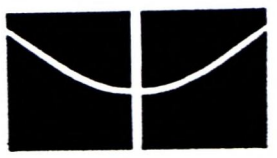
FUNDAÇÃO UNIVERSIDADE DE BRASÍLIA

ANÁLISE AUTOMÁTICA DO SISTEMA LINGÜÍSTICO PORTUGUÊS

O SISTEMA LINGA E SUAS APLICAÇÕES



ANÁLISE AUTOMÁTICA DO SISTEMA LINGÜÍSTICO PORTUGUÊS

O SISTEMA LINGA E SUAS APLICAÇÕES

DISSERTAÇÃO APRESENTADA AO DEPARTAMENTO DE LETRAS E LINGÜÍSTICA DO INSTITUTO DE EXPRESSÃO E COMUNICAÇÃO DA UNIVERSIDADE DE BRASÍLIA COMO REQUISITO PARCIAL À OBTENÇÃO DO TÍTULO DE MESTRE EM LINGÜÍSTICA.

POR

NELMO ROQUE SCHER

JUNHO - 1985

BANCA EXAMINADORA:

Profª Dra. Lúcia Maria Pinheiro Lobato

Prof. Dr. Jaime Robredo

Profª Dra. Stella Maris Bortoni Ricardo
(Presidente)

AGRADECIMENTOS

À Profª Stella Maris Bortoni Ricardo, pela orientação decisiva e competente;

À Profª Lúcia M.P. Lobato, por seu apoio reconfortador e por sua orientação segura quanto à sintaxe gerativa;

Ao Prof. Johann Haller - amigo de todas as horas - iniciador e incentivador da Lingüística Computacional na UnB;

Ao Prof. Jaime Robredo, por seu incentivo ao intercâmbio das ciências na UnB;

Ao Prof. João Ferreira, pela revisão final deste trabalho e pelas palavras de carinho que tanto me encorajaram;

Aos professores e colegas do curso, especialmente aos integrantes da equipe do LINGA: Antônio Batista Pereira, Carlos Alberto de Oliveira, Simone Borges Bastos e Mary Lyn Kelso;

À CAPES, pelo apoio financeiro de parte do curso;

Ao Ernesto Capocci (in memoriam).

A meus pais,
que me incentivam
na busca pela Vida.
A meus irmãos e irmã,
pelo carinho sempre renovado.
À Maria do Carmo, por amor.

E o Verbo se fez carne...

Jo 1.14

LISTA DE ABREVIATURAS

ADV	-	Advérbio
AILA	-	Associação Internacional de Lingüística Aplicada
ALPAC	-	Language and Machines-computation in Translation and Linguistic Automated Procession Advisory Committee, National Academy of Sciences
ATT	-	Adjetivo em função de adjunto adnominal
AUI	-	Forma infinitiva de um verbo auxiliar
AUX	-	Forma flexionada de um verbo auxiliar
AVP	-	Advérbio
CNP	-	Categoria gramatical (número e pessoa)
COM	-	Vírgula
CON	-	Conjunção coordenativa
CONDOR	-	Comunication in natural language with dialoge - orient retrieval
DET	-	Determinante (artigo definido e indefinido)
DEM	-	É a partícula 'demais' que Haller distinguiu de outras formas adverbiais
ELP	-	Elemento de uma locução preposicional
FAC	-	Função Sintática da Categoria Gramatical
FASIT	-	Fully Automatic Syntactically Based Indexing System
FDIC	-	Dicionário de Freqüência
GER	-	Gerúndio
HTAB	-	Tabela das Homografias Sintáticas
INF	-	Infinitivo em Função Verbal
ITA	-	Instituto Tecnológico da Aeronáutica
INPE	-	Instituto Nacional de Pesquisas Espaciais
LC	-	Linguística Computacional
MDI	-	Forma infinitiva de um verbo modal
M/G	-	Modo/Gênero
NAT	-	Numeral em função de adjunto adnominal
NDIC	-	Dicionário de Formas Nominais
NML	-	A supercategoria do adjetivo como núcleo do sintagma nominal
NNO	-	Numeral como núcleo de um sintagma nominal
NOM	-	Núcleo de um sintagma nominal

NEG - Advérbio de negação
OPR - Pronome oblíquo proclítico
PAR - Particípio em função verbal
PRD - Adjetivo ou particípio em função precadicativa
PRP - Preposição simples
PNT - Ponto final de uma frase
QNT - Quantificador (na acepção da gramática gerativa)
QAT - Categorias integrantes da supercategoria do adjtivo (cf. Lemle 1984:150)
REL - Pronome relativo
SUB - Conjunção subordinativa
SEP - Elementos gráficos (parênteses, travessão, barra, etc.)
SPR - Pronome do caso reto
SATAN - Sistema de Análise Automática de textos de Saarbrucken
SPIRIT - Sistema para indexação e recuperação de informações textuais
SA - Sintagma Adjetival
SAdv - Sintagma Adverbial
SN - Sintagma Nominal
SP - Sintagma Preposicional
SV - Sintagma Verbal
TIM - Tempo das Formas Verbais
VDIC - Dicionário de Verbos
VTAR - Morfologia Verbal
VRB - Forma verbal flexionada
VBM - Forma flexionada de um verbo modal

CONVENÇÃO

* - Indica inclusão no glossário (pp 127-32)

LISTA DE QUADROS

RESUMO

A Lingüística Computacional tem tido influência considerável sobre o desenvolvimento da teoria lingüística nos últimos anos, pois induz os lingüistas a definirem com a maior precisão possível as regras e as idiossincrasias de um sistema lingüístico para que o mesmo possa ser submetido à análise automática e, então, receber as mais diversas aplicações, como, por exemplo, a indexação automática, a tradução automática, a criação automática de um tesauro de uma ciência, aplicações na Informática, nos meios modernos de comunicação, no ensino de línguas estrangeiras, etc.

Nesta dissertação, objetivou-se apresentar o desenvolvimento de um sistema de análise automática da língua portuguesa, denominado LINGA ("Linguistic Analysis"). Esse sistema está organizado em tabelas, dicionários, rotinas e algoritmos obedecendo à distinção entre as categorias lexicais e não-lexicais.

Os passos da análise lingüística operados pelo LINGA são: segmentação de um texto em frases e das frases em palavras; busca nos dicionários desenvolvidos no sistema, análise morfológica, sintática e das homografias.

A partir das definições das categorias lexicais e não-lexicais, o sistema LINGA define a posição sintática das palavras por meio de uma sofisticada rotina estruturada a partir dos traços configuracionais dos itens lexicais. Assim, toda a estrutura gramatical nos níveis morfológicos e sintáticos é abrangida por procedimentos pré-estabelecidos.

Com o desenvolvimento de procedimentos que analisam a estrutura morfossintática da língua portuguesa, o sistema

LINGA está apto a diversas aplicações. Como objetivo prático desta dissertação oferecemos contribuições da análise lingüística para a área da Teoria da Informação. Especificamente tratamos da análise do sintagma nominal (SN) e de suas implicações na indexação automática.

ABSTRACT

Computational Linguistics has had a large influence in the development of the linguistic theory in the last years because it contributes to the accuracy in the definition of the linguistic rules.

Its main applications have been in indexing systems, automatic translation, development of scientific thesauri, modern system of mass midia, language teaching, retrieval of information, etc.

The purpose of the present dissertation is to describe in detail the system of automatic analysis of the Portuguese language - LINGA. This system comprises tables, dictionaires, routines and algorithms wich are organized according distinction between lexical and non-lexical categories.

The procedures of linguistic analysis performed by LINGA are: text segmentation into sentences and sentence segmentation into words; research in the dictionary and morphological analysis; syntactic analysis and analysis of homographies.

After the definition of the lexical and non-lexical categories, the LINGA system defines the syntactic position of the words through a sophisticated structured routine wich takes into account the configurational features of the lexical items. It covers therefore all the morphological and syntactic structure of the grammar.

With the development of procedures to analyse the morpho-syntactic structure of the Portuguese language, the LINGA system can have several applications. The practical purpose of this dissertation was to offer contributions of linguistic analysis to the theory of information. We deal in particular with the noun phrase analysis and offer a discussion of its implication to indexing systems.

ÍNDICE

INTRODUÇÃO

A importância do uso do computador se acentua cada vez mais na análise das línguas naturais. É de suma relevância conhecer esse importante instrumento de serviço para o exame, classificação e análise dos fenômenos lingüísticos.

O nosso interesse pelo assunto partiu do exame de dados fornecidos pelo computador na análise de textos durante o curso de lingüística computacional (LC) realizado na Universidade de Brasília, no período de 1982-83. Sentimos, então, a importância de formalizar as regras do sistema lingüístico, nos níveis morfológico e sintático, de tal sorte que fosse possível processar estas regras pelo computador. O processamento automático das regras do sistema lingüístico pelo computador permite as mais diversas aplicações. Dentre essas podemos citar, entre outras, a indexação automática, a obtenção de resumos de textos, a criação de thesauri de ciências, o ensino de línguas, a tradução automática, a análise estilística de textos literários a partir da estrutura morfossintática das frases, aplicações na informática e na inteligência artificial, etc.

Temos como objetivos em nosso trabalho:

- apresentar um resumo do estado da arte na LC em alguns centros da Europa, dos Estados Unidos e, especificamente, do Brasil;

- descrever sistematicamente todos os dicionários, tabelas, rotinas e procedimentos do sistema de análise automática da língua portuguesa, denominado LINGA;

- demonstrar como os resultados da análise lingüística teórica são apreendidos e aproveitados no desenvolvimento da LC que, num processo circular, contribui, por sua vez, para tornar a análise lingüística mais precisa e formalizada;

- discutir a estrutura do SN português, à luz de várias abordagens da teoria gerativa, enfatizando o fato de que o modelo padrão da regra de expansão do SN é redundante com as informações do léxico, que no LINGA estão codificadas na rotina HTAB.

No capítulo 1 descrevemos o estado da arte apresentando um resumo das aplicações e desenvolvimento da LC nos Estados Unidos e na Europa. Apresentamos, também, nesse capítulo o início dos trabalhos nessa área no Brasil.

No capítulo 1 descrevemos o estado da arte apresentando um resumo das aplicações e desenvolvimento da Lingüística Computacional (LC) nos Estados Unidos e na Europa. Apresentamos, também, nesse capítulo o início dos trabalhos nessa área no Brasil.

