Sistemas de informação em linguagem natural: em busca de uma indexação automática

Marcílio de Brito

Resumo

Este artigo aborda o tratamento automático de linguagens naturais, particularmente a descrição do conteúdo informacional de textos, para melhorar sua indexação e preencher os requisitos dos sistemas de informação documentária, a partir de elementos fornecidos pela estruturação dos sintagmas nominais (SN). Uma nova ferramenta para análise morfossintática foi criada e desenvolvida com a linguagem de programação Starlet, baseada na teoria de Gramáticas Afixos, gramáticas em dois níveis, resultante do trabalho anterior de C.H. A. Koster. Usando-se gramáticas em dois níveis, aumentou-se a capacidade descritiva desta nova linguagem e produziu-se um simples e elegante modelo que possibilitou uma representação mais detalhada dos procedimentos de análise. Um corpo maior constituído de textos da Agence France Presse (AFP News Brieves) foi usado para testar o analisador morfossintático. Os resultados demonstraram claramente a capacidade das gramáticas em dois níveis para alcançar a formalização de fenômenos lingüísticos. As vantagens importantes deste método repousam na capacidade de se ter controle mais específico sobre a aplicação das regras de análise. Uma descrição mais sintática conduza programas mais bem adaptados ao meio computadorizado e às necessidades lingüísticas.

Palavras-chave

Recuperação da informação; Indexação automática; Tratamento automático da linguagem natural; Gramáticas Afixos.

Artigo extraída de parte da tese de doutorado intitulada *Réalisation d'un analyseur morpho-syntaxique pour la reconnaissance du syntagme nominal. Utilisation des grammaires affixes,* aprovada pela Université Claude Bernard, Lyon I, França, em 1991.

INTRODUÇÃO

O processo de pesquisa de informações nos leva à constatação de uma relação complexa entre escrever, comunicar e descrever a língua para exprimir um pensamento. Na verdade, o que se deseja é poder representar os conhecimentos materializados por atos de linguagem em um texto. Essa representação, no contexto de nossos-trabalhos, passa pela aplicação de métodos informáticos capazes de veicular propriedades intrínsecas à linguagem natural e à gramática que lhe é própria. O objeto do trabalho que apresentamos consiste substancialmente em transpor automaticamente um texto, em linguagem natural, para uma metalinguagem de análise gramatical. Esta análise nos permitirá observar a ordem estrutural dos constituintes da frase, que pode diferir da ordem linear. A frase pode então ser descrita por meio de relações de regência entre as unidades últimas. Descrever as funções sintáticas realizadas em um enunciado significa indicar as dependências existentes entre os elementos deste enunciado.

Por volta dos anos 60, surgem alguns semanticistas dedicando-se a pesquisas descritivas. Os problemas até então limitados são tratados, de preferência, sob o ângulo dos "campos morfossemânticos"[1] ou sob o ângulo lexicológico não ligado à morfologia, trabalhando-se essencialmente sobre textos circunscritos: descrição semântica do vocabulário da habitação, dos animais domésticos[2], de móveis[3], da moda[4,5] e do vocabulário político-social[6]. Daí aparece a possibilidade de uma organização fundada sobre as relações de antonímia de origem afetiva. Pouco a pouco essas relações atêm-se às propriedades físicas dos objetos denotadas por palavras estudadas, tira-se proveito do exame dos códigos ou linguagens documentárias destinadas aos pesquisadores sobre objetos ou textos de origens arqueológicas. A pesquisa sobre a teoria semântica ficará principalmente marcada pelos trabalhos de Greimas[4].

Muitos pesquisadores dedicaram suas carreiras em busca de uma melhor representação da linguagem natural e em particular ao seu tratamento automático. É preciso considerar a variedade dos componentes que intervém na linguagem – morfológicos, lexicais, sintáticos, semânticos, lógicos ... – e ressaltar as articulações entre os elementos de diferentes níveis. É preciso ainda distinguir entre os modelos que visam ao tratamento da linguagem em geral, os modelos mais limitados, porém ambiciosos, e por fim as modelizações de aspectos específicos fundamentadas em observações de casos particulares. Para esses últimos, as observações revelam uma parte significativa da realidade, não se atendo apenas a fenômenos isolados.

Pensar que seria possível evitar essa dificuldade, pelo emprego de uma experiência imaginária, que consistiria em tentar representar o efeito eventual do enunciado, *se* ele fosse pronunciado sem contexto, é enganar a si próprio. O que se chama uma ocorrência sem contexto é apenas uma ocorrência inserida em um contexto artificialmente simplificado, a significação constatada nessas condições não é necessariamente aquela que permitirá compreender aquelas registradas em contextos naturais. É evidente que seremos levados a uma descrição desse tipo, se tentarmos, artificialmente, colocarmo-nos fora de qualquer emprego efetivo*.

* Nossa proposta de análise não visará a reproduzir qualquer fato de significação, mas a chegara uma descrição morfossintática capaz de identificar e representar da melhor maneira a estrutura mínima do discurso, o sintagma nominal.

Uma primeira solução seria uma tradução para uma metalinguagem universal, isso, porém, parece-nos ainda utópico. Uma segunda solução consistiria em criar relações ligando os principais tipos de ocorrências entre si* . É bem verdade que a preocupação de produzir regras gerais, aplicáveis a essas ocorrências, assim como a casos particulares, faria aparecer a necessidade de uma metalinguagem. Assim, decidir qual é a **significação** de um enunciado fora de suas ocorrências possíveis é ultrapassar o terreno da experiência e da constatação e passar à elaboração de uma hipótese que precisa ser justificada.

Nós, humanos, sabemos que em alguma parte de um texto existe uma significação, conhecimentos, que podemos facilmente extrair por meio de operações naturais como a leitura. Os tratamentos automáticos ainda não atingiram o estado de representar convenientemente essas relações. Dessa forma, partimos, e ainda por um bom momento, à coleta de materiais, a exemplo do que fizeram Lamark e Darwin, esperando poder um dia elaborar novas teorias.

Decifrar textos antigos, escritos em línguas desconhecidas, utilizando alfabetos desconhecidos, é um exemplo particularmente instrutivo. Uma percepção intuitiva nos diz que esses textos contêm informações, quer sejamos capazes ou não de extrair. Esse sentimento é tão forte, quanto a convicção de que existe um significado em um jornal escrito em coreano, mesmo se não se compreende estritamente nada de coreano. Uma vez que o manuscrito ou a língua de um texto foram decifrados, ninguém mais pergunta onde reside a significação, enquanto ela está no texto, e não no método de decifragem, da mesma forma como a música está no disco, e não no toca-discos! Pode-se justamente identificar os mecanismos de decodificação por suas qualidades em não acrescentar significação aos sinais ou aos objetos que eles tratam; eles só fazem revelar a significação intrínseca desses sinais ou desses objetos.

