Pesquisas Espaciais desenvolveu-se um protótipo (1) de processador de linguagem natural (LN) que, pelo seu arcabouço teórico e por suas características de flexibilidade e transportabilidade, permite ser adaptado para fins de geração de um (pré)índice e de um elementar (pré)resumo. Os fundamentos teóricos estão centrados na Lingüística de Texto (van DIJK, 1977; Marcuschi, 1983) e já foram exaustivamente discutidos em Oliveira (1986a, 1986b, 1987, 1988, 1989) e Rosado e Lopes (1988). A implementação foi centrada sobre conhecimento exclusivamente lingüístico (Lingua Portuguesa do Brasil) o que enseja, através da inclusão de métodos estatísticos (Siqueira e Costa, 1987), a depreensão, seleção e hierarquização de possíveis descritores de dado texto, cabendo a um bibliotecário genérico fazer a depuração do resultado final. Com relação a geração de (pré)resumos, embora o trabalho seja mais intenso e problemático, já existem estudos (Schank and Abelson, 1977) e aplicações (Scott and Souza, 1989) que permitem uma modelagem satisfatória.

(1) Conforme tese de doutorado "IDEAL, uma Interface Dialógica em linguagEm naturAL para sistemas especialistas" de Carlos Alberto de Oliveira.

2. PARA UMA ABORDAGEM LINGÜÍSTICO-PRAGMÁTICA

Primeiramente, por que este trabalho transita por, no mínimo, duas diferentes áreas de conhecimento, a decodificação terminológica está vinculada, no que concerne à Documentação, a Robredo (1978) e, no que concerne à Lingüística, a Dubois et alii (1978).

Em segundo lugar, alguns pressupostos teóricos, no que concerne à área da Lingüística, devem ser considerados aqui para a melhor explicitação da proposta de um sistema gerador de (pré)índices e/ou (pré)resumos. Acrescente-se que o uso do prefixo "pré" para índices e resumos, vincula-se ao fato de que, neste caso, o resultado da análise lingüístico-pragmática não tem o intúito de ser o definitivo, mas sim, o de ser uma fase intermediária entre o que pode ser e o que efetivamente deve ser: uma espécie de esboço "bom". Necessitar-se-á sempre o concurso de bibliotecários para validar, corrigir desvios, eliminar incongruências, refinar deduções efetivadas, etc.

2.1. O PROCESSO TRADUTOLÓGICO

Os seres humanos comunicam-se através de sistemas de signos e, em especial, o sistema de signos lingüísticos. Este, o signo lingüístico, conforme Couto (1983), é uma realidade bifacial, isto é, consta de uma Expressão (dado concreto, sensorial) e de um Conteúdo (dado abstrato, conceitual), os quais estão em relação, conforme Figura 1.

Fig. 1 - A relação sígnica.
FONTE: Couto (1983), p. 31

Segue-se, pois, que a própria relação é um dado, ou seja, a significação. Através dessa relação é que se pode fazer o trânsito Expressão-Conteúdo, ou melhor, pode-se codificar/decodificar o signo, isto é, traduzí-lo (Figura 2).

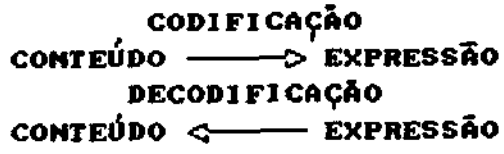

Fig. 2 - O processo de codificação/decodificação.

Nesse contexto, a tradução pode ser intra ou inter-códigos.

Tomando-se como exemplo a frase "se as condicões da zona frontal são favoráveis", indicadores morfológicos, lexicais, sintáticos e semânticos permitem depreender da frase que: a) pela taxe do SE, ele deve introduzir uma frase condicional; b) a relação entre AS CONDIÇOES e A ZONA FRONTAL intermediada pela preposição DE permite inferir que a segunda expressão é um "complemento" da primeira; as "terminações" +ÇÕES e +AL permitem inferir que as palavras as quais estão agregadas podem ser, respectivamente, a nominalização de um verbo e um adjetivo. Isto possibilita a transformação dessa frase em outras, tais como, " se a zona frontal se condiciona favoravelmente" ou "se a favorabilidade das condições da zona frontal é verdadeira", caracterizando-se assim a tradução intra-código, isto é, dentro da própria língua.

Já a Figura 3 dá uma visão da tradução inter-códigos: o elemento sol no domínio A (fenômemos meteorológicos, por exemplo) pode estabelecer uma relação de antonímia com chuva, ou seja, "não fazer sol implica chover", o que necessariamente nem sempre é verdade no mundo real; no domínio B (astronomia, por exemplo), a relação pode se estabelecer de maneira indireta com lua, ou seja, ambos são particularizações de uma classe de objetos; no domínio C (divisão do espaço de tempo 24 horas, por exemplo), as relações estabelecidas são de antonímia com lua e de sinonímia com dia. Pode-se notar que, conforme mudam-se os domínios, mudam-se também as relações que dão significação a dada "palavra".