No capítulo II procuramos situar a LC como um ramos da própria ciência da lingüística, que se utiliza de um instrumento de trabalho altamente sofisticado para formular e testar as regras lingüísticas nas enormes massas de informações que já se apresentam hoje aos usuários. Segundo Rector (1984:1) "o computador nada mais é do que um meio de comunicação de massa, que surge em função do fenômeno da industrialização". Em razão disso, desconhecê-lo em seu funcionamento e em suas aplicações como instrumento de trabalho nos mais variados campos da ciência moderna é, no mínimo, parar no tempo e ignorar o progresso da ciência.

Concordamos com Pais (1981:20) quando afirma que "o sistema lingüístico é gerador e veículo de significação e um instrumento de produção, transmissão, armazenamento e recuperação da informação". Sua assertiva nos alenta no nosso propósito de formalizar as regras morfossintáticas do sistema lingüístico de tal modo que possam ser submetidas à análise por computador.

Consideramos, ainda no capítulo II, o desenvolvimento da LC nos países desenvolvidos e em vias de desenvolvimento e o crescente interesse que despertam suas aplica-

ções na tradução automática, na biblioteconomia, na informática, etc. Nesse capítulo retomamos também a discussão da aplicação da LC à língua portuguesa e historiamos os primeiros trabalhos na área.

No capítulo III procuramos mostrar os intrincados processos de formalização das regras de um sistema lingüístico até que o mesmo esteja em condições de ser submetido a um processamento eletrônico. A tarefa mais trabalhosa é converter as regras lingüísticas no formato preciso e detalhado exigido pelo computador. Devemos lembrar que o computador é máquina, portanto, de todo obtuso: nada processa sem que os dados estejam previamente armazenados em programas. A LC, segundo Bott (1976: 206) "tem tido influência considerável sobre o desenvolvimento da teoria lingüística nos últimos anos, pois tem forçado os lingüistas que trabalham com computadores a formular suas regras de maneira mais precisa do que teriam feito em outras circunstâncias".

Descrevemos no capítulo III, sucintamente, os sistemas SATAN, CONDOR, SPIRIT, FASIT e o método PARSING com o objetivo de mostrar o desenvolvimento de sistemas de análise lingüística automática, em diferentes códigos lingüísticos, que utilizam metodologias de organização e funcionamento distintos.

Especificamente, a partir da seção 3.6 objetivamos apresentar o sistema de análise automática da língua portuguesa - o sistema LINGA - descrevendo seus objetivos, sua metodologia, sua estrutura e seus procedimentos. Esse

sistema foi desenvolvido por Johann Haller em colaboração com uma equipe de alunos dos cursos de Pós-graduação em Lingüística e Biblioteconomia da Universidade de Brasília, da qual o autor desta dissertação é integrante. Procuramos situar os fundamentos teórico-lingüísticos que sustentam a elaboração do programa, de suas tabelas, de suas rotinas e de seus procedimentos.

Tendo em vista os objetivos da LC, que são formular de maneira precisa e objetiva as regras do sistema lingüístico e testá-las numa grande massa de dados, um curso de LC não deve ser entendido como uma simples aplicação de recursos da área da computação à análise lingüística. Tomamos a LC como um ramo da disciplina lingüística que desenvolve seus fundamentos teórico-metodológicos próprios, nutrindo-se no desenvolvimento da teoria lingüística e da teoria da informação. Nesse sentido, a presente dissertação se define como um trabalho de caráter lingüístico cujo principal objetivo é trazer uma contribuição aos processos de análise lingüística, na medida em que aponta os caminhos nos quais o analista poderá ser auxiliado pela máquina em seu trabalho.

Em razão do exposto no parágrafo anterior, no capítulo IV nos concentramos em um aspecto da estrutura lingüística: o sintagma nominal (SN). Escolhemos o SN por diversas razões, a saber: 1) existem diversas análises na teoria gerativa a respeito do sintagma nominal, que acompanham a evolução que essa teoria teve nas décadas de 60, 70 e 80;

2) o estudo do SN vem ao encontro de nosso objetivo prático, qual seja o de oferecer contribuições da teoria lingüística para a indexação automática; 3) a análise do SN nos permite discutir as implicações da teoria $\bar{X}$ (xis barra) na formulação da rotina HTAB do sistema LINGA.

No capítulo V, preocupamo-nos em estabelecer os elementos do SN que preenchem os requisitos lingüísticos para a definição de um descritor simples e composto.

Quanto à indexação automática, baseamo-nos em Robredo (Robredo 1982, 1983) que desenvolvem uma técnica que em muito se assemelha aos objetivos do sistema LINGA no que tange a esse assunto.

Esperamos poder contribuir para o sistema desenvolvido por Robredo com a inserção de procedimentos de análise lingüística operados pelo LINGA.

CAPÍTULO I

O CONTEXTO DA PESQUISA

O estudo da LC segundo Haller (1983), tem suas origens nas primeiras tentativas da tradução automática, feitas nos Estados Unidos nos anos 50, pela IBM e pela Universidade de Georgetown. Foram feitos investimentos altos nesse desenvolvimento até o relatório "ALPAC" "(Language and Machines-Computation in Translation and Linguistic Automated Language Processing Advisory Committee, National Academy of Sciences"), encomendado pelo governo americano a Bar-Hillel, em 1968. Este relatório causou um corte drástico nas atividades da tradução automática,mas recomendou esforços nas pesquisas básicas da descrição de línguas, na computação e na própria LC.

Estes esforços começaram lentamente em diversas universidades da Europa e dos Estados Unidos. Ao mesmo tempo, surgiram outras aplicações possíveis do processamento de textos em línguas naturais tais como à indexação automáti-

ca, ao ensino programado, à comunicação homem-máquina em geral, à automatização de escritórios, etc. (Haller, 1983).

No Brasil, foram poucas até hoje as pesquisas no ramo da LC as primeiras tentativas são do ITA numa pesquisa sobre a entropia da língua portuguesa de autoria de Maria Tereza Biderman e Paltônio Daun Fraga (Biderman, 1977). Recentemente, surgiu uma equipe na PUC, do Rio de Janeiro, como conseqüência de um seminário de Vitorino Ruas e de Andreewsky, da Universidade Paris-Sud, que pretende adaptar o sistema SPIRIT descrito na seção 3.4 ao português, (Andreewsky, 1983).

Na Universidade de Brasília, está sendo desenvolvido, desde 1980, um programa experimental de análise automática da língua portuguesa. Existe uma versão preliminar, em constante aperfeiçoamento, em linguagem COBOL, no sistema Borroughs 6700, do CPD da UnB. Calaborou neste programa com Haller uma equipe de alunos da pós-graduação em lingüística, da pós-graduação em biblioteconomia e do Centro de Processamento de Dados. Na seção 3.6 descreveremos a estrutura e funcionamneto deste sistema denominado LINGA ("Linguistic Analysis"). Este sistema foi criado e desenvolvido para fins de pesquisa, sem deixar de ter em vista as possíveis aplicações práticas acima mencionadas. Cabe ressaltar que a maior ênfase nas pesquisas foi dedicada à análise de textos para a possível definição e extração de seus descritores simples e/ou compostos.

O contexto da pesquisa que aqui apresentamos, envolve, aparentemente, campos distintos do conhecimento humano: a LC e a gramática gerativa. Por LC entendemos a própria lingüística que tem como objeto o estudo da estrutura de uma língua dada, auxiliada neste intento pelos avanços científicos e tecnológicos proporcionados pelo computador que, em última análise, é um sofisticado instrumento de trabalho a serviço do homem. No nosso caso específico, interessa-nos analisar os avanços da LC no estrangeiro, sobretudo os trabalhos de Fischer (1981), de Haller (1982), de Andreewsky em colaboração com P. Binquet, F. Debili, C. Fluhur e Ponderoux (1982) e de Dillon e Gray (1983) que desenvolveram sistemas de análises lingüísticas de línguas diferentes e com diferentes objetivos. Interessam-nos, também, as tentativas de análise do sistema lingüístico da língua portuguesa já realizadas no Brasil: a proposta de Andreewsky (1982) de adaptar o sistema SPIRIT à língua portuguesa e, sobretudo interessa-nos especificamente, apresentar o sistema LINGA acima mencionado.

Quanto à parte lingüística propriamente dita e que fundamenta especificamente esta pesquisa, seguimos as propostas da gramática gerativa e discutimos as várias abordagens que o sintagma nominal sofreu com a evolução da teoria gerativa para, então, podermos discutir a possibilidade do desenvolvimento de um algoritmo a partir da regra do SN português.

Evidentemente, a discussão sobre esse tema no Capítulo 4, nos indicará a necessidade ou não do desenvolvimento e implementação desse algoritmo no sistema LINGA.

CAPÍTULO II

A LINGUÍSTICA COMPUTACIONAL

A "lingüística computacional", tradução do inglês "computational linguistic", é um conceito relativamente antigo, de cerca de 40 anos, que engloba praticamente tudo o que se fez até hoje com língua natural e computador. Como as pesquisas na área da língua portuguesa são reduzidas e, assim mesmo, muito diversificadas, esta denominação pode ser ser usada para designar a lingüística matemática, a lingüística lógica, a lingüística mecânica, a lingüística automática, a lingüística algorítmica (Rosa, 1982) etc., desde que se faça uso de um computador. Mais recentemente, surgiram novas denominações influenciadas pela teoria da ciência da computação, tal como enganharia lingüística (Haller,1983).

Um pressuposto da teoria é que todo computador que executa um programa está testando um modelo da realidade feito pelo analista e pelo programador. Em síntese, o computador deve substituir parcial ou totalmente alguma capacidade humana, de modo que o produto ou os serviços oferecidos sejam aceitos pelos usuários como iguais ou superiores

aos produzidos pelos braços ou cérebros. Só assim, uma parte da ciência da linguagem (ou seria tecnologia como se interroga Haller) poderá ser denominada de "lingüística computacional".

O grande suporte científico que a LC pode oferecer à lingüística propriamente dita é a capacidade de formular e testar as regras do sistema lingüístico numa grande massa de dados num espaço de tempo relativamente pequeno. O computador poderá testar as implicações das regras lingüísticas com o texto de um jornal ou livro, uma vez que haja disponibilidade de textos em meios eletrônicos. Daí a importância do desenvolvimento desta ciência para se poder fazer frente à enorme massa de informações que já se está apresentando hoje e que cresce a cada dia que passa.