Essa capacidade em fazer aparecer explicitamente cada uma das etapas de uma prova dentro de um mesmo quadro rígido é a característica principal dos sistemas formais, de tal maneira, que qualquer matemático possa verificar mecanicamente o trabalho de um outro. Aí, porém, encontramos o inconveniente de um quadro representativo em que não se pode criar uma nova regra a cada novo caso encontrado. Para esse dilema, existe, contudo, uma saída: a formalização de uma metateoria. As regras derivadas (os metateoremas) seriam agora os teoremas de um sistema formal maior, em que seria legítimo deriva-tos como teoremas, ou seja, teoremas da metateoria formalizada. Esses teoremas poderiam então ser utilizados para acelerar a derivação dos teoremas do cálculo de proposições. Essa idéia pode ser interessante, porém suscita imediatamente uma outra, a metametateoria, e assim por diante.

Em uma interpretação não significativa, não existe qualquer relação isomorfa aparente entre os teoremas de um sistema e a realidade. Assim, os teoremas podem parecer tão verdadeiros, quanto os não-teoremas. As interpretações significativas, ao contrário, indicam uma correspondência entre os teoremas e as verdades, ou seja, um isomorfismo entre os teoremas e uma parte da realidade. Apesar de, inicialmente, os símbolos estarem desprovidos de sentido, eles adquirem inevitavelmente uma "significação" a partir do momento em que um isomorfismo é descoberto. Não obstante, vale ressaltar que há diferenças entre a significação no sistema formal e na língua - nesta última, uma vez que se apreende o sentido de uma palavra, novas asserções podem ser fabricadas. A significação torna-se de certa forma ativa, pois ela engendra uma nova regra de criação de frases. O domínio de uma língua possui dessa maneira uma capacidade de evolução. Em um sistema formal, ao contrário, os teoremas são predefinidos por meio de regras de produção. Podemos escolher as "significações" em função de um isomorfismo (a condição de encontrar um) entre os teoremas e as asserções verdadeiras. Isso, contudo, não nos permite sair do sistema e acrescentar novos teoremas aos antigos. É o que constitui a "exigência da formalidade", ou seja, nunca se pode agir externamente às regras estabelecidas.

Como a língua é para nós o suporte de descrição dela mesma, poderíamos deduzir que *a* partir daí é possível dotar-se de meios para descrever o pensamento, porque é através de fenômenos lingüísticos que o fazemos naturalmente.

As classificações documentárias são conjuntos de morfemas ligados entre si por relações paradigmáticas* graças às quais esses conjuntos constituem classes diferentes, portanto línguas artificiais simplificadas, interessantes a mais de um título, e com atributos de análise lingüística. Por conseguinte, os caminhos que adotamos para a análise automática da linguagem natural têm-se verificado eminentemente lingüísticos.

O problema da documentação caracteriza-se por sua complexidade, devido ao fato de que ele não permite seu tratamento por métodos matemáticos, ou mesmo simplesmente científicos. A informática participa desse processo com uma pequena parte na solução teórica da documentação, ela só intervém quando todos os problemas mais incertos estão resolvidos (ou decididos).

O critério essencial da modernidade nos métodos de tratamento da informação não reside no emprego de equipamentos sofisticados e modernos, mas na adoção de uma forma nova de se colocarem os problemas.

A análise de um documento escrito compreende primeiramente as operações necessárias para que esse documento bruto possa ser utilizado (tratado) convenientemente por um sistema documentário qualquer. Ela consiste em um conjunto de operações destinadas a indexar o documento, a descrever seu conteúdo informacional, respeitando as condições impostas pelas linguagens utilizadas. A indexação, tal qual nós a vemos, é uma tradução lexical das unidades da língua, ou ainda uma tradução sintática, quando se trata de exprimir as relações entre as diferentes partes do discurso (que descrevem seu conteúdo, os desertores).

Diante dessas poucas informações, fica claro que nossa intenção aqui é simplesmente apresentar uma visão diferente, fundada sobre uma descrição mais rica dos fenômenos lingüísticos e que estão na origem de nossas reflexões sobre o tratamento automático da informação.

* Esta solução consiste em mostrar as diferentes relações (paradigmáticas e sintagmáticas) entre morfemas (Saussure), ao contrário da primeira, substancialmente gramatical (Chomsky).

* Paradigmáticas no sentido da Saussure e em particular as relações inerentes às classificações existentes.

No processo de análise automática de documentos, duas categorias são identificadas segundo suas naturezas:

1) pura escolha dos elementos existentes nos documentos (ex. KWIC - Key Word In Context);

2) transformação do conteúdo do documento.

A seleção não modifica o documento, apenas rearranja o documento segundo critérios diferentes. Essas operações são seguidas ou não de cálculo. Em todos os casos, a seleção opera por consulta a tabelas ou dicionários (dicionário negativo ou antidicionário).

Os melhores resultados de uma análise pelo método de índice de permutação se verificam quando utilizados sobre títulos e até mesmo sobre resumos. Nesse caso ainda, é preciso que os títulos representem convenientemente o conteúdo dos artigos.

O método de seleção automática de frases[7] é mais representativo do conteúdo de um documento. Os critérios fundamentais utilizados na extração de frases "representativas" são dados por cálculos de freqüência e de proximidade das palavras.

Simmons[8] introduz, nesse método, a identificação automática dos sinônimos feita pela consulta a uma lista de sufixos.

Os métodos lingüísticos nos levam primeiramente a analisar os trabalhos empreendidos para traduzir automaticamente um texto em linguagem natural, sob a forma de gratos escritos em uma metalinguagem de análise gramatical.

ANÁLISE GRAMATICAL AUTOMÁTICA

1) Análise por constituintes - permite observar a ordem estrutural dos constituintes, que pode ser diferente da ordem linear real. A frase pode assim ser descrita por meio de árvores onde figuram as relações de regência, entre as unidades últimas (o vocabulário terminal). Os métodos de análise gramatical diferem, segundo a escolha das unidades últimas. Alguns tomarão a palavra como unidade última, outros descerão ao nível dos radicais, afixos, desinências e muitas vezes níveis ainda inferiores.

2) O método de estemas, que descreve as funções sintáticas realizadas em um enunciado, indica as dependências existentes entre os elementos deste enunciado. Desse método, conhece-se 6 regra: uma palavra possui um só regente, mas pode reger várias.

Quadro 1 - Categorias fundamentais de tratamento segundo a natureza do documento.

Operações	Seleção	Transformação
Produtos Finais	1) índice de permutação (KWIC), escolha de frases extraídas de documentos originais	2) Grafos integrais de frases naturais 3) Conjuntos de termos definidos em um léxico organizado (com ou sem organizações sintáticas) 4) Resumo escrito em linguagem natural

3) A análise preditiva nasceu dos trabalhos de I. Rhodes[9], ao analisar frases em russo. Em seguida, Salton e Lemmon[10] aplicaram-na *à* análise documentária de textos em inglês. Aqui, c texto é tratado palavra por palavra. O programa guarda em memória uma lista de estruturas sintáticas esperadas ou possíveis. Cada possibilidade apresentada é comparada a uma lista chamada "reservatório de predições".