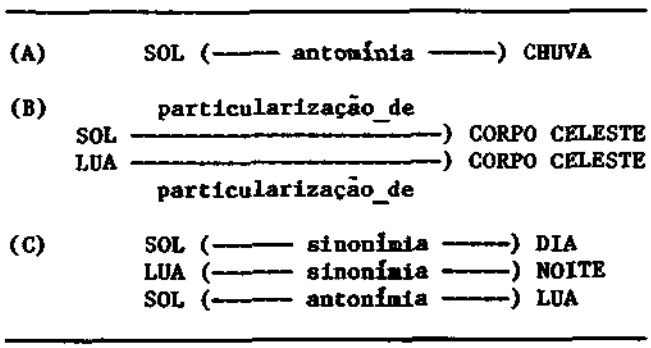

Fig. 3 - Possíveis relações da "palavra" sol em diferentes domínios.

Desde que as relações (significações) valem dentro do código no qual o signo está inserto (para cada domínio, um elemento pode encetar diferentes relações com os demais elementos desse mesmo domínio), a tradução inter-códigos, para se realizar, deve basear-se num conjunto de relações conhecidas que delimitem e permitam o trânsito codificação/decodificação inter-domínios.

2.2 - A LINGÜÍSTICA DE TEXTO

A Lingüística de Texto (LT) "é uma lingüística dos sentidos e processo cognitivos e não da organização pura e simples dos constituintes da frase" (Marcuschi, 1983, p. 14) e, por isso, será adotada com o objetivo principal de, tomando frases em LN: traduzir os conhecimentos nelas insertos para dada representação de conhecimento; verificar, por inferências próprias ou interação com o usuário (produtor de tais frases), se tais conhecimentos constituem um "tecido" coerente.

De uma forma geral, texto "consiste em qualquer passagem, falada ou escrita, que forma um todo significativo, independente de sua extensão. Trata-se pois de uma unidade de sentido, de um contínuo comunicativo contextual que se caracteriza pela coerência e pela coesão, conjunto de relações responsáveis pela tessitura do texto" (Fávero e Koch, 1983, p. 25). Indicam a coesão as maneiras como os elementos constituintes do universo textual estão ligados dentro da linearidade, isto é, numa espécie de semântica da Expressão, na qual se estudam as maneiras como são usados padrões formais na transmissão de conhecimentos e sentidos (Fávero, 1985, p. 148); já a coerência, resultado de processos cognitivos que se operam entre usuários, é indicada pelos conceitos e relações que subjazem a Expressão, isto é, situa-se no terreno do Conteúdo.

São princípios da LT aqui adotados os seguintes: fatores de coesão - os substituídores e dentre estes, a anáfora e a elipse; os seqüenciadores e dentre estes, o aspecto, a disjunção e a conjunção; fatores de coerência - algumas relações lógicas e modelos cognitivos que se fundamentam na necessidade de continuidade de sentidos de um texto. E esta continuidade expressa-se através de conceitos e relações. "A configuração destes, subjacente ao texto, constitui o universo textual, que pode não estar de acordo com a versão estabelecida do 'mundo real', visto que abrange toda a constelação de produção e recepção, o que faz com que o texto contenha muito mais do que a soma das expressões lingüísticas que o compõem. Ele incorpora (...) os conhecimentos e a experiência do dia-a-dia" (Koch, 1985, p. 159); fatores de pragmática - o usuário e as intencionalidade, informatividade e intertextualidade.

2.3 - CONSIDERAÇÕES SINTÁTICO-SEMÂNTICAS

O ser humano pode verbalizar a mesma informação de muitas formas diferentes. Assim: na Figura 4(A) temos o mesmo Conteúdo veiculado por duas Expressões que, tendo a mesma estruturação sintática, se diferenciam apenas pela forma de duas lexias (2), as quais têm a mesma função sintática e semântica; entre a Figura 4(A) e 4(B), o mesmo Conteúdo é veiculado por diferentes estruturações sintáticas e diferentes lexias.

Na Figura 4(C) tem-se o mesmo Conteúdo que difere na estruturação sintática apenas. Por fim, na Figura 4(D), o mesmo Conteúdo e diferentes estruturações e diferentes formas. Combinando-se os quatro exemplos, percebe-se o número de Expressões possíveis para veicular o mesmo Conteúdo, usando-se como lexemas (2) apenas NUV+, APROXIM+, FORM+ e EXIST+.