A contribuição da ciência de informação, LC e da inteligência artificial está abrindo horizontes nunca imaginados para a compreensão da estrutura dos sistemas lingüísticos.

A armazenagem e recuperação da informação podem melhorar nossa compreensão de como as pessoas organizam e usam a informação em língua natural e de como essas informações podem constituir-se numa base de dados computadorizados e, assim, serem recuperados ou serem submetidos a um processamento criando-se um sistema especializado que permita a derivação e sintetização do mesmo, segundo Walker (1981:349).

A LC pode contribuir para o desenvolvimento de novos tipos de tecnologia de recuperação da informação e assim facilitar a comunicação efetiva da informação desejada entre criadores e usuários humanos. Walker (1981:350) acentua que a LC é relevante de duas maneiras:

> First, it can provide some practical techniques for interacting with and controlling the operations of a computer through natural language. By making access to computer-based data and text files conversational, we can both increase their utility and make it easier to observe how people work with them. Second, it addresses the general issue of communication in natural language: recent research has made it clear how complex the processes associated with human understanding really are and how much more there is that we need to know.

O objetivo de facilitar a comunicação em língua natural entre seres humanos e computadores complementa a preocupação mais geral da LC com a modelação da estrutura e uso da língua. O objetivo central da LC "is to identify and formalize all the many complex factors entailed in human communication" conforme Walker (1981:351). Assim, a descrição de uma sentença fornece um conjunto de análise ou descrições estruturais que indicam a maneira pela qual uma seqüência de palavras pode ser agrupada para formar uma interpretação sintaticamente válida. No caso específico desta dissertação, interessa-nos a estrutura do SN português pela razão de a categoria N ser a que contém maior informação semântica sobre um texto, conforme Biderman (1977:30-38).

A compreensão de como as palavras são combinadas em frases e orações exige regras de composição que estabelecem suas relações umas com as outras, no contexto de uma

interpretação semântica apropriada. Por exemplo, na seguinte seqüência de palavras "solicitação de aluguel de apartamento para aluno de graduação" o significado se altera e se especifica com o acréscimo sucessivo de palavras para formar uma seqüência maior em torno de seu núcleo. No presente estágio, o sistema LINGA examina a contexto morfossintático de cada categoria. Num próximo estágio o sistema estará apto a analisar a estrutura dos constituintes imediatos da oração. Este último aspecto já está em fase de estudos e implementação por Carlos Alberto de Oliveira em seu projeto de pesquisa para doutoramento apresentado ao Instituto Nacional de Pesquisas Espaciais-INPE, de São José dos Campos, SP.

Procurando responder ao crescente interesse despertado pela LC, Walker (1981:351) chama a atenção para o relacionamento entre palavras e frases:

> it has become increasingly clear that it is necessary to incorporate into a computational model of language understanding not only general information about the relationship between words, phrases and sentences and the world, but, in addition, specific knowledge about a particular domain of context of application.

Percebe-se pela extensão e complexidade da tarefa a ser feita que há ainda um difícil caminho a ser percorrido até se alcançar resultados satisfatórios na análise da estrutura de uma língua dada.

Segundo esse autor, tem havido poucas tentativas de analisar textos de modo a derivar sua estrutura. Sager e

Schank, citados por Walker, (1981:352) desenvolveram sistemas experimentais para a elaboração de resumos de artigos científicos ou artigos de jornais. São protótipos experimentais que demonstram "the progress toward the goal of being able to process a text automatically to derive a representation of its content". A LC já está começando a abordar estes problemas e a ajuda essencial da pesquisa em inteligência artificial será fundamental para sua solução.

2.1 A lingüística computacional nos países desenvolvidos

Atualmente, existem várias pesquisas em desenvolvimento em universidades, tanto na Europa e nos Estados Unidos, como na Rússia (Apresjan, 1980) e na China segundo Haller (1983). Alguns destes sistemas desenvolvidos já encontraram sua aplicação na indústria ou em outros campos de aplicação. Os únicos centros conhecidos na aplicação de sistemas de análise automática são as Forças Armadas dos Estados Unidos, que há muito usam um sistema de tradução automática da língua russa para o inglês, e a Comissão da Comunidade Européia, onde tradutores usam textos traduzidos pela máquina como apoio a seu trabalho.

O processamento automático de textos, gradativamente começa a ocupar parte considerável de palestras e gru

pos de trabalhos em congressos de lingüística aplicada, como o "AILA", projeto da "Quinta Geração" do Japão, o Projeto de Palo Alto, na Califórnia, EUA, para onde foram lingüistas como Jackendoff e especialistas em inteligência artificial com Koy, Winograd, Woods, etc. Revistas americanas como The Journal of the ACM e outras especializadas em Biblioteconomia concedem um espaço crescente à LC, conforme Haller (1983).

2.2 A lingüística computacional em países em vias de desenvolvimento

Segundo Haller (1983) três pontos são importantes quando se considera a viabilidade de pesquisas na LC em países em vias de desenvolvimento: investimentos baixos; mão-de-obra disponível, interdisciplinaridade com a informática e biblioteconomia, por exemplo, como fomento ao crescimento científico.

Como se trata de uma "tecnologia de ponta" que pode ser desenvolvida com poucos recursos, é de suma importância dominá-la como própria para que não seja necessário adquiri-la a peso de ouro e, ainda assim, sem poder adaptar a "caixa preta" às próprias necessidades e às peculiaridades específicas de cada sistema lingüístico.

2.3 A lingüística computacional aplicada à língua portuguesa

As primeiras pesquisas da aplicação da LC à língua portuguesa datam dos anos 60, com trabalhos feitos no Instituto Tecnológico da Aeronáutica para determinar a entropia da nossa língua (Biderman, 1977: 30). O ITA desenvolveu ainda a pesquisa viabilidade de estudos literários via computador-estatística e estilo, trabalho que possui fundamentalmente o mérito do pioneirismo. Esse trabalho pioneiro contém ótimos resultados que interessam não só à estatística literária, mas, também, à informática da língua portuguesa. Os pesquisadores estudam a distribuição das classes de palavras e dos símbolos lingüísticos (sinais de pontuação) em obras de Alencar, Machado de Assis, Aluísio de Azevedo e Jorge Amado.

Entre os estudos estatístico-literários mais antigos, situa-se a Análise computacional de Fernando Pessoa (ensaios de estatística lexical), tese de doutoramento defendida na Universidade de São Paulo, por Biderman, em 1969. Segundo a autora o "exame das listas de frequências" de todo o vocabulário utilizado na obra de Fernando Pessoa evidenciou que, ao nível das altas freqüências léxicas (palavras de valor instrumental como preposição, artigos, palavras semiplenas e/ou semigramaticais como pronomes e advérbios), não existia diferença sensível entre os quatro heterônimos do autor analisado. Só aparece diferença significativa nas classes de palavras referentes ao universo externo à língua, ou seja: substantivos, adjetivos e verbos" (cf. Biderman 1977:33-34). Essas palavras de significação plena correspondem a 61,76% da obra de Fernando Pessoa.

Em França, na Universidade de Toulouse, a equipe do Prof. Jean Roche, montou um pequeno centro de estudos do vocabulário português, que vem funcionando desde 1965 (Biderman, 1977). No estudo Du Vocabulaire Poétique Brésilien, Roche utilizou vários "index verborum" de poetas brasileiros, representantes de escolas diferentes, procurando distinguir estilos individuais dos diferentes poetas estudados e maneiras literárias de escolas.

A caracterização do estilo de cada poeta se verifica no domínio das palavras de baixa freqüência. Daí a importância dessas para a indexação. Roche tentou estabelecer contrastes entre a combinatória das classes substantivo/verbo; substantivo/adjetivo; verbo/advérbio. Há um aumento constante na razão fundamental substantivo/verbo - base e medida do dinamismo de um texto - à proporção da evolução do romantismo aos contemporâneos. João Cabral de Melo Neto, para cada dois substantivos, emprega pelo menos um verbo. Daí o dinamismo revelado por sua poesia. Carlos Drummond de Andrade não se alinha a nenhuma tendência, emprega um máximo de verbos e um mínimo de adjetivos, conforme as conclusões da análise de Roche.

Biderman, em sua tese de livre-docência defendida na Universidade de São Paulo em 1974 - A categoria do gênero -, estuda a permanência das formas categoriais do gênero nas cinco principais línguas românicas: português, espanhol, francês, italiano e romeno. Trata-se de um estudo teórico e quantitativo sobre a origem e a evolução da categoria gê-

nero, desde o latim até os tempos modernos, incluindo um estudo estatístico detalhado sobre as formas de masculino x feminino das cinco línguas referidas (op. cit. 35).

O Frequency Dictionary of Portuguese Words (FDPW), executado por John Duncan, que o apresentou como tese de doutoramento na Stanford University, USA, em 1972, divide-se em duas partes: a) a primeira relaciona as palavras de um corpus de 5.142 palavras com ocorrência superior a quatro, em ordem alfabética; b) a segunda parte reproduz as mesmas palavras, alistadas em ordem decrescente, relativamente aos parâmetros de uso, freqüência e dispersão (Biderman, 1977:35).

Este estudo mostra que as 100 palavras mais freqüentes da língua compõem 61,98% do corpus escolhido. As 1000 palavras mais freqüentes correspondem a 84,57% do corpus escolhido. Portanto, com as 1000 palavras mais freqüentes do português temos praticamente o material lexical essencial para qualquer tipo de comunicação.

Uma gramática transformacional bilíngüe (português-inglês) gerada automaticamente pelo computador: A Computer Validated Portuguese to English Transformational Grammar, de James Wyatt, da University of Texas at Arlington não foi bem sucedida porque gerava resultados inaceitáveis conforme Biderman (1977).

Paltônio Daun Fraga desenvolveu um modelo de análise automática do português com vistas à tradução automáti

ca, já com resultados satisfatórios, conforme Biderman (1977:38). Esse modelo realiza a segmentação das formas; identifica os morfes, aplica as regras da gramática e efetua a transferência dos valores das variáveis morfológicas e/ou sintáticas, definidas por padrões, modelos, ou formatos chamados sintáticos.