Experiências no tratamento de documentos conduzem a evidências marcadas pelas irregularidades e variações de pontos de vista para um mesmo indexador*, ou de um indexador para outro, e a possibilidade de confiar esta análise a autômatos. Por sua vez, a análise automática apresenta problemas delicados sobre os quais dirigiremos nossas proposições à luz da teoria das Gramáticas Afixos[11].

Antes, porém, é necessário conhecermos um pouco mais sobre os aspectos da indexação de documentos, vista sob o ângulo da teoria lingüística.

O trabalho de indexação é uma tarefa árdua e constantemente questionada, visto ser fruto da experiência daquele que a executa. O estatuto da palavra nos ajudará a melhor compreender seus valores. Vejamos primeiramente as distinções que faremos sobre a palavra "palavra" em nosso texto.

As palavras da língua, ou do dicionário, remetem unicamente aos seus significados, elas não designam referências*.

As palavras do léxico. Constituem o conjunto de palavras da língua. As palavras do léxico são igualmente palavras da língua ou do dicionário.

As palavras do discurso. No discurso, as palavras da língua são utilizadas para constituir unidades capazes de designar coisas. Atenção especial para os nomes próprios, que, apesar de se apresentarem como palavras isoladas, possuem a capacidade de estabelecer uma relação direta com seu objeto.

As palavras em terminologia. No léxico, como em terminologia, o que encontramos são **palavras,** mas certamente não se trata das mesmas **palavras.** Em terminologia, as palavras estão ligadas a coisas. As palavras que os lexicógrafos designam como substantivos são, na realidade, predicados, eles falam de qualidades, e não de substâncias.

Essas definições são de grande importância para o esclarecimento do processo referencial e não possuem necessariamente propriedades exclusivas. É talvez por isso que se prestam freqüentemente a confusões.

Em lógica formal, diz-se que as palavras do discurso correspondem a uma lógica **extensional,** enquanto às palavras da língua corresponde uma lógica **intensional,** e uma se opõe a outra (figura 1).

* Aquele que faz a indexação.

* A função referencial, também chamada **denotação,** produz-se entre o símbolo e a referência, ou seja, um objeto.

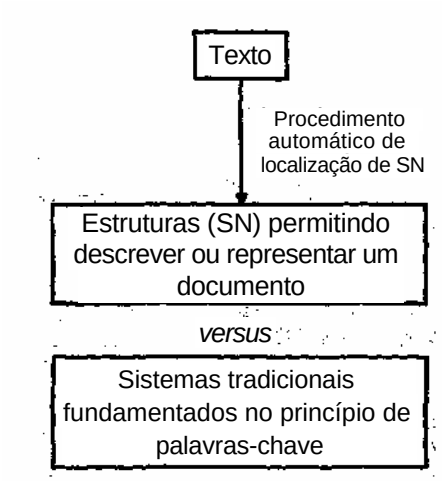


Figura 1 - Representação do texto.

Em face do dilema estabelecido entre palavras-chave e estruturas lingüísticas, a situação que consiste em descrever o conteúdo dos documentos com palavras do léxico parece-nos assim pouco sustentável, por ser incapaz de promover a transição léxico/discurso. É preciso procurar do lado do discurso os elementos necessários à representação dos temas do discurso. A busca pela função referencial do desertor nos leva naturalmente à unidade mínima do discurso, que é o sintagma nominal (SN). As razões que nos levam ao SN são muitas e complexas, não cabendo aqui nos estendermos sobre o assunto. Para maiores esclarecimentos, remetemos o leitor aos trabalhos do professor Le Guem[12] citados na bibliografia.

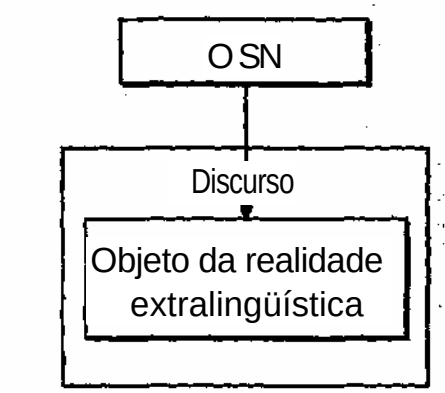


Figura 2 - Representação do sintagma nominal.

DA LINGUAGEM NATURAL À LINGUAGEM ARTIFICIAL: A FORMALIZAÇÃO

Um modelo é por excelência uma estrutura lógica ou matemática formalizada, utilizada para ressaltar um conjunto de fenômenos que, mesmo não tendo uma ligação de causalidade unívoca, possuem entre eles certas relações. A formalização é a operação que prepara um modelo explícito para o calculo lógico ou para operações dedutivas inequívocas.

Inspirado na notação X-barra de Chomsky, A. Berrendonner* normalizou as configurações arborescentes para a representação do sintagma nominal. A respeito dessa gramática, utilizaremos aqui as seguintes convenções:

- O lado esquerdo de cada regra é separado do seu lado direito por uma seta ($\rightarrow$), a concatenação é representada pelo símbolo (+).

- O vocabulário terminal (vt).

* Nós atribuimos as reflexões lingüísticas desse modelo a M. Le Guem, professar da Universidade de Lumière-Lyon II. Foi a partir de suas convicções lingüísticas a respeito do sintagma nominal (SN) que uma primeira gramática para um analisador morfossintático do francês escrito foi criada por M. Berrendonner. (Grammaire pour un analyseur: aspects morphologiques. 1983).

Vt-= {F-NOM, F-NOM-PRP, F-NOM-PRO, F-NAN, F-ADJ, D, D-DEF. D-NUM, D-IND, W-QUA, W-AAJ, P, P-DE}

F-NOM: os nomes.
F-NOM-PRO: os nomes-pronomes.

F-ADJ: os adjetivos.
D-DEF: os predeterminantes definidos.

D-IND: os outros predeterminantes.
W-AAJ: os advérbios modificadores de adjetivos (de intensidade).
P-DE: a preposição /de/.

F-NOM-PRP: os nomes próprios.
F-NAN: os nomes que podem ser, segundo o contexto, nome ou adjetivo.
D: os predeterminantes.
D-NUM: os predeterminantes numerais cardinais e similares.
W-QUA: os advérbios de quantidade.
P: as preposições

- o vocabulário não-terminal (Vn)
  Vn={N", N', N, A", A', A, D', Ep, Sp^n, sp}.

N" é o axioma e representa a categoria dos sintagmas nominais N" domina N', que domina N.
A" é o sintagma adjetival.
Ep é a expansão preposicional.
Sp^n é o sintagma preposicional.