Nesse contexto, "Tomando a capacidade potencial de uma expressão de transmitir conhecimentos ou conteúdos, dizemos que o sentido é a realização atual desse conhecimento no texto. O texto é uma atualização seletiva de significações potenciais para possibilitar um só sentido.(...) O sentido deve manter uma continuidade, caso contrário o texto é incompreensível. Esta continuidade de sentido forma a coerência do texto e se expressa em conceitos e relações." (Marcuschi, 1983, p. 46).

(A)	**Há nuvens se aproximando** **Existem nuvens se aproximando**
(B)	**A distância das nuvens para o local de destino no instante anterior é maior que a distância entre nuvens e destino no instante atual**
(C)	**nuvens em aproximação** **aproximação de nuvens** **nuvens se aproximando**
(D)	**nuvens // formações nebulosas**

Fig. 4 - Diferentes Expressões de um mesmo Conteúdo.

Dessa forma, Expressões do(s) usuário(s) podem ser mapeadas para uma única ou diferentes estruturas cognitivas subjacentes.

---

(2) "Na terminologia de B. Pottier, a **lexia** é a unidade de comportamento do léxico. Opõe-se a **morfema**, menor signo lingüístico, e a **palavra**, unidade mínima construída. É, portanto, a unidade funcional significativa do discurso. A lexia simples pode ser uma palavra (...). A lexia composta pode conter várias palavras em via de integração ou integradas: quebra-gelo. A lexia complexa é uma seqüência estereotipada: a cavalo (...)". Ainda, serão adotados aqui os termos **lexema** (morfemas léxicos), pertencentes a inventários ilimitados e abertos (**léxico**), e **gramemas** (morfemas gramaticais). Os lexemas são unidades dependentes, necessitando do concurso do(s) gramema(s) para sua atualização. Outrossim, o lexema é provido de conteúdo sêmico. (Dubois et alii, 1978, p. 360-361)