Como se percebe pelos modelos de análise da língua portuguesa, referidos em Biderman (1977), estamos apenas no limiar de uma nova ciência que abrange, não somente a tradução automática, mas também se dedica a objetivos tais como o ensino de línguas estrangeiras (Malmberg, 1974:234—235), documentação e informação, elaboração de "index verborum" e thesauri, de dicionários de tipo especial, e ainda simulação dos mecanismos da inteligência humana no que toca ao processamento da linguagem.

CAPÍTULO III

SISTEMAS AUTOMÁTICOS DE ANÁLISE LINGÜÍSTICA

Descreveremos a seguir cinco sistemas diferentes de análise lingüística através do computador tendo como base as estruturas da língua alemã (Sistema SATAN e CONDOR); da língua francesa (Sistema SPIRIT); da língua inglesa (Sistema FASIT e método PARSING) e da língua portuguesa (Sistema LINGA).

3.1 O método do PARSING

Segundo Haller (1983), o método usado nas primeiras tentativas da análise automática de textos foi o PARSING direto, que identificava as palavras no dicionário, e, depois de atribuídas todas as funções sintáticas possíveis, criava todas as cadeias que resultavam das combinações dessas funções sintáticas. Cada cadeia analisada era submetida à correção por uma gramática existente no computador.

Além de ser pouco econômico, o método PARSING implicava uma grande dificuldade de formulação do algoritmo e

do programa, já que muitas vezes era necessário rever posições assumidas como definitivas.

3.2 Sistema SATAN

O sistema SATAN - "Sistema de análise automática de textos de Saarbrücken" (Saarbrücken Automatische Textanalyses) - da Universidade de Saarbrücken, para a análise de textos em língua alemã, descrito em Zimmermann (1980), propõe uma classificação das palavras segundo um esquema muito detalhado de mais de 150 categorias sintáticas. Como primeiro passo de análise nesse sistema são tratados os elementos gramaticais do sistema lingüístico que, no caso do sistema LINGA, constam da tabela FDIC. Através desses elementos são reduzidas as homografias de elementos como as palavras "como", "que", "o", "a", etc.

O processo de análise sintática da palavra no SATAN consiste no exame de seus vizinhos à esquerda e à direita. É um processo de exclusão que objetiva encontrar a função correta da palavra. A elaboração das regras morfossintáticas num sistema automático é um processo muito difícil e, no caso das homografias, se as regras são muito abrangentes, sobram muitas delas, em prejuízo da correção sintática; se são muito detalhadas, pode acontecer que se exclua uma seqüência de palavras permitida. Para resolver este problema, no SATAN atribui-se uma "probabilidade" a cada par de classe que se estabelece depois de análises estatísticas de textos. É um método empírico que no

sistema SPIRIT é mais aperfeiçoado, como veremos na seção 3.4 desta dissertação.

3.3 O sistema CONDOR

O objetivo geral do CONDOR ("Comunication in Natural Language with Dialogue-Oriented Retrieval") é o processamento de dados estruturados e não estruturados dos sistemas de documentação e informação cujas estruturas inerentes precisam ser previamente determinados por análise lingüística. Devido ao fato de existirem vários tipos e classes de usuários e também ao formato dos dados que precisam ser recuperados é necessário que todos os diálogos entre o usuário e o sistema de informação sejam possíveis tanto em linguagem formal de comando quanto em língua natural, conforme Fischer (1981:179):

> to avoid bottlenecks in data colection and the manual intelectual analysis of textual data it is essencial to develop ways and means of automating this operations, i.e., methods of linguistics analysis and texts classifications

Percebe-se, portanto, que a partir da análise lingüística, a função do CONDOR é permitir, sem dificuldades, a comunicação interativa com especialistas em processamento de dados e com qualquer outra área científica, bem como com usuários amadores.

O sistema CONDOR teve seu desenvolvimento efetuado em duas etapas, em Munique. A 1ª etapa, de 1973 a 1977,

prendeu-se ao desenvolvimento de modelos básicos, métodos e princípios de projetos para a concretização de futuros sistemas de base de dados para a recuperação de informações. A 2ª etapa iniciou-se em 1978 com a reformulação do projeto e início de testes-piloto, conforme Fischer (1981).

A análise automática* no sistema CONDOR se processa basicamente obedecendo à estrutura morfossintática da língua. Para os substantivos e verbos, que constituem as categorias léxicais, construiu-se um algoritmo que a partir das desinências indica a classe a que pertencem as palavras. Para a indicação de preposição, pronomes, artigos, conjunções, advérbios e também os verbos anômolos e auxiliares compilou-se um dicionário fixo com cerca de 800 entradas, uma vez que estas classes de palavras, com exceção dos verbos, possuem estrutura fixa. Assim, obedecendo a informações morfografemáticas, determina-se a função sintática ocupada pelas palavras.

3.3.1 O sistema de gerência de dados do CONDOR

O sistema CONDOR (Fischer 1981:181) desenvolveu seu próprio sistema de gerência de dados (Data Management Systems-DMS), que faculta o armazenamento, gerência e recuperação de dados com várias estruturas e extensão arbitrárias. É possível, por exemplo, dividir textos em capítulos, em parágrafos, em sentenças e mesmo em palavras isoladas e assim serem recuperadas.

3.3.2 Análise lingüística do CONDOR

A análise lingüística no sistema CONDOR tem importância fundamental e um duplo propósito:

a) determinar descritores potenciais no texto;

b) determinar relações sintáticas e semânticas entre descritores.

A análise, assim proposta em Fischer (1981 : 182), objetiva reduzir as palavras às suas raízes, de modo que a classificação temática do arquivo de dados não contenha termos idênticos ou similares pertencentes a grupos diferentes. Assim procedendo, todos os descritores de um texto estarão disponíveis. O sistema objetiva ainda a criação de opções unificadas para a representação e recuperação de palavras compostas, de sintagmas nominais e de sintagmas verbais complexos.

A análise das propriedades morfológicas de uma palavra em um texto e, por sua vez, na oração é significativa; porém, para estabelecer o seu peso lingüístico, é necessário determinar as relações intra-oracionais entre seu descritor e outros descritores, o que é realizado pela análise da estrutura da oração. As relações semânticas entre as palavras são analisadas no sistema CONDOR, a partir do dicionário de função das palavras com cerca de 800 entradas. Os resultados dessa análise lingüística são usados de acordo com a aplicação determinada do sistema e, também, de acordo com as exigências especificadas pelo usuário.

Segundo a análise lingüística operada pelo sistema CONDOR, a cada objeto de análise é atribuída uma lista de descritores com um peso lingüístico* determinado. De acordo com a freqüência em que ocorrem, pode inferir-se a possibilidade de relações temáticas entre eles. Essas relações temáticas constituem o campo semântico específico do objeto (texto) analisado.

O sistema CONDOR desempenha as seguintes operações de processamento de dados:

a) armazenamento da estrutura de informação desejada;

b) armazenamento do texto original;

c) armazenamento de dados estruturais em listas invertidas junto com as proposições dos dados do texto obtidos por análises lingüísticas;

d) armazenamento das raízes com seus pesos lingüísticos específicos;

e) armazenamento de raízes no thesaurus específico.

A recuperação de dados no sistema CONDOR obedece às seguintes operações:

a) uso de uma linguagem formal de comando para encontrar dados estruturados e proposições no texto;

b) uso de formas de busca auto-definidas para se encontrar dados estruturados e passagens do texto;

c) uso do thesaurus de raízes e cadeias de classificação para objetos desejados (busca superficial);

d) uso de resultados totais nas análises lingüísticas para a busca de precisão dentro dos objetos desejados;

e) "polimento" dos resultados para permitir a retomada de busca por qualquer método desejado.

A recuperação no sistema CONDOR pode ser efetuada em linguagem natural através de orações interrogativas. É a busca superficial*. Essas orações são submetidas a uma análise lingüística como se fossem o objeto de análise. Encontrando-se as raízes das palavras das perguntas, o usuário* poderá proceder à busca de precisão*, manipulando as chaves de busca do sistema, conforme a prioridade desejada.

Ao se completar a busca superficial, ou a busca de dados estruturados, um certo número de arquivos dos objetos estarão disponíveis ao usuário. Para permitir uma redução desse conjunto de objetos, a busca de precisão, que opera exclusivamente com os resultados de análise lingüística, foi introduzida no sistema CONDOR. O objeto dessa busca de precisão é oferecer partes selecionadas do texto que tratam do problema, definido na pergunta do usuário. Em contraste com a busca superficial, a busca da precisão usa não só as raízes de palavras das perguntas, mas também as relações lingüísticas existentes entre as chaves de busca que permitem a comparação entre as estruturas das perguntas e as estruturas do texto. A busca de precisão pára a nível de oração e depende da qualidade dos resultados da análise lingüística.

Aperfeiçoando-se a análise lingüística e encontrando-se uma forma unificada de representação para todos os dados estruturados e não estruturados, o modo de realização de sistemas de recuperação de informações poderá ser finalmente liberado, conforme Fischer (1981:186).

Como se percebe, o sistema CONDOR é um sistema complexo que tenta combinar a análise lingüística de textos com um processamento de recuperação de informações muito sofisticado. Talvez pelo alto grau de complexidade, o sistema CONDOR ainda não parece adequado para aplicação a nível comercial, constituindo-se, no entanto, num excelente laboratório de pesquisas lingüísticas aplicadas à ciência da informática.

3.4 O sistema SPIRIT

O sistema SPIRIT é um sistema de indexação automática baseado em métodos lingüísticos e estatísticos com o objetivo de processar documentos em língua natural. Este sistema foi desenvolvido por Andreewsky, em colaboração com P. Binquet, F. Debili, C. Fluhr e B. Ponderoux, do "Centre National de la Recherche Scientifique" (CNRS) francês.

O sistema SPIRIT, segundo Andreewsky (1982), permite o armazenamento e a interrogação em língua natural e, com os tratamentos lingüísticos a todos os níveis dos textos introduzidos, aliados a tratamento estatístico, permite ainda a realização de uma indexação automática ponderada de

documentos. Define-se aqui a indexação automática como a técnica de processamento eletrônico de documentos que visa à recuperação dos mesmos a partir de informações relativas ao seu conteúdo. Trata-se, mais especificamente, de obter os documentos que contêm o maior número de informações relativas a uma dada pergunta do usuário.