Eis a gramática do sintagma nominal (SN)[12] que nos serviu de modelo para a realização do analisador morfossintático.

- sintagmas nominais:
  [1] N" $\rightarrow$ N"+N'
  [4] N" $\rightarrow$ D'+N'
  [5] N" $\rightarrow$ NOM-PRO
  [5] N" $\rightarrow$ NOM-PRP
- sintagmas adjetivais:
  [6] A" $\rightarrow$ A' + SP^n
  [7] A" $\rightarrow$ A'
- expressões nominais:
  [8] N' $\rightarrow$ N + SP"
- expressões predeterminativas:
  [11] N' $\rightarrow$ N
  [12] D' $\rightarrow$ - DEF + D-NUM
  [13] D' $\rightarrow$ P-DE + D-DEF
  [13'] D' $\rightarrow$ W-QUA + P-DE + D-DEF
  [13"] D' $\rightarrow$ W-QUA + P-DE
  [14] D' $\rightarrow$ D
- centros adjetivais:
  [15] A' $\rightarrow$ W-AAJ+A
  [15'] A' $\rightarrow$ A + EP
  [16] A' $\rightarrow$ A
- centros nominais:
  [17] N $\rightarrow$ + EP
  [18] N $\rightarrow$ N+A"
  [19] N $\rightarrow$ A"
  [20] N $\rightarrow$ A' + N
- nominais:
  [21] N $\rightarrow$ F-NOM
  [22] N $\rightarrow$ F-NAM
  [23] A $\rightarrow$ F-NAM
  [24] A $\rightarrow$ F-ADJ
- sintagma preposicional:
  [28] Sp $\rightarrow$ P' + N"
- seqüência de sintagma preposicional:
  [29] $SP^n \rightarrow$ Sp + Sp^n
  [30] $SP^n \rightarrow$ Sp
- expansão preposicional:
  [31] Ep $\rightarrow$ P' + N'
  [32] P' $\rightarrow$ P'

Através desse modelo lingüístico e de sua representação sob a forma de Gramáticas Afixos (gramáticas em dois níveis), veremos uma aplicação prática na realização do analisador morfossintático.

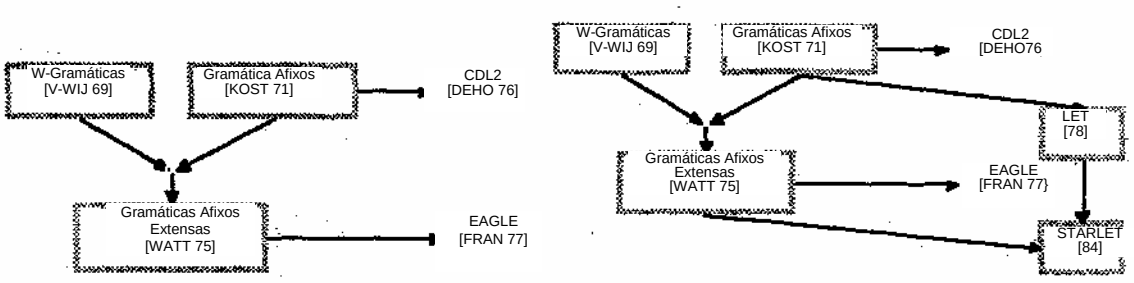
Antes, porém, devemos conhecer um pouco mais sobre o surgimento das Gramáticas Afixos.

AS GRAMÁTICAS AFIXOS

Após ter participado na definição da linguagem de programação Algol 68, C.H.A. Koster[11] introduz as Gramáticas Afixos para a realização de um tradutor associado ao Algol. Derivadas das W-gramáticas (Van Wijngaarden[13]), as Gramáticas Afixos apresentam um interesse particular para a descrição tanto sintática, quanto semântica das linguagens de programação.

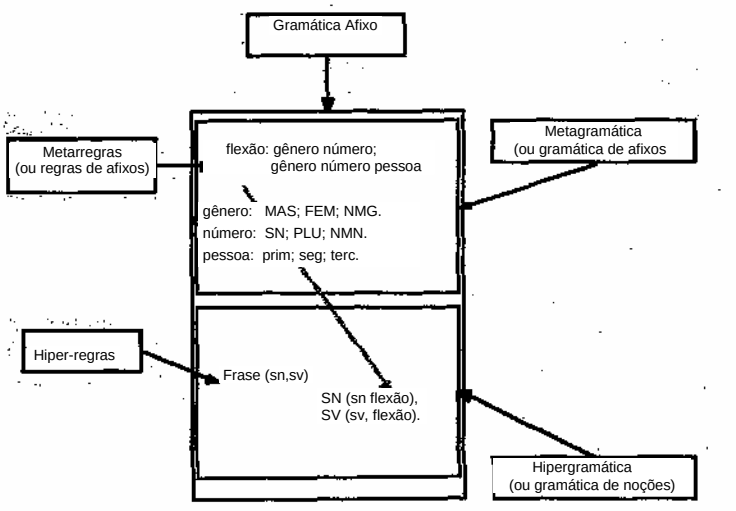
As Gramáticas Afixos são gramáticas em dois níveis e possuem simultaneamente a potência descritiva e a regularidade das W-gramáticas, introduzindo uma orientação do fluxo de informação que permite associar à gramática um analisador sintático contextual. Permitindo guiar a análise sintática por restrições contextuais, é possível suprimir a dicotomia entre análise sintática não contextual e análise semântica estática. Nesse sentido, o trabalho de Watt[14] consistiu em aproximar as Gramáticas Afixos e as W-gramáticas pela definição das Gramáticas Afixos Extensas.

Em 1984 J. Beney, professor no Laboratório de Informática da Universidade IN5A (Lyon-França), apresenta uma nova linguagem de programação, Starlet, baseada no princípio das Gramáticas Afixos Extensas. Starlet possui um compilador dotado de interpretação algorítmica e visa a produzir rapidamente tradutores mais confiáveis. Herdando os conceitos da teoria da compilação e dos sistemas gramaticais, Starlet aproxima-se das linguagens de programação lógica pelo seu mecanismo de unificação e pelo tratamento não determinista (figura 3).



[KOST 71] KOSTER, C.H.A *op. cit.* [11]
[V-WIJ 69] VAN WIJNGAARDEN, A. *et alii. op. cit.* [13]
[WATT 75] WATT, D. A. *op. cit.* [14]
[FRAN 77] FRANZEN, H. *et. alii. op. cit.* [15]
[DEHO 76] DEHOTTAY, J. P. *et. alii. op. cit.* [16]

Figura 3 - O aparecimento de Starlet



a flecha (→) ao centro, mostra a utilização do afixo "flexão" como variável.

Figura 4 — Ilustração de uma aplicação Starlet.

Uma aplicação em linguagem Starlet pode ser ilustrada pela figura 4, onde se evidencia a atuação da metagramática sobre a hipergramática.