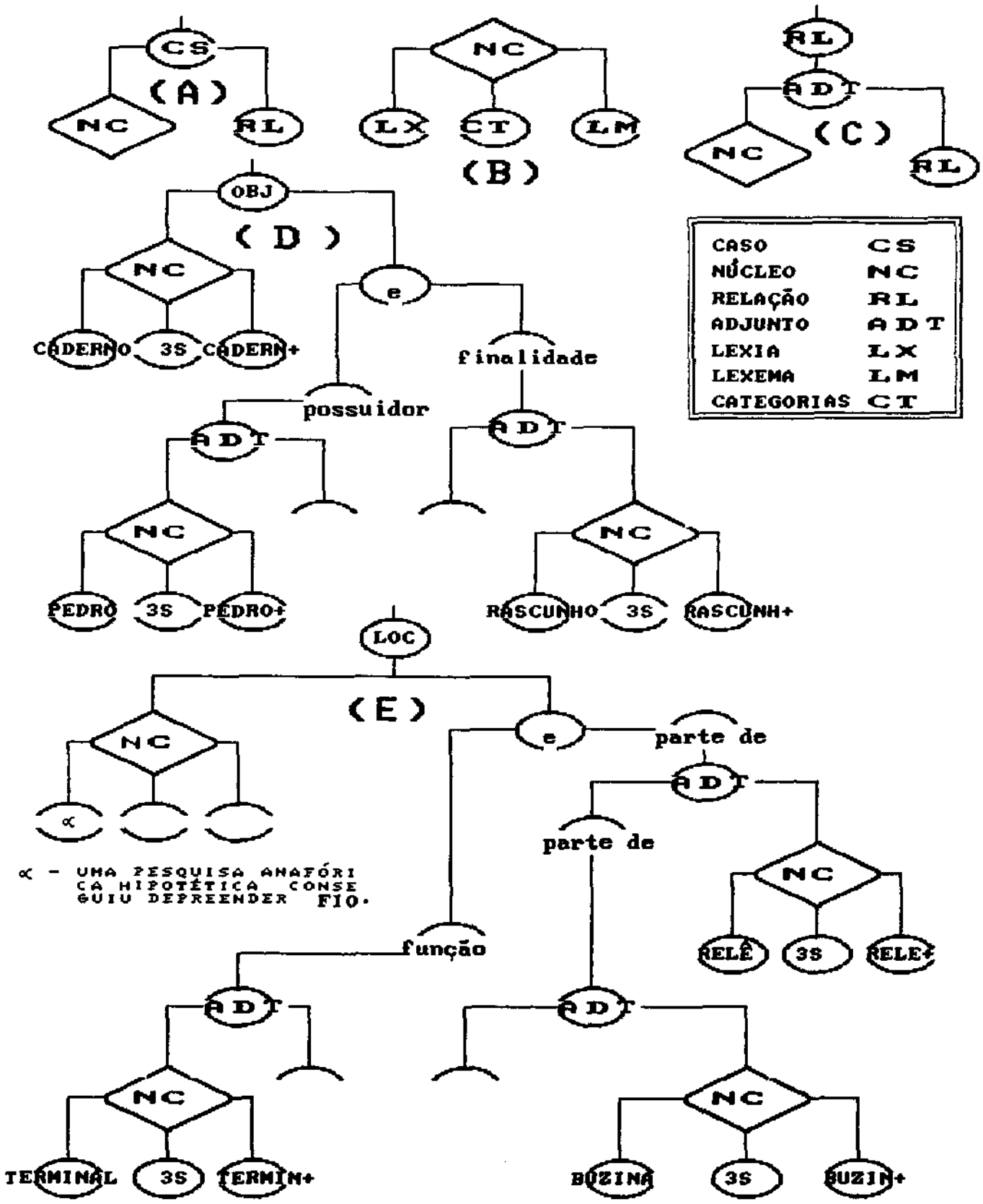

Fig.5 - Estruturas de representação intermediária de conhecimento lingüístico-pragmático.

Observando-se, então, que o usuário é um dos participantes na construção de um texto, nada mais importante que levá-lo em conta quando do processo de compreensão de LNs. Logo, frases em LN devem, por isso, serem tratadas como componentes de um texto, o qual se constrói na interação, especialmente quando advirem dúvidas sobre qual Conteúdo se fundamenta dada Expressão.

2.4 - O PROCESSO ANALÍTICO-INTERPRETATIVO

Torna-se claro, a partir dos pressupostos comentados, que é factível, através de uma abordagem exclusivamente lingüístico-pragmática, depreender de um texto escrito o "tecido cognitivo" que o gerou.

O conhecimento que subjaz ao sistema-protótipo (1) de processamento de LN são aqueles inerentes a LPB (regras morfossintáticas, regras semânticas, gramemas (2) e uma gramática de casos). Incluindo-se ainda o conhecimento pertinente à condução de objetivos dialogais. Este é necessário porque, sendo o usuário uma fonte de conhecimento integrante do sistema, torna-se exigível um processo interativo bem estruturado.

Não há, a "priori", qualquer outro conhecimento agregado ao sistema: durante o processo de análise-interpretação é que o conhecimento lingüístico-pragmático do domínio será incorporado ao conhecimento inicial.

Assim, é através do processo interativo com o usuário que cada "palavra", cada frase, cada expressão adquirirá somente uma interpretação, já que é ele (o usuário) quem, em última análise, determinará as significações do domínio.

Após o término do processo de análise-interpretação de cada frase, serão exibidas interpretações possíveis para a mesma até aquele momento. Aceita uma delas, (re)compor-se-á o "tecido cognitivo" até que todo o texto tenha sido processado. Adotou-se uma estrutura de representação intermediária que permite guardar as relações significativas que vigem no domínio: esta representação pode ser vista na Figura 5.

Na Figura 5(A) tem-se o caso com suas duas únicas ramificações: o núcleo e a relação. O núcleo é o centro do caso, a sua parte essencial (sua ocorrência é sempre no nível mais alto da representação). A relação conecta ao caso outros elementos do segmento da frase que não núcleos e estabelece significação. Na Figura 5(B), o núcleo comporta a lexia em análise, suas categorias e o lexema pertinente. Já a relação pode ser vazia (só existe o núcleo no caso), pode agregar um adjunto, ou ainda, pode ser uma conjunção/disjunção de relações. O adjunto, como na Figura 5(C), comporta um núcleo e uma relação.

Na Figura 5(D), representa-se "o caderno de rascunho de Pedro". Percebe-se que tal representação será diferente para "o terminal do relê da buzina", como na Figura 5(E). Nesta, inclusive, se estabelece uma necessidade de pesquisa anafórica, visto o núcleo da estrutura não estar preenchido (elíptico).

Observe-se que esta forma de representar intermediariamente o conhecimento expresso em LN permite: a) uma tradução intra-código, ou seja, dar nova estrutura superficial à frase de entrada, o que é fundamental para a fase de geração de (pré)resumos; b) uma tradução inter-códigos, desde que seja conhecida a linguagem de descrição da nova forma de representação a ser gerada, ou seja, o (pré)índice.

3.- GERAÇÃO DE (PRÉ)ÍNDICES E DE (PRÉ)RESUMOS

Saliente-se que comentar-se-á aqui uma indexação profunda a partir do documento completo em linguagem natural livre.

De uma maneira geral, um usuário genérico poderá submeter seu texto editado ao gerador em pauta e este, então, interagindo, fornecer-lhe-á um (pré)índice e um (pré)resumo. Estes poderão subsidiar o usuário no refinamento do resumo oficial que faz parte do seu documento.

Porém, é num centro de documentação informatizado que este aplicativo mais será de valia. O texto em "disquete" ou fita magnética seria processado e a bibliotecária poderia, a partir de uma listagem dos resultados, decidir a melhor configuração do índice e decidir a melhor redação do resumo.