3.4.1 Componentes do sistema

O sistema SPIRIT consiste nos componentes seguintes:

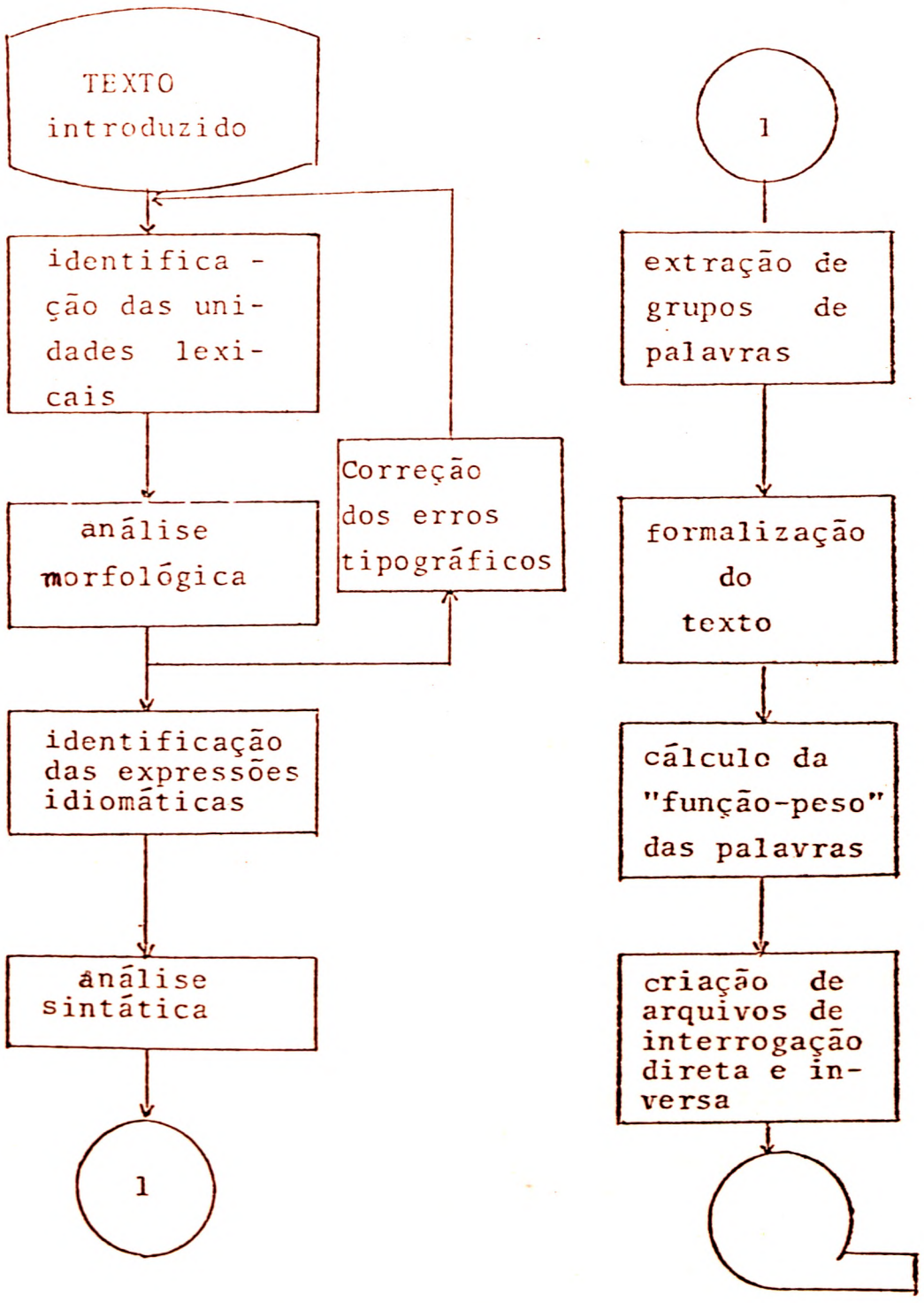
1. Um dicionário: que permite a análise morfológica dos textos. Em particular, permite o reconhecimento da sinonímia, das variações paradigmáticas de uma palavra. Fornece, além disso, todos os valores gramaticais de uma palavra.

2. Algoritmo de análise sintática: determina a categoria correta no texto de uma palavra ambígua como "para", resolvendo fenômenos da homografia.

3. Algoritmo de análise semântica: No sistema SPIRIT a análise semântica reduz-se à identificação correta da relação palavra-designado, que se tenta determinar em função do contexto. (Ver fluxograma abaixo, cf. Andreewsky, 1982).

QUADRO 3.0

FLUXOGRAMA DO SPIRIT



(Cf. Andreewsky, 1982)

3.4.2 O método de aprendizado do sistema SPIRIT

Através desse método o sistema SPIRIT "tenta reproduzir automaticamente os mecanismos que se supõe serem criados no cérebro humano, quando do aprendizado de uma língua natural" (Andreewsky, 1982). Permite decidir se uma frase é correta ou não; ambígua ou não.

O método inicialmente define um conjunto de categorias gramaticais. Na versão francesa, há 176 categorias gramaticais, o que bem dá a dimensão de sua complexidade e conseqüente dificuldade.

O método funciona fornecendo-se ao computador um texto de aprendizado com cerca de 5000 palavras, com todas as indicações gramaticais. A seguir constrói-se um dicionário de acúmulo onde as palavras do texto de aprendizado são classificadas por ordem alfabética e seguidas de todas as categorias em que aparecem no dito texto. Num terceiro passo, constrói-se um texto ambíguo a partir de texto de aprendizado e compara-se, enfim, o texto ambíguo com o texto de aprendizado e obtêm-se automaticamente as regras corretas de precedência entre os itens lexicais (ponderadas por freqüência ou não) binárias, ternárias, etc.

O sistema SPIRIT fornece ao computador análises semânticas corretas pelos chamados filtros lingüísticos*, que são os elementos gramaticais do sistema lingüístico.

A entropia* de uma palavra normalizada com respeito a um conjunto de documentos é uma quantidade destinada a avaliar o caráter discriminativo dessa palavra, no sentido de que quanto mais sua entropia é baixa mais informativa é a palavra.

3.4.3 A adaptação do SPIRIT à língua portuguesa

Constituem aspectos essenciais, para se proceder a uma adaptação da língua a uma outra língua, a análise das ambigüidades existentes nessa língua assim como a possiblidade de levantá-las por métodos de aprendizado que levem em conta as propriedades posicionais das classes de palavras.

Essa adaptação ao inglês e ao espanhol revelou-se perfeitamente factível. Provavelmente, os métodos estatísticos e de filtragem dos quais se serve o sistema SPIRIT para o francês poderão ser utilizados para o português, segundo Andreewsky.

3.4.4 Análise sintática no SPIRIT

A análise sintática no SPIRIT é fundamental para levantar todas as ambigüidades existentes numa língua, por causa do emprego de uma mesma palavra em funções distintas. Como o computador encontrará no dicionário duas ou mais categorias para tal palavra, será necessário levantar a ambigüidade procurando-se determinar a categoria gramatical cor

reta da palavra que corresponde ao seu emprego no texto. As matrizes de procedência indicam no SPIRIT a relação de vizinhança das categorias gramaticais e assim desfazem as ambigüidades existentes.

3.5 O sistema FASIT

3.5.1 Objetivo do sistema FASIT

O objetivo da indexação automática é obter uma representação compacta adequada de um documento para a recuperação. O FASIT (Fully Automatic Syntactically Based Indexing System) identifica o conteúdo de unidades textuais sem fazer uma análise gramatical completa e sem utilizar critérios semânticos e agrupa estas unidades em conjuntos de termos quase sinônimos.

Basicamente, a indexação automática representa um documento usando critérios lingüísticos ou estatísticos para selecionar palavras ou frases significativas do texto.

3.5.2 Indexação automática no FASIT

A dimensão maior na qual os sistemas de indexação automática diferem é o seu grau de sofisticação lingüística. Em contraste com sistemas de indexação baseados na derivação, que têm como base a palavra reduzida a sua raiz, a busca de uma solução para ambigüidade/sinonímia é tentada através de uma análise lingüística mais profunda do texto no FASIT.

O sistema FASIT tenta obter os resultados de uma análise gramatical com o objetivo de indexação decompondo o texto em frases adequadas, sem uma análise gramatical complexa e agrupando essas unidades em conjunto de termos quase sinônimos, sem componentes semânticos. No entanto, é improvável que se alcancem melhores resultados em um sistema de indexação automática sem uma profunda análise sintática ou mesmo semântica, como se percebe pela análise do sistema CONDOR, LINGA e SPIRIT.

3.5.3 Componentes do FASIT

O FASIT baseia-se na idéia de que palavras ou frases significativas pertencem a certas categorias sintáticas ou combinações de categorias. Após reduzir as palavras do texto a categorias, o sistema seleciona conceitos baseados em padrões pré-definidos de categorias. Então, reduz as variações destes conceitos a uma forma predominante para o agrupamento.

A indexação pelo FASIT consiste em duas operações principais: a primeira é a seleção de conceitos; a segunda é o agrupamento desses conceitos.

A seleção de conceitos processa-se pela atribuição de categorias sintáticas a palavras utilizando-se de um dicionário de exceções de palavras e de um dicionário de sufixos: processa-se ainda pela eliminação de ambigüidades de palavras com categorias múltiplas realizadas através das desinências das palavras antes e depois da palavra ambígua.

A atribuição de categorias sintáticas apropriadas às palavras começa com o dicionário de exceções que contém cinco tipos de palavras, a saber:

1. Classes fechadas, tais como os artigos; conjunções; preposições; pronomes; verbos auxiliares, que são poucos em número, mas ocorrem com muita freqüência, e ainda a pontuação;

2. Palavras de alta freqüência e aplicação geral cujas terminações (desinências irregulares) não se adaptam às categorias determinadas pelo dicionário de sufixos nominais e adverbiais;

3. Substantivos e adjetivos de aplicações tão gerais que se tornam inúteis para a indexação. São caracterizados como substantivos comuns e adjetivos fracos;

4. Nomes próprios;

5. Palavras específicas para uma aplicação cujo comportamento sintático está em desacordo com os padrões gerais do sistema.

Qualquer palavra, dessa forma, não encontrada no dicionário de exceções é classificada pelo seu sufixo em categoria nominal ou verbal.