O interesse de Starlet para o tratamento da linguagem natural está principalmente no fato de que as metarregras permitem a produção de novas regras de gramática da mesma maneira como a gramática engendra a linguagem. Trata-se sobretudo de um método potente para exprimir regras, sem atribuir excessivo poder ao formalismo. Cabe assinalar que as metarregras permitem explicar as regularidades da língua para as quais as regras não contextuais são incapazes de retratar.

Em particular, as relações entre as diferentes estruturas de frases são postas em evidência pelo fato de que elas são produzidas por regras distintas, mas que derivam da mesma metarregra. De certa forma, as metarregras estão mais próximas das transformações com a diferença de que as metarregras operam sobre regras e as transformações operam sobre árvores.

Dentro do modelo de representação formal da linguagem escrita, o grau de abstração permitido para a representação de estruturas profundas é ainda muito fraco. Observaremos que essas estruturas continuarão a ser expressas nos mesmos termos formais que as estruturas superficiais:

- as categorias (N", N', N...) são mencionadas como representações em dois níveis; estruturas profundas e superficiais são assim analisadas dentro das mesmas categorias sintáticas;

- as relações utilizadas para exprimir a combinatória das categorias são igual mente da mesma ordem, em dois níveis: estruturas profundas superficiais são representadas pelos mesmos tipos de relações.

Esse formalismo, para as estruturas de superfície, é o mesmo das estruturas profundas chomskyanas. Ele se constitui essencialmente das regularizações que não colocam em discussão os conceitos gramaticais de superfície.

A função das estruturas profundas, que é permitir a generalização da combinatória sintática superficial, conduz à substituição das noções particulares da gramática de superfície (as categorias e *as* relações dentro da estrutura do sintagma) por outras noções gramaticais mais gerais.

Descobre-se, assim, que, com as condições de Chomsky, uma classe de condições lingüísticas das quais a simulação deve ao mesmo tempo estar dentro do modelo gramatical e constituir um metadiscurso sobre as regras de transformação. Ele propõe, enfim, uma solução que é admitir que a componente transformacional de uma gramática deve ser organizada em dois níveis de discurso (figura 5):

- o primeiro nível sustenta um discurso simulatório sobre o objeto "língua", enumerando um conjunto de regras, as transformações, que são expressões cujas partes constitutivas (SN. P, /de/ ...) referem-se diretamente às partes da língua para descrever seu funcionamento combinatório;

- no segundo nível, constituindo um metadiscurso com relação ao precedente, o modelo deve conter as expressões cuja função *é* especificar a maneira como as expressões são . aplicadas. Essas expressões são metarregras em relação às regras de transformações, uma vez que seus átomos constitutivos devem ser os nomes das transformações. Essas metarregras devem ser enunciadas com base em símbolos que se referem não somente à língua, mas às transformações, ou seja, ao discurso primário simulando a língua.

Dessa organização chomskyana, tiraremos um proveito particular, aproximando o concerto de meta-algoritmo à metagramática de um sistema gramatical em dois níveis. Nesse sistema, as instruções teriam por função precisar as condições às quais se aplica cada transformação e predizer algoritmicamente sua aplicação. As asserções desse meta-algoritmo seriam, de um lado, as estruturas profundas (circunstâncias de aplicação das regras de transformação), que seriam descritas pela Gramática de Afixos, e, de outro lado, o conjunto de regras de transformação de nível 1, consideradas como operações virtuais, susceptíveis de serem executadas sobre estruturas profundas (descritas pela gramática de noções). O papel do meta-algoritmo consiste em gerar, por suas metarregras, um programa transformacional particular (ou vários) adaptado a cada estrutura profunda e capaz de convertê-la em uma estrutura de superfície bem formada.

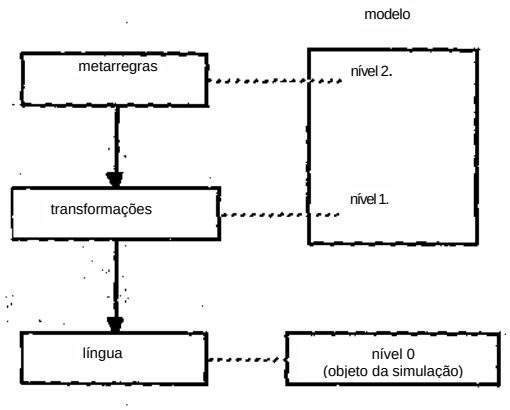


Figura 5 - A organização chomskyana.

Assim para Chomsky, é a posição do SN dentro da estrutura profunda que permitirá inferir sua função regente para o modelo de análise do sintagma nominal. A posição de cada elemento da estrutura sintática do discurso nos permitirá deduzir sobre a identificação das unidades sintagmáticas mais largas (N").

A análise morfossintática de um texto consiste essencialmente em aplicar um processo que, pela análise de formas sobre a superfície do texto, procura tirar um máximo de informações, permitindo uma estruturação do texto por reagrupamento das unidades sintáticas. Chamaremos essas informações sintáticas recuperadas na superfície do texto de **"conhecimentos sintáticos".** Para tratar esses conhecimentos, é natural utilizar-se das noções gramaticais, ou seja, partindo das regras que regem as possibilidades de associações de palavras entre si segundo suas características lexicais. São essas as noções que nos permitirão identificar e analisar unidades lingüísticas tais como o sintagma nominal (SN).

O ANALISADOR MORFOSSINTÁTICO

Seria vão pensar na criação de um analisador universal. A concepção de ferramentas informáticas especializadas no tratamento de acervos fisicamente não limitados e linguisticamente restritos seria mais realista. Por conseguinte, nosso analisador é destinado a operar com uma gramática particular e governada pela natureza dos textos de entrada.

Nosso interesse maior é poder identificar, no resultado da análise morfológica, um número máximo de estruturas regulares. Para fazê-lo, não devemos nos contentar em etiquetar as formas de superfície do texto por meio de traços metalingüísticos, mas executar, sobre essas formas, operações de regularização, constituindo em trazer as exceções a casos genéricos correspondentes.

O exemplo mais típico e representativo desse tratamento é a disjunção da amálgama /do/ em /de + o/. Isso faz com que um caso particular possa ser tratado como um caso genérico de preposição seguida de predeterminante. Este tipo de substituição contribui para simplificar a árvore de representação das formas do texto, ao mesmo tempo em que permite prevenir uma ambigüidade na análise. À forma */o/* será dada à categoria "predeterminante" com exclusão da interpretação pronominal.

Atendo-se ao princípio de redução do complexo ao simples, uma operação de regularização consiste, portanto, em trazer

uma forma de superfície a uma forma profunda, desaparecendo a primeira. Com esse método, pretende-se restringir as regras de análise a um número mínimo de operações, cada uma dotada de um rendimento máximo, evitando-se igualmente deixar para etapas posteriores problemas morfológicos que o analisador não sabe tratar.