3.1 - O PROCESSO

No processo de análise-interpretação, a LT toma o "layout" do texto como um dado de conhecimento. Títulos, grifos, destaques e outros indicadores de "importância" são levados em conta na atribuição de pesos aos candidatos a integrantes de um índice ou resumo.

O processo interativo se incumbe de dirimir dúvidas que, se deixadas a cargo, apenas, do motor de inferências do sistema, onerariam por demais o processo. Além disso, a solução do problema pode estar vinculada a conhecimento extra-lingüístico ou a conhecimento que o sistema ainda não assimilou ou possui. Por exemplo, cite-se a frase "Recebi uma foto de Franca". Seria necessário um volume muito grande e bem estruturado de conhecimento de mundo para decidir pelas interpretações possíveis, o que talvez nem um falante-ouvinte-interlocutor da língua em questão poderia resolver. Somente o gerador de tal frase (o usuário) teria a chave para a solução.

"2.2.2 Classificação e seleção

Quando o bibliotecário ou o documentário colocam um pacote de fichas em ordem alfabética por autores, ou as colocam no lugar conveniente num fichário por assuntos, estão realizando uma operação de classificação.

Tentemos analisar estas operações, para entender como o computador realiza, seguindo as instruções que lhe dá o programa, diversas operações de classificação. Consideremos a ficha representada na Figura 2.13, na qual podem ser reconhecidos os elementos tradicionais da catalogação de um artigo de revista, isto é, o número de registro, os autores, o título do artigo, a referência bibliográfica (indicando o volume, o ano, o número e a primeira e última páginas).

Quando se trata de classificar as fichas por autores, o documentarista ou o bibliotecário consideram, de uma maneira quase inconsciente, a segunda linha da ficha, prescindindo dos demais elementos que nela figuram (título, referência). Quer dizer, realizam uma seleção da informação contida na ficha e, logo, procedem à classificação por autores.

Estas duas operações de seleção e classificação são fundamentais, já que a elas se reduzem praticamente, um grande número das operações que se realizam em computador no campo da informática documentária."

Fig. 6. Fragmento de texto.
FONTE: Robredo (1978, p. 31)