A ambigüidade sintática de palavras em um ambiente de recuperação que usa partes das sentenças pode ter sua solução tratando os adjetivos-substantivos ambíguos como sendo substantivos. As ambigüidades em verbo-substantivo têm mais probabilidade de serem eliminadas com relação aos verbos.

Quando estas palavras ambíguas são classificadas como substantivos, elas são quase sempre mais significativas do que quando são usadas como verbo para a indexação. Por ex.: a luta - ele luta; a casa - ele casa.

A segunda operação do FASIT - agrupamento de conceitos - reduz os conceitos selecionados a formas canônicas para a consolidação de formas sinônimas. A criação de formas canônicas reduz substancialmente o número de conceitos únicos, mas é insuficiente ao relacionar conceitos similares em significado. Por exemplo, conceitos com mais de uma palavra como "curso de educação contínua" é um aspecto de conceito mais amplo de que "educação contínua". No FASIT, o conceito com mais de uma palavra é preservado usando mais a derivação extensiva e a superposição de conceitos. O uso de categorias sintáticas para palavras individuais e para combinar palavras em frases reduz o problema da ambigüidade. As frases do FASIT têm um alto grau de especificidade enquando que um alto nível de exaustividade na indexação é alcançado pelo agrupamento de conceitos através da combinação parcial de palavras e frases.

O referimento e aperfeiçoamento da análise, no FASIT, podem ser feitos em cada uma de suas etapas principais, mantendo ao mesmo tempo sua generalidade. O aumento do dicionário de exceções para incluir um vocabulário mais geral e especialmente verbos de alta freqüência, é um tipo de aperfeiçoamento. A crescente eliminação de ambigüidades e

um controle melhor dos agrupamentos de conceitos, deve melhorar também os resultados, concluem Dillon e Gray (1983: 107).

O FASIT é de uso geral e, uma vez implantado, é de fácil manutenção. Os resultados da recuperação indicam que a idéia básica do FASIT de que termos significativos do texto podem ser identificados e agrupados usando-se critérios sintáticos é válida.

3.5.4 Conclusão

O sistema FASIT é um sistema experimental desenvolvido como pesquisa universitária com um dicionário de freqüência e massa de dados muito reduzida. Não se tem conhecimento de uma aplicação em grande escala em sistemas de informação e recuperação. No caso do sistema LINGA, o mesmo é destinado à aplicação na recuperação e informação, possuindo um sofisticado sistema de análise lingüística.

3.6 O sistema LINGA

O projeto do sistema de análise lingüística da língua portuguesa denominado LINGA ("Linguistic Analysis") tem suas origens nas pesquisas de Haller realizadas na empresa Siemens, em Munique, Alemanha Federal (cf. Fischer, 1981 e Haller, 1981).

Sendo o objetivo central da LC identificar e formalizar todos os complexos fatores envolvidos no processo de comunicação humana, procurou-se estruturar, no sistema LINGA, os fatos gramaticais da língua portuguesa para que então possam ser utilizados para os mais variados fins, como a indexação automática, tradução automática e outros.

O sistema LINGA está estruturado em sub-rotinas que objetivam agilizar a análise lingüística de textos. Assim sendo, as regularidades e irregularidades do código linguístico são organizadas em tabelas e rotinas de tal forma que abrangem toda a estrutura gramatical da língua portuguesa nos níveis morfológico e sintático. Em princípio adotou-se a classificação das palavras proposta por Pottier (1975) e (1978) que propõe uma distribuição dos morfemas em duas classes: os morfemas lexicais ou lexemas, que integram os elementos de um conjunto inacabado e aberto, que são os nomes, adjetivos e os verbos; os morfemas gramaticais ou gramemas, que integram os elementos de um conjunto finito, fechado. São os prefixos, sufixos, desinências, artigos, preposições, conjunções, etc. O critério de Pottier se baseia não só na função distintiva, mas também no traço semântico dos morfemas.

Guéron (1982), no âmbito da teoria $\bar{X}$, propõe uma modificação da teoria lexicalista de Chomsky e Jackendoff, segundo a qual as categorias lexicais são o nome (N), o verbo (V), o adjetivo (A) e a preposição (P). Guéron distin-

gue no conjunto de traços que determinam as categorias sintáticas, duas grandes classes: os traços lexicais (N, V, A) que determinam as categorias lexicais* e os operadores, que determinam as categorias gramaticais e que são os determinantes, complementadores, etc. Propõe igualmente a existência de certas classes de itens lexicais "mistos", interpretados ao nível da forma lógica, seja como constituinte gramatical ou como constituinte lexical. As preposições nesta proposta são consideradas como classes mistas juntamente com os pronomes clíticos.

Para Guéron "les mots lexicaux et leurs projections dénotent un object (Nom), une action (Verbe), ou une propriété (Adjectif). Les mots grammaticaux, par contre, ne dénotent rien, mais déclenchent opérations interprétatives ou morphosyntaxiques. Nous appelons donc ces constituants des opérateurs". Nessa proposição, operadores são todos os constituintes ou morfemas não-lexicais. Cada operador possui um traço [+Q] que o identifica como não-lexical. Apesar das abordagens teóricas diferentes, as propostas de Pottier e de Guéron se assemelham quanto à descrição da estrutura lingüística. Dentre as palavras lexicais interessa-nos especificamente a categoria do nome para a indexação automática. Os operadores estão definidos e arrolados no FDIC (ver 3.6.2).

Seguindo os princípios da correção lingüística, rapidez na análise, economia de tempo na computação, o sistema LINGA foi estruturado basicamente a partir dos elementos gramaticais - os operadores - por se constituírem num

conjunto finito de elementos e por encontrarem seu referente no próprio sistema lingüístico (cf. Baylon, 1979:183-4). Dado este fato, procurou-se organizar os elementos gramaticais na tabela FDIC (Dicionário de Freqüência) (V. quadro 3.9) e na tabela HTAB (V. quadro 3.11) e objetivou-se estabelecer as regularidades das 33 categorias morfossintáticas estabelecidas no sistema LINGA. A organização dessas duas tabelas consumiram anos de estudo e testes durante os cursos de LC no mestrado em lingüística da UnB.

3.6.1 Análise lingüística no LINGA

A análise lingüística operada no sistema LINGA consiste em submeter as palavras de uma frase a três tipos de procedimentos analíticos básicos. Primeiro: são confrontadas com dicionários fechados para uma triagem onde são analisadas e definidas as palavras gramaticais; segundo: são confrontadas com dicionários abertos para uma triagem onde são analisadas e definidas as palavras lexicais; terceiro: após a definição das palavras gramaticais e lexicais, ocorre a definição sintática das palavras, obedecendo às configurações que cada categoria pode ter no léxico. Com estes procedimentos, o sistema LINGA objetiva cobrir a estrutura morfossintática da língua portuguesa.

Como vimos, para que seja possível a análise das palavras de um texto pelo sistema LINGA, há a necessidade de segmentá-lo em frases que, por sua vez são segmentadas em palavras que, então, sofrem os procedimentos analíticos descritos no parágrafo introdutório desta seção.

A segmentação do texto em frases e das frases em palavras pode parecer um problema desprezível para qualquer falante/ouvinte da língua natural. No entanto, tratando-se de análise de texto por computador, a situação muda de aspecto. Para o computador, cada cadeia de "bytes"* é considerada como uma possível unidade lingüística. Para o sistema LINGA, qualquer conjunto de dígitos (letras, números ou sinais gráficos), reunidos aleatoriamente será considerado pa

ra efeitos de análise. Poderão ser palavras para o sistema LINGA:

- uma cadeia aleatória de letras, por exemplo: XPTO;
- uma cadeia alfanumérica terminada em letras, por exemplo: 359A;
- uma cadeia de símbolos, por exemplo: (-/?).

Para reconhecer o limite entre uma palavra e outra, o programa considera os seguintes símbolos: . , : ; () ? " " ou um espaço em branco entre duas palavras.

Os seguintes símbolos indicam o fim de uma frase:

- ponto final (se não se tratar de um elemento da lista de abreviações; (v. quadro 3.5).
- dois pontos;
- ponto e vírgula;
- ponto de interrogação e exclamação;
- uma combinação de símbolos com mais de 28 dígitos (convencionou-se no sistema LINGA este número, porque é pouco provável que uma palavra tenha mais de 28 dígitos (letras) na língua portuguesa).

A solução para o ponto indicador de abreviação das palavras foi a inserção no sistema de uma tabela contendo a lista das abreviações mais usuais no código escrito. Assim, uma frase é analisada em sua estrutura normal e não sofre a interrupção da análise ao se defrontar com o ponto indicador de abreviação.

Para o caso das formas verbais pronominais mesoclíticas e enclíticas hifenadas, foi elaborada por Carlos

QUADRO 3.1

FORMAS PRONOMINAIS HIFENADAS

Alberto de Oliveira uma tabela com todas as terminações possíveis (ver quadro 3.1) porque essas formas não são abrangidas pela rotina da morfologia verbal (VTAB) (quadro 3.6) que, especificamente, cobre as formas verbais regulares flexionadas e não-flexionadas sem pronomes enclíticos ou mesoclíticos. O procedimento da rotina da tabela das formas verbais pronominais hifenados consiste em converter as formas hifenadas em formas não-hifenadas para que a definição morfossintática do verbo e do pronome seja possível. Por exemplo, as formas verbais hifenadas abaixo terão as seguintes formas depois de serem submetidas à rotina da tabela:

- dar-lhes-iam — dariam lhes

O sistema LINGA converte as formas hifenadas em formas enclíticas para efeitos de análise morfossintática, embora o usual sejam as formas proclíticas. Por exemplo:

contar-nos-ão contarão nos (no sistema LINGA)
nos contarão (forma usual)

Não havendo confronto com uma dessas possibilidades assinaladas na tabela das formas pronominais hifenadas, a palavra será considerada um substantivo composto em razão da presença do hífen.

No quadro 3.2 (segmentação) temos o resultado da segmentação de uma sentença processada pelo sistema LINGA.