A ANÁLISE MORFOSSINTÁTICA

A análise morfossintática do texto se passa em dois níveis: no primeiro, há consulta direta ao léxico; no segundo, há um pré-tratamento morfossintático.

No pré-tratamento de análise morfossintática, somente os aspectos de análise morfológica são abordados (figura 6).

Pré-tratamento morfossintático local

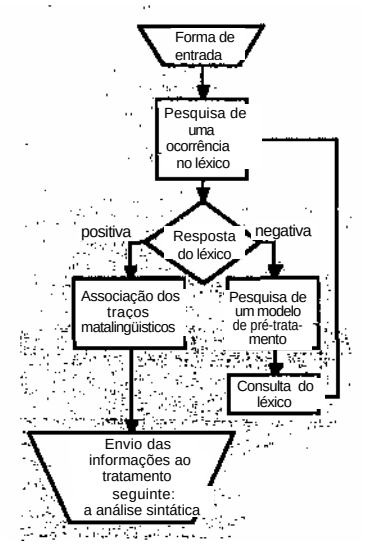


Figura 6 - Pré-tratamento morfossintático.

Na verdade, um pré-tratamento local de natureza morfossintática precede brevemente a análise morfológica para detectar, nas seqüências de formas, algumas propriedades sintáticas. Por exemplo, */este/,* seguido de um pronome relativo, é de natureza pronominal, e não predeterminativa. Uma analise centrada sobre uma só forma de superfície não seria sensível a um tal contexto (figura 7).

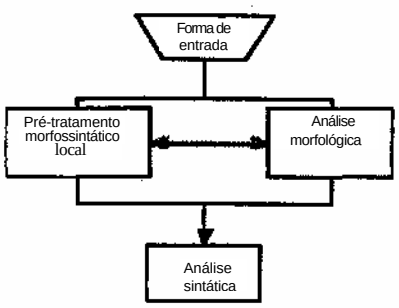


Figura 7 - A análise morfológica.

Esse tratamento permite a extração de algumas ambigüidades por atribuição "contextual" das categorias às formas de superfície. É nesta fase que se procede também a substituição das amálgamas.

Os efeitos desta operação de regularização do texto têm por conseqüências uma redução do inventário de formas do léxico, assim como as categorias necessárias à análise. Diversas propriedades sintáticas poderão aparecer como resultado desse desmembramento das amálgamas.

Vejamos, então, como se inicia a análise dentro desse contexto.

O texto é lido palavra por palavra. Para as formas desconhecidas, um módulo à parte é ativado para que se dê a entrada de novas formas no léxico (figura 8).

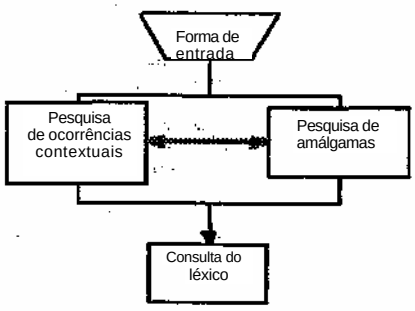


Figura 8 - Ocorrências contextuais.

Ainda nessa fase alguns movimentos retroativos *backtrack* são executados para que se possam detectar ocorrências sintáticas particulares *(cf.* locuções preposicionais).

Por uma série de razões, tanto informáticas como lingüísticas, a análise flexional só será proposta, uma vez que todas as etapas que utilizam a consulta direta do léxico ver-se-ão esgotadas, verificando-se antecipadamente se a forma em questão não é na verdade apresentada pelo léxico como uma forma canônica*

O tratamento mortos sintático, ao qual fazemos alusão aqui, constitui-se de uma série de modelos, repertoriando todos os componentes flexionais possíveis, de forma a reduzir uma forma flexionada em uma forma canônica presente no léxico (figura 9).

Análise morfossintática

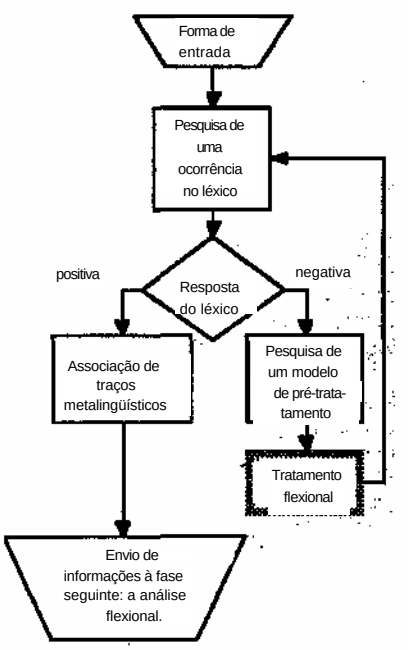


Figura 9 - O tratamento flexional.

* Entendemos por forma canônica a forma que é considerada como modelo, norma ou padrão. Assim diferenciaremos as ocorrências entre formas derivadas ou flexionadas.

O mecanismo aplicado permite reduzir uma forma flexionada (gênero e número) como /maestrinas/ à sua base /maestr/, permitindo chegar à sua forma canônica /maestro/ (masculino singular) presente no léxico. À forma resultante agruparemos as interpretações lingüísticas enviadas pelo léxico ou deduzidas das regras de análise. O conjunto dessas informações se compõe de:

- uma categoria gramatical (F para os nomes e adjetivos);
- uma subcategoria (NOM para os nomes próprios e comuns);
- valores flexionais em gênero (FEM) e em número (PLU);
- valores semânticos (ANI/INA) animado/inanimado.

Partindo das informações recuperadas na superfície do texto, o sistema se encarrega de construir uma árvore sintática. É o resultado da fase de análise sintática. Os nós dessas árvores portam informações bem importantes para o estabelecimento de relações entre as diversas partes do texto. Essas unidades, os SN, são os objetos da realidade extralingüística, são elas que por um lado descrevem o conteúdo dos documentos e por outro revelam as inter-relações existentes, (figura 10).

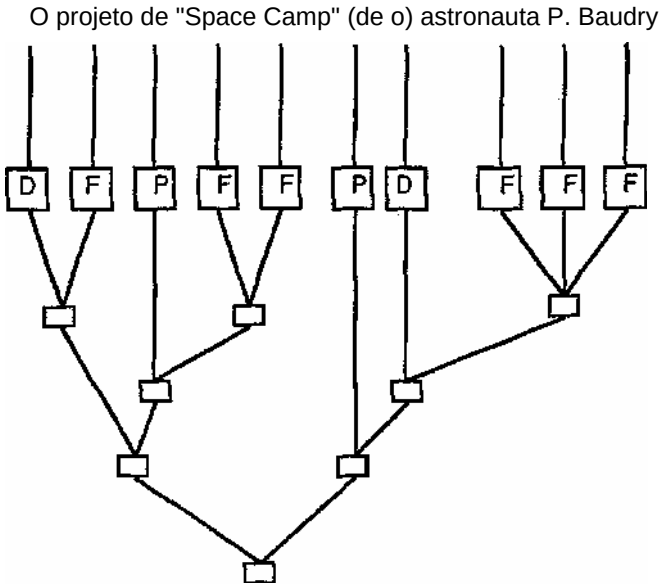


Figura 10 - Representação dos valores gramaticais terminais

As unidades explícitas, os nós da árvore de análise, são validadas pelas expressões de afixos, para descrever as informações sintáticas na superfície do texto (figura 11).