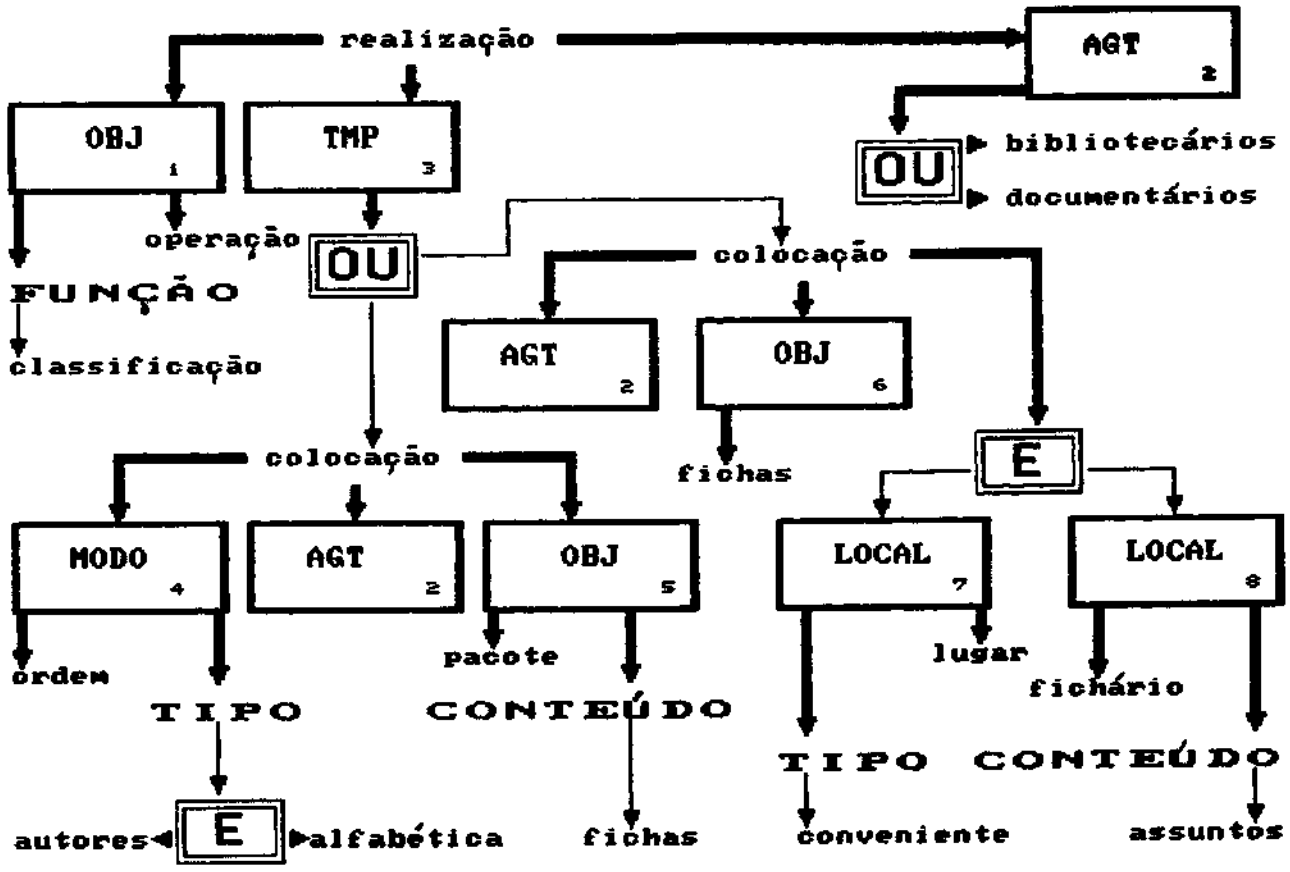

Fig. 7 - Uma análise do primeiro parágrafo da Figura 6.

Tomando-se, então, para efeito demonstrativo, o fragmento de texto da Figura 6, pode-se ter uma idéia de como funciona a análise lingüístico-pragmática.

O título, analisado como tal, já predispõe as duas lexias ali existentes em conjunção como fortes candidatas ao índice.

O primeiro parágrafo do texto é na realidade uma só frase a qual está subordinada uma segunda. ou seja, o gramema "quando", embora introduza a primeira frase a ser analisada, indica a existência de uma outra, em nível superior.

Dessa forma, o caso TMP (tempo) desdobra-se na disjunçao de duas frases: "o bibliotecário ou documentário ou [colocam um pacote ...] ou [colocam no lugar ...]".

Por regras semânticas (conhecimento lingüístico do sistema), as lexias "bibliotecário" e "documentário" são definidas, interativamente, como Agentes da ação-processo COLOCAR, nominalizada esta como "colocação". O Objeto da ação-processo é "um pacote de fichas". Aqui, a relação prepositiva, realizada oir "de", é questionada e o usuário responde, por hipótese, que "ficha é o conteúdo do pacote". A relação prepositiva introduzida por "em", questionada, insere o Modo como a ação-processo realiza o Objeto: "ordem do tipo por alfabeto e por autores".

A segunda frase da disjunção tem como Objeto "fichas". Esta lexia é depreendida da busca anafórica em "as colocam no lugar ...". As relações prepositivas seguintes definem, por interação, os LOCAIS "lugar do tipo conveniente" e "fichário cujo conteúdo é assunto" (3) .

A análise da frase principal depreende o Objeto "operação do tipo classificação". Notar que tanto a lexia "operação" quanto "classificação" são nominalizações dos verbos OPERAR e CLASSIFICAR, podendo-se, se assim fosse necessário, questionar a estrutura de tais ações: "quem classifica o quê?".

Ao término da análise desse primeiro parágrafo, cujo resultado se mostra na Figura 7, já se pode perceber duas ocorrências da lexia "ficha" em dois níveis diferentes: relacionada ao núcleo ("pacote") do Objeto 5 e como núcleo do Objeto 6. A lexia "classificação" também já ocorreu duas vezes: no título e relacionada ao núcleo ("operação") do Objeto 1.

O Agente, neste caso particular, raramente tem prioridade ou valor, sendo os Objetos os mais realçados.

O resto do texto segue, então, raciocínio análogo ao já discutido. Assim, após o término do processo o sistema já "apreendeu" conhecimento sobre o domínio, tal que poderá evitará futuros questionamentos redundantes e repetitivos.

3.2 - O RESULTADO

A Figura 8 representa um estágio intermediário dos resultados de uma análise do tipo em pauta sobre o texto da Figura 6. Pode-se observar que o número de ocorrências de dada lexia e/ou expressão está relacionada ao nível de profundidade (subordinação) que o elemento em análise tem na representação lingüística.

		número de ocorrências por níveis		
		N1	N2	N3
1	computador	2	-	-
2	operação classificação seleção	4	1	-
3	classificação operação de fichas tipo por autores	3	3	-
4	autores classificação de tipo ordem por	1	2	1
5	fichas ordem alfabética por autores local fichário pacote de conteúdo informação	2	2	-
6	seleção informação operação de	2	1	-

Fig. 8 - Exemplo de fase intermediária na geração de (pré)índice.

Nessa Figura 8, o peso final de cada item é uma média ponderada: no exemplo em pauta, têm-se apenas 3 níveis e foram atribuídos a eles os pesos 3, 2 e 1, em ordem crescente e respectivamente.