QUADRO 3.2

SEGMENTAÇÃO

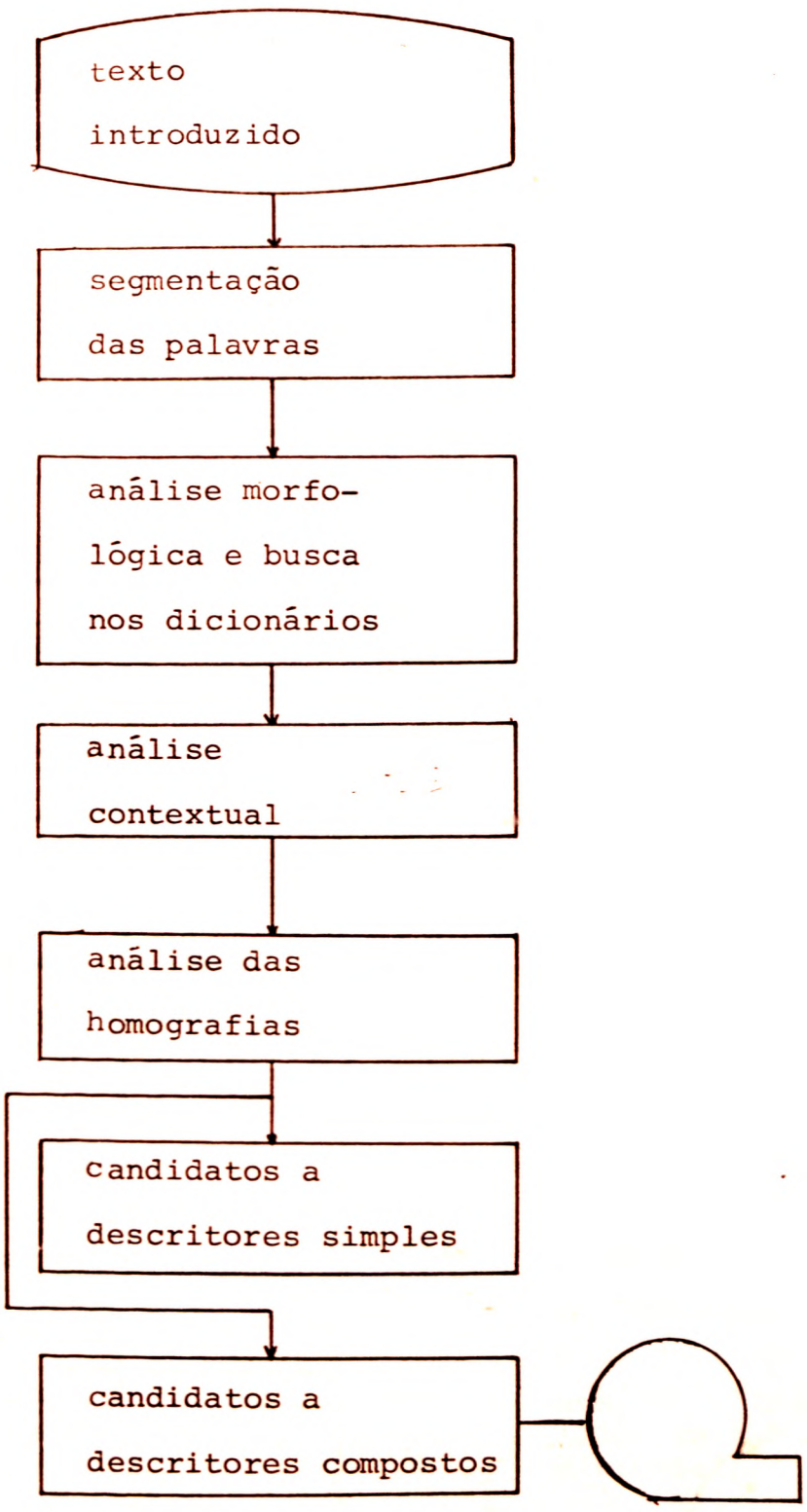
001	010	000	INTRODUCAO
002	013	000	----------
003	004	000	ESTE
004	006	000	ESTUDO
005	009	000	BASEIA-SE
006	002	000	EM
007	002	000	UM
008	008	000	TRABALHO
009	005	000	SOBRE
010	001	000	O
011	002	000	SN
012	009	000	PORTUGUES
013	005	000	FEITO
014	003	000	POR
015	007	000	BEATRIZ
016	005	000	NUNES
017	002	000	DE
018	008	000	OLIVEIRA
019	005	000	LONGO
020	001	032	,
021	003	000	UNB
022	002	000	EM
023	004	000	1981
024	001	031	.

Temos na primeira coluna, o número, em ordem crescente, da palavra segmentada. No caso deste exemplo, temos 24 segmentos. Na segunda coluna, temos o número de letras (dígitos) de cada palavra, por exemplo: a palavra "INTRODUÇÃO" tem 10 letras; os dígitos que sublinham INTRODUÇÃO são treze, etc. A terceira coluna destina-se à definição das categorias morfossintáticas, que serão discutidas a partir do quadro 3.12 No presente estágio da análise, somente as categorias 32 (vírgula) e 31 (ponto) aparecem indicadas, pois sua definição não implica análises ulteriores. Na quarta coluna, aparecem as palavras que serão submetidas aos procedimentos analíticos do LINGA.

A fim de facilitar a compreensão da seqüência das rotinas do LINGA, apresentamos a seguir o seu fluxograma (quadro 3.3) e a relação das rotinas da análise morfológica (quadro 3.4) com uma descrição sucinta de cada rotina.

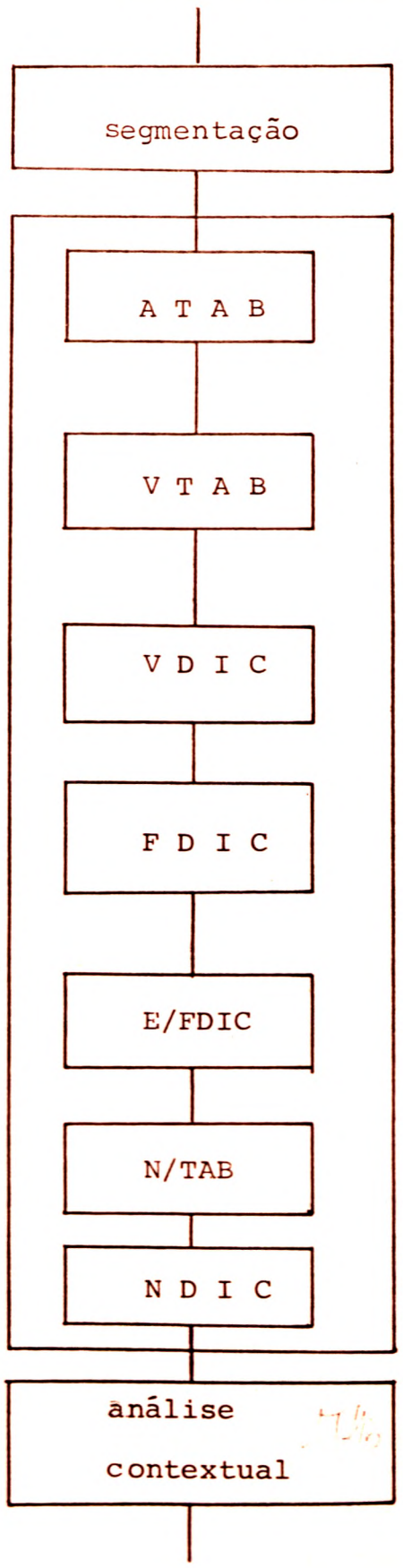
QUADRO 3.3

FLUXOGRAMA DO LINCA



QUADRO 3.4

ROTINAS DA MORFOLOGIA NOMINAL



Obs.: O quadro 3.4 (rotina da morfologia nominal) é um detalhamento da etapa da 'Análise Morfológica e busca nos dicionários' do quadro 3.3 (Fluxograma do LINGA).

3.6.1.1 Análise morfológica e busca nos dicionários

Após o processo da segmentação do texto, a frase é submetida à rotina da tabela ATAB - tabela das abreviações (quadro 3.5 organizada pelo autor desta dissertação. O objetivo desta tabela é evitar que o programa LINGA considere o ponto indicador de abreviação como um ponto final de frase. Este fato acarretaria evidentes prejuízos para a análise automática das frases.

O funcionamento desta tabela processa-se da seguinte forma: na etapa da segmentação do texto em frases, o sistema LINGA identifica como limite de frase o ponto final. Acontece que o ponto final ocorre também como indicador de abreviações. Portanto, a listagem das abreviações é necessária para que não ocorra a interrupção da análise da frase. Por exemplo, na abreviação de conforme, cf., ocorre, como também nas outras abreviações desta listagem, a presença do ponto característico. No entanto, não ocorre a interrupção da análise da frase, devido ao confronto desta abreviação com a que está armazenada na tabela. No caso de haver uma abreviação composta, tal como V.Sa. (de Vossa Senhoria) o procedimento é o mesmo daquele que descrevemos acima; apenas repete-se o procedimento de V. e Sa. como se fossem abreviações isoladas. Se não houvesse confronto com as abreviações constantes da tabela (quadro 3.5), o sistema consideraria como ponto final de frase a ocorrência do ponto de uma abreviação ainda não listada na mesma.