No âmbito da nossa tese, foi-nos possível analisar mais de 200 textos da Agence France Presse (AFP News Brieves), dos quais se extraíram os SN. Esses resultados, comparados às extrações manuais, mostraram por um lado a precisão com que as estruturas procuradas foram selecionadas e por outro um índice de equivalência automático/manual até então nunca alcançado por sistemas automáticos*.

Essas unidades (os SN), assim recuperadas, constituem o conjunto de estruturas que remetem diretamente aos objetos da realidade extralingüística de que fala o texto; em outras palavras, elas representam o conteúdo informacional do documento tratado.

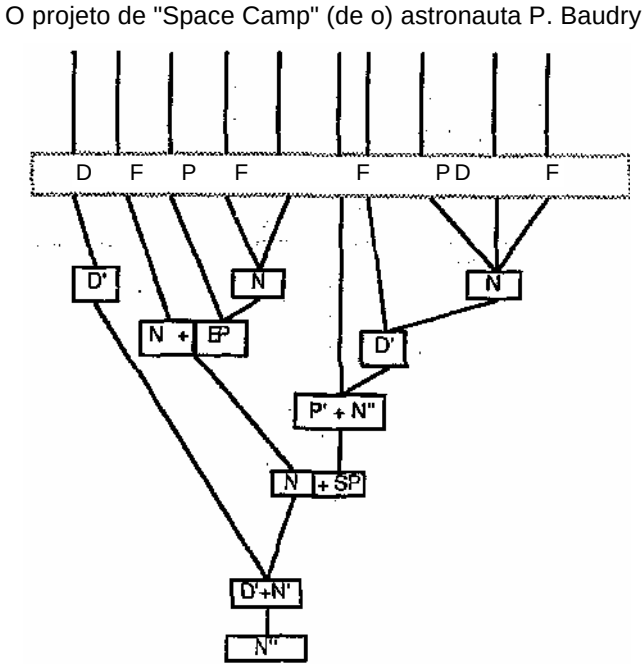


Figura 11 - Representação da árvore de análise.

* Esses resultados encontram-se repertonados e analisados em Brito, M. de, *Réalisation d'un analyseur morpho-syntaxique pour la reconnaissance du syntagme nominal. Utilisation des gramaires affixes.* Université Claude Bernard, Lyon I, França, 1991. (Tese de doutorado em Informática Documentária).

CONCLUSÃO

Sobre a análise de textos em linguagem natural, a redescoberta das Gramáticas Afixos sob a forma da linguagem Starlet vem mostrar novos horizontes para o tratamento da informação. A fidelidade e a reprodutibilidade dos resultados alcançados mostra a qualidade e o nível de refinamento das análises. Enfim, nossos meios e métodos assemelham-se, de muito perto, ao papel dos compiladores, para os quais o processo de análise sintática é minuciosamente controlado e estreitamente ligado às ações adotadas pelos usuários (o programador). As informações dadas pelo sistema devem permitir ao usuário intervir de maneira precisa, quer seja para corrigir um erro, ou para redirigir as ações do programador no âmbito de uma aplicação.

A utilização das Gramáticas Afixos nos mostrou ainda que poderíamos aumentar a qualidade dos resultados das análises morfossintáticas por meio de uma descrição gramatical mais bem adaptada, mais fina e mais fiel ao modelo lingüístico proposto. Graças aos métodos próprios às gramáticas em dois níveis, pode-se escrever programas compactos e potentes, revelando com riqueza de detalhes as relações expressas pela teoria lingüística associada. Assim, foi-nos possível mostrar que:

- o tratamento dos problemas lingüísticos pode se realizar com grande especificidade, de forma concisa e elegante;

- a análise dos resultados pela análise das regras de gramática utilizadas na avaliação do problema permite, por diversas maneiras, melhor formalizar os fenômenos lingüísticos, melhor compreendê-los, identificar as fontes de erro dentro da análise com mais rapidez, ou simplesmente verificar a eficácia da gramática;

- ainda pela forma descritiva das Gramáticas Afixos, chegamos à constatação de que é possível analisar textos em ausência completa de léxico para a categoria dos nominais. O ambiente sintatico, ricamente expresso pela descrição dos afixos, permite a obtenção de uma representação de sintagmas pelo reconhecimento dos indicadores de superfície; a classe de nominais, sendo uma classe "aberta", pode ser deduzida pela análise de outros componentes ambientais;

- os sistemas de indexação automática para bases em texto integral possuem aqui um instrumento que lhes permitirá fundar a nova geração de sistemas de recuperação da informação.

Trabalhos semelhantes a esse já foram realizados na Holanda pela equipe de pesquisa do professor C.H.A. Koster, porém jamais algo comparável havia sido feito sobre línguas originariamente latinas. Nós esperamos com essa experiência fazer brotarem maiores incentivos à concepção de novas ferramentas de tratamento da informação, indexação assistida por computador e muitos outros.

REFERÊNCIAS BIBLIOGRÁFICAS

1. GUIRAUD, P. De la grive au maquereau: le champ morpho-syntaxique des noms de l'animal tacheté, *Le Français moderne,* n.34,1966.

2. MOUNIN, G. Un champ sémantique: la dénotation des animaux domestiques. *La linguistique,* n.1, 1965.

3. POTTIER, B. *Recherches sur l'analyse sémantique en linguistique et en traduction mécanique,* 1963.

4. GREIMAS, A.J. *La mode en 1830: Essai de description du vocabulaire vestimentaire d'après les journaux de model de l'époque.* Paris: Sorbonne (thèse dactylographiée), s.d.

5. BARTHES, R. *Le système de la mode,* Paris, Seuil; 1967.

6. DUBOIS, J. *Le vocabulaire politique et social en France de 1869 à 1870,* Paris: Larousse, 1963.

7. LUHN, H.P. The automatic creation of lilerature abstracts. *IBM Journal of Research and Development,* 1958.

8. SIMMONS, F., McCONLOGUE, K. L. Maximum depth indexing for computer retrievel of English language data, *American Documentation,* 1963.

9. RHODES. *A newapproach to the mechanical syntatic analysis of Russian.* National Bureau of Standards, 1959.

10. LEMMON., A. Report on a syntactic analysis program for information retrieval. In: SALTON, *Information Storage and Retrieval,* 1962.