Aqueles ítens cujo valor ultrapassar um limite (média aritmética dos pesos finais considerando-se todas as lexias do texto) são

(3) Observe-se que a lexia "fichário" é composta pelo lexema "fich+", candidatando-se a ser questionada, se um refinamento da análise for necessário: "qual a relação entre ficha e fichário?". Nesse caso, ficha poderia ser o conteúdo de fichário e assunto o conteúdo de ficha.

candidatos em potencial para figurarem no índice. Observar que os ítens 2 e 6 recebem um acréscimo ao peso final em virtude de serem ambos um dado de conhecimento do texto: pertencem ao título.

Lexias que ocorrerem, no mínimo uma vez em, pelo menos, dois níveis são sempre candidatas. No entanto, só serão consideradas aquelas que tiveram ocorrências no nível de maior peso. No caso de tal não acontecer, o peso resultante do ítem será somado àquele item ao qual se agregar: é o caso, na Figura 8, de "ordem", "pacote" e "informação".

O resultado será então oferecido nos moldes da Figura 8 (sem o número de ocorrências e com referência às páginas) e o bibliotecário é quem determinará a configuração final do índice. Dessa intervenção sobre o resultado é que será gerado também o (pré)resumo.

No caso da Figura 8, tomando-se como hipótese que o bibliotecário a tenha aceito como está e levando-se em conta os ítens de maior peso e relacionados diretamente, poder-se-á gerar o seguinte (pré)resumo: "O computador realiza operações. As operações podem ser de classificação e de seleção. A classificação pode ser de fichas e por autores. A seleção pode ser de informação da ficha."

Deve-se ter em mente que a descrição do processo e dos resultados feita tão sucintamente pode deixar transparecer que tal seja trivial. Não o é e, ainda, o nível de complexidade é bem maior do que a primeira vista possa parecer. Cite-se, como exemplo, a determinação dos níveis de profundidade a serem levados em consideração e os pesos a serem atribuídos a tais níveis para que espelhem mais de perto a realidade do texto.

4. À GUISA DE CONCLUSÃO

Ao se propor e demonstrar a factibilidade de construção de um gerador de (pré)índices e de (pré)resumos, a partir de uma abordagem lingüístico-pragmática, objetiva-se, lançando-se mão de métodos e técnicas avançadas da Ciência da Computação (inteligência Artificial), subsidiar a automação em centros de documentação.

Ademais, o volume de conhecimento "apreendido" durante o processo de análise-interpretação pode vir a servir como subsídio para a construção futura de um gerador automatizado de thesaurus.

A expectativa atual é de tornar operacional um protótipo para a parte de geração de (pré)índices a médio prazo e para a parte de (pré)resumos a longo prazo. Espera-se com isso que, conseqüentemente, tempo maior seja deixado para que a capacidade humana nesses centros informatizados exercite o que lhe é mais próprio e inerente: a visão crítica e a decisão.

5. REFERÊNCIAS BIBLIOGRÁFICAS

COUTO, H. H. do. Uma introdução à semiótica. Rio de Janeiro, Presença, 1983.

van DIJK, T. A. Gramáticas textuais e estruturas narrativa. In: CHABROL, C. (ed.) Semiótica narrativa e textual. São Paulo, Cultrix, 1977. p. 196-229.

DUBOIS, J.; GIACOMO, M.; GUESPIN, L.; MARCELLESI, C.; MARCELLESI, J.-B.; MEVEL, J.-P. Dicionário de Lingüística. São Paulo, Cultrix, 1978.

FÁVERO, L. L. Mesa redonda: lingüística textual. Grupo de Estudos Lingüísticos, 10(1):146-152, 1985.

KOCH, I. G. V. Mesa redonda: lingüística textual. Grupo de Estudos Lingüísticos, 10(1):153-161, 1985.

MARCUSCHI, L. A. Lingüística de Texto: o que é e como se faz. (Dissertação de Mestrado em Lingüística) - Universidade Federal de Pernambuco, Recife, 1983. (Série Debates, 1).

OLIVEIRA, C. A. de. Uma proposta de classificação verbal como unidade geradora de texto. Grupo de Estudos Lingüísticos, 12(1):281-294, 1986a.

OLIVEIRA, C. A. de. Linguagem natural: um sistema de relações? In: CONGRESSO BRASILEIRO DE AUTOMÁTICA, 6., Belo Horizonte, 1986. Anais. Belo Horizonte, UFMG, 1986b. v. 1, p. 215-219.

OLIVEIRA, C. A. de. A morfologia e a sintaxe: um enfoque integrado baseado no conhecimento lingüístico. In: SIMPÓSIO BRASILEIRO DE INTELIGÊNCIA ARTIFICIAL, 4., Uberlândia, 1987. Anais. Uberlândia, UFUb, 1987. v. 1, p. 187-196.

OLIVEIRA, C. A. de. O tratamento automáticao de LN em processos de aquisicao de conhecimento. In: SIMPÓSIO BRASILEIRO DE INTELIGÊNCIA ARTIFICIAL, 5., Natal, 1988. Anais. Natal, UFRN, 1988, v. 1, p. 84-93.

OLIVEIRA, C. A. de. A sintaxe, a semântica e pragmática: um enfoque integrado baseado no conhecimento lingüístico textual. In: SIMPÓSIO BRASILEIRO DE INTELIGÊNCIA ARTIFICIAL, 6., Rio de Janeiro, 1989. Anais. Rio de Janeiro, PUCRJ, 1989. v. 1, p. 219-233.

ROBREDO, J. Documentação de hoje e de amanhã. Brasília, DF, Associação de Bibliotecários do Distrito Federal, 1978

ROSADO, P.; LOPES, G. Interfaces de língua natural com capacidade para aprenderem novos vocábulos, seu significado e para se adaptarem a novos utilizadores: uma experiência. In: SIMPÓSIO BRASILEIRO DE INTELIGÊNCIA ARTIFICIAL, 5., Natal, 1988. Anais. Natal, UFRN, 1988. v. 1, p. 138-148.

SCOTT, R. D.; SOUZA, C. S. de. Concialiatory planning for extended descriptive texts. In: SIMPÓSIO BRASILEIRO DE INTELIGÊNCIA ARTIFICIAL, 6., Rio de Janeiro, 1989. Anais. Rio de Janeiro, PUCRJ, 1989. v. 1, p. 234-248.

SCHANK, R. C.; ABELSON, R. P. Scripts, plans, goals and understanding. Lawrence Erlbaum, 1977.

SIQUEIRA, I.; COSTA, A. E. Interação homem-máquina diante de interface em linguagem natural com português escrito. In: SIMPÓSIO BRASILEIRO DE INTELIGÊNCIA ARTIFICIAL, 4., Uberlância, 1987. Anais. Uberlândia, UFUb, 1987. v. 1, p. 175-185.

MINISTÉRIO DA CIÊNCIA E TECNOLOGIA
INSTITUTO DE PESQUISAS ESPACIAIS

PROPOSTA PARA PUBLICAÇÃO

- ☐ DISSERTAÇÃO
- ☐ TESE
- ☐ RELATÓRIO
- ☒ OUTROS

IDENTIFICAÇÃO

TÍTULO

PROPOSTA PARA UM GERADOR DE (PRÉ)ÍNDICES E DE (PRÉ)RESUMOS, A PARTIR DE UMA INTERPRETAÇÃO LINGÜÍSTICO-PRAGMÁTICA

AUTOR(ES)

CARLOS ALBERTO DE OLIVEIRA

ORIENTADOR

CO-ORIENTADOR

DISS. OU TESE

DIVULGAÇÃO

☒ EXTERNA ☐ INTERNA ☐ RESTRITA

EVENTO/MEIO

PRÉ

☒ CONGRESSO ☐ REVISTA ☒ OUTROS

LIMITE	DEFESA	CURSO	ÓRGÃO
__/__/__	__/__/__		

REV. TÉCNICA

NOME DO REVISOR

MARCIANA Leite Ribeiro

RECEBIDO	DEVOLVIDO	ASSINATURA
28/09/90	02/10/90	M. L. Ribeiro

APROVAÇÃO

NOME DO RESPONSÁVEL
FRANCISCO E. VITOLA
Chefe do Núcleo de Computação

APROVADO	DATA	ASSINATURA
☐ SIM ☐ NÃO	__/__/__	

REV. LINGUAGEM

Nº	PRIOR.	RECEBIDO	NOME DO REVISOR
		__/__/__	

PÁG.	DEVOLVIDO	ASSINATURA
	__/__/__	

DATILOGRAFIA

OS AUTORES DEVEM MENCIONAR NO VERSO INSTRUÇÕES ESPECÍFICAS, ANEXANDO NORMAS, SE HOUVER

RECEBIDO	DEVOLVIDO	NOME DA DATILÓGRAFA
	__/__/__	

Nº DA PUBLICAÇÃO:	PÁG.:
CÓPIAS:	Nº DISCO: LOCAL:

DIRETOR

AUTORIZO A PUBLICAÇÃO

☐ SIM
☐ NÃO __/__/__

OBSERVAÇÕES E NOTAS

ACEITO PARA SER PUBLICADO NOS ANAIS DO IV SEMINÁRIO DE AUTOMAÇÃO EM BIBLIOTECAS e CENTROS DE DOCUMENTAÇÃO.

PALAVRAS-CHAVE: TEXTO, RESUMO, ÍNDICE, LINGÜÍSTICA, INTELIGÊNCIA ARTIFICIAL, LINGUAGEM NATURAL

APÓS A PUBLICAÇÃO, ESTA FICHA SERÁ ARQUIVADA JUNTAMENTE COM A VERSÃO ORIGINAL

INPE 103