11. KOSTER, C.H.A. Affix Grammars. *ALGOL 68 implementation,* 1970.

12. LE GUERN, Michel. Un analyseur morpho-syntaxique pour l'indexation automatique. *Le Français Moderne,* juin, 1991.

13. VAN WIJNGAARDEN, Aad, MAILLOUX, a, PECK, J.D.L KOSTER, C.H.A. *Report on the algorithmic language ALGOL 68. MR 101,* Amsterdam: Mathematisch Centrum, 1969.

14. WATT, D.A. *Analysis-oriented two-level grammars* Berlin: Technical University of Berlin, 1975. (Ph.D. thesis, Glasgow, 1974).

15. FRANZEN, H., HOFFMANN, B., POHL, B., SCHMIEDECKE, I.R, *The EAGLE parser generator:* an experimental step towards a practical compiler-compiler using two level grammars. In: 5TH ANNUAL III CONFERENCE. France: Guide I, 1977. p. 397-420.

16. DEHOTTAY, J.P., FEUERHAHN, H., KOSTER, C.H.A., STAHL, H.M. *Syntaktische beschreibung von CDL2.* (Internal report) Berlin: Technical University of Berlin, Sept. 1976. multigr.

BIBLIOGRAFIA CONSULTADA

BENEY, *Jean. Présentation de STARLET/GL.* INSAL, Laboratoire d'lnformatique Apliquée, Juillet, 1989. Revisé en Février 1990. 58p. (Documentação interna).

BENEY, Jean, BOULICAUT, J-François. Des spécifications grammaticales à la programmation logique: le compromis Starlet In: Actes des journées AFCET. *Nouveaux Langages pour le Gènie Logiciel.* Evry: BIGRE-GLOBULE. n.45. octobre 1985. p.81-88.

BERRENDONNER, Alain. *Grammaire pour un analyseur:* aspects morphologiques. Université de Fribourg (CH), Grenoble-ll, Lyon-l, Lyon-ll, 1979. 103p. Document de travail du groupe SYDO.

BOUCHÉ, Richard. *Valeur référentielle et langage d'indexation dans les systèmes d'informations documentaires.* In: COLLOQUE SUR ARCHIVES ET TEMPS RÉEL. Lille: CREDO (Univ. Lille-lll)/ADBS/archives du Nord, 28 novembre 1988.12p. multigr.

BOUCHÉ, Richard. Le syntagme nominal, une nouvelle approche des base de données textuelles. In: ACTES du colloque terminologie et industries de la langue. *META Journal des traducteurs,* Montréal, v.34. n.3, september, p.429-434. 1989.

BOULICAUT, J-François. Méta-compilation et programmation: des règles méthodologiques pour fiabiliser la construction de programmes. *Génie logiciel et Systèmes Experts,* n. 11, mars, p.36-4a 1988.

CLEAVELAND, J.C., UZGALIS, R.C. *Grammars for programming languages.* Netherlands: Elsevier North-Holland, 1977.154p. (Programming languages series nº 4).

COLMERAUER, Alain, KANOUI, Henry, VAN CANEGHEN, Michel. PROLOG: Bases théoriques et développements actuels. *Technique et Science Informatiques,* Gauthier-Villars, v.2, n.4, p.271-312.1983.

COYAUD, Maurice. *Linguistique et documentation:* les articulations logiques du discours. Langue et langage, Librairie Larousse, 1972. 173p.

DUPONT, Pierre. *Eléments logico-sémantiques pour une analyse du francais.* Lyon: Université Lumière Lyon-ll, 1983. 580p. (Thèse d'État).

FRECON, Louis. *Pratique des grammaires affixes:* Réalisations & questions ouvertes. *In:* ATELIER LYON/NIJMGEN sur les Grammaires Affixes. Les Hautannes, St Germain au Mont d'Or, 26-29 juin 1989. 16p.

KOSTER, C. H.A. *Two level grammars* In: Advance course in compile construction. Lecture Notes In Computer Science, 21, Springer-Verlag, 1974. p.146-156.

LAINÉ, Sylvie. *Extraction et sélection de descripteurs complèxes dans un ensemble de textes pour leur indexation automatique.* Lyon: Université Claude Bernard. Lyon-l, juin 1982. 137p. Thèse de Docteur-Ingénieur en mathématiques (informatique).

LE GUERN, Michel. Sur les relations entre terminologie et lexique. In: ACTES du colloque terminologie et Industries de la langue. *META Journal des traducterus,* Montréal, v.34, n.3, Septembre, 1989. p.340-343.1989.

LENNON, Martins, PEIRCE, D.S., TARRY, B.D., WILLETT, P. An evaluation of some conflation algorithms for information retrieval. *Journal of Information Science,* North-Holland, march, p.177-183. 1981.

METZGER, J-Paul. *Syntagmes nominaux et information textuelle.* Lyon, Université Claude Bernard - Lyon-l, octobre, 1988. 325p. (Thèse de Docteur d'Etat Es Sciences).

NEF, Frédéric. *La logique du langage naturel.* Paris Editions Hermès, 1989. 63p.

PEREIRA, Fernando C.N., WARREN, David H. D. Definite clause grammars for language analysis: a survey of the formalism and a comparison with augmented transition networks. *Artificial intelligence,* North-Holland, v.13, p.231-278. 1980.

SABAH, Gerard. *L'intelligence artificielle et le langage:* représentations des connaissances. Paris: Hermès, v.1,1988. 352p.

Artigo aceito para publicação em 18 de dezembro de 1992.

Marcílio de Brito

Doutor em Informática Documentária pela Université Claude Bernard Lyon-l, França, é funcionário do Serviço Brasileiro de Apoio às Pequenas e Médias Empresas (Sebrae) e professor visitante da Universidade de Brasília, Departamento de Ciência da Informação e Documentação.

Information systems in natural languages: looking for an automatic indexing

Abstract

This paper deals with the automatic treatment of natural languages, particularly the informational description of texts in order to improve their indexing and match the requirements of documentary information systems from noun phrase structured elements. A new tool for morpho-syntactic analysis was created and developed with the programming language Starlet based on the theory of Affix Grammars, two-level grammars, which resulted from C.H.A. Koster's early work. Using two-level grammars increased the descriptive power of this new language and produced a simple and elegant frame that allowed a more detailed representation of the analysis procedures. A large corpus of texts from Agence France Presse (AFP News Brieves) was used to test the morpno-syntatic analyser. The results clearly demonstrated the power of two-level grammars to reach linguistics phenomena formalization. The main advantages of this method lay in the ability to have stricter control on analysis rules. A better syntatic description leads to programs better adapted to computerized environment and linguistics needs.

Key words

Information retrieval; Automatic indexing; Automatic treatment of natural languages; Affix Grammars.