QUADRO 3.5

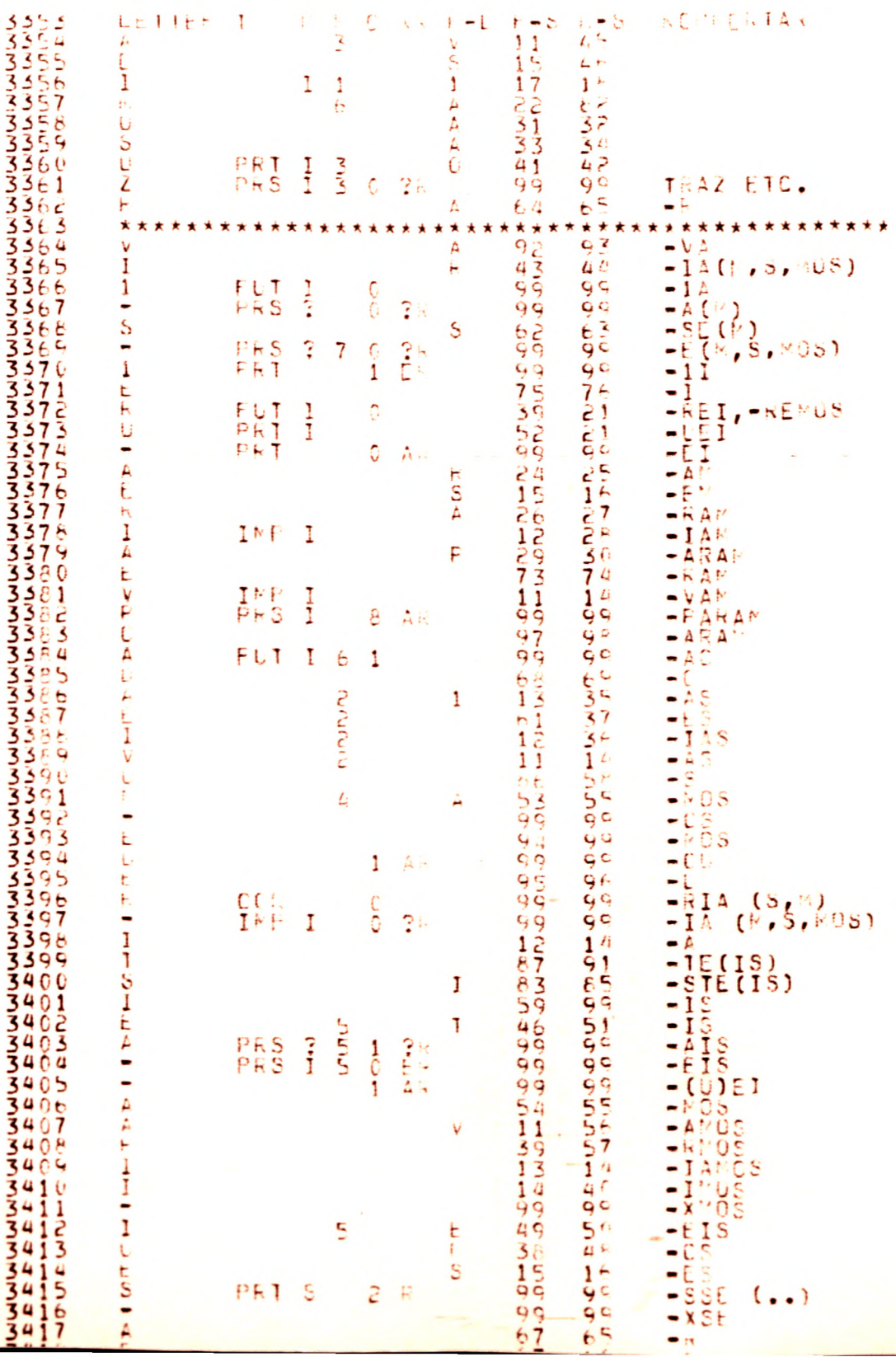
TABELA DAS ABREVIAÇÕES (ATAB)

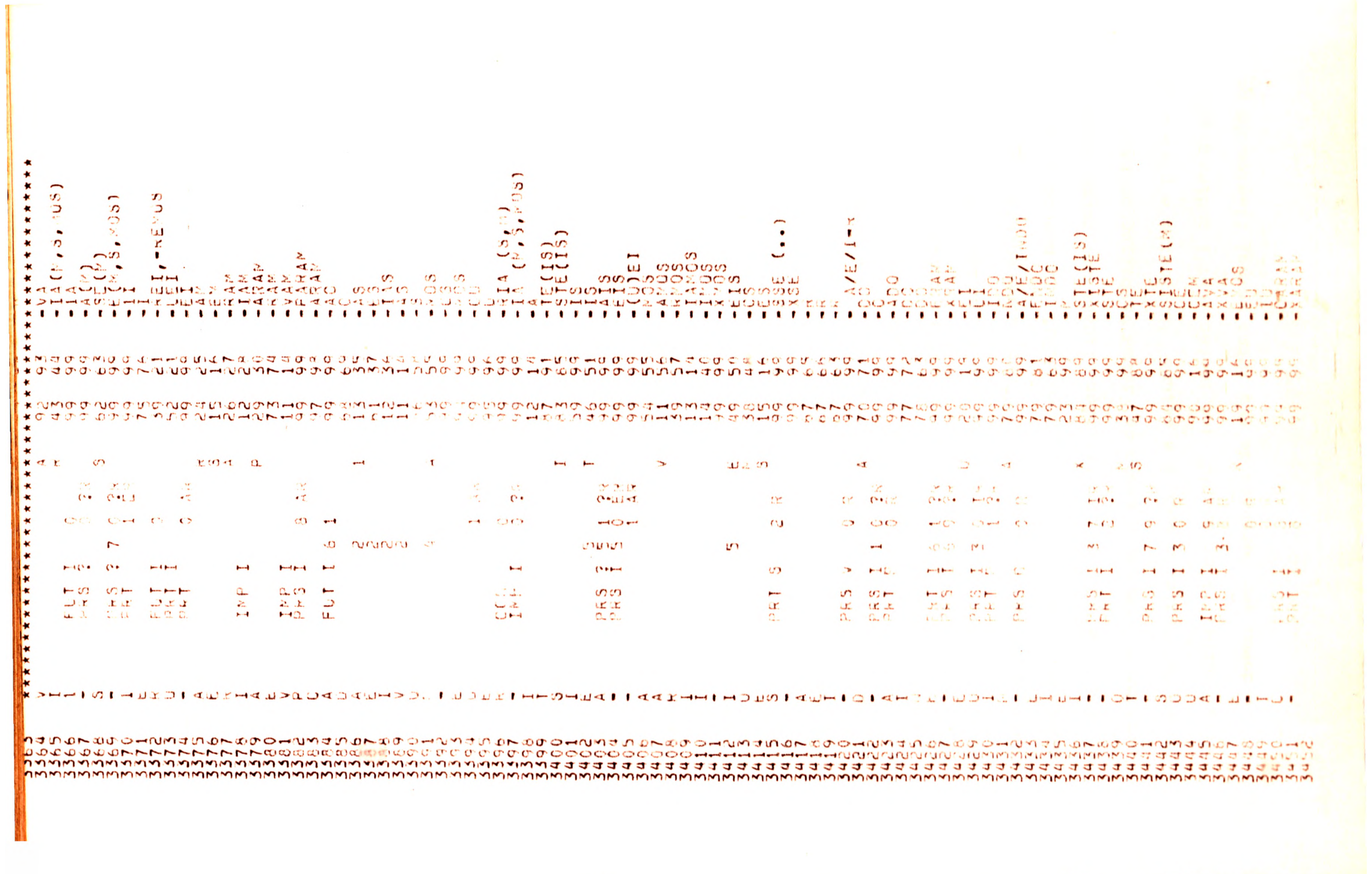
100	A	AUTOR/ALTEZA
200	AF	AFUD
300	B	BURO
400	C	CENTRO
500	CF	CONFRONTAR/CONFERIR
600	CIA	COMPANHIA
700	D	DOM/DONA
800	DL	DIGNISSIMO
900	DRA	DOUTORA
1000	DR	DOUTOR
1100	E	EMINENCIA
1200	EL	ELEMENTO
1300	ETC	ETCETERA
1400	EX	EXEMPLO
1500	EXA	EXCELENCIA
1600	EXAS	EXCELENCIAS
1700	EXMA	EXCELENTISSIMA
1800	EXMAS	EXCELENTISSIMAS
1900	EXMO	EXCELENTISSIMO
2000	EXMOS	EXCELENTISSIMOS
2100	EXT	EXTENSO
2200	F	FEMININO/FOLHA(S)/FORMA(S)
2300	FR	FRANCES
2400	G	GENERO
2500	GAL	GENERAL
2600	GEN	GENERO
2700	I	ISTO (E)
2800	IE	IBIDEM
2900	ID	IDEM
3000	ILMA	ILUSTRISSIMA
3100	ILMO	ILUSTRISSIMO
3200	IND	INDICATIVO
3300	INF	INFINITIVO
3400	INGL	INGLES
3500	INT	INTRANSITIVO
3600	IT	ITALIANO
3700	KG	KILOGRAMA
3800	KM	KILOMETRO
3900	L	LITRO
4000	M	MAJESTADE (VOSSA)/METRO
4100	MM	MERITISSIMO
4200	N	NUMERO/NORTE
4300	P	PAGINA
4400	PA	PARA
4500	PE	PERNAMBUCO
4600	PG	PAGO
4700	PP	PAGINAS
4800	PROFA	PROFESSORA
4900	PROF	PROFESSOR
5000	Q	QUEIRA (VER)
5100	QTO	QUANTO
5200	QTOS	QUANTOS
5300	REVMA	REVERENDISSIMA
5400	REVMO	REVERENDISSIMO
5500	SA	SOCIEDADE ANONIMA
5600	S	SUBSTANTIVO
5700	SS	SUA SANTIDADE
5800	SR	SENHOR
5900	SRA	SENHORA
6000	SRAS	SENHORAS
6100	SRS	SENHORES
6200	T	TRANSITIVO
6300	TB	TAMBEM
6400	V	VERBO

A tabela da morfologia verbal - VTAB - (quadro 3.6) organizada por Hans Haller, objetiva analisar as palavras a partir de suas desinências para identificar uma possível forma verbal.

QUADRO 3.6

TABELA DA MORFOLOGIA VERBAL





A fim de cobrir toda a morfologia verbal, a tabela VTAB está estruturada da seguinte forma: nas dez primeiras linhas da coluna 'LETTER' estão arroladas todas as letras que podem ocorrer em posição final nas formas verbais da língua portuguesa. Se uma palavra terminar em outra letra que não conste da lista, esta não será uma forma verbal. Identificadas as últimas letras, são examinadas as letras imediatamente precedentes às mesmas, que estão arroladas na coluna P-L. Havendo a possibilidade da ocorrência das duas letras finais de uma forma verbal, a rotina VTAB tem seqüência em seus procedimentos (exemplo: VA, As, AO, etc.). A coluna P-S indica a etapa seguinte de análise, a partir das linhas 11 a 99, para identificar a letra que poderá estar precedendo as duas últimas letras de uma forma verbal. (Observe-se que as informações codificadas nos números que aparecem na coluna P-S e N-S, nas linhas 1 a 11, referem-se a procedimentos da rotina do algoritmo que serão discutidos sucintamente adiante). A análise tem prosseguimento na coluna N-S, nas linhas 11 a 99. Esta etapa é implementada somente se for identificada a letra final de uma possível forma verbal, mas não sua precedente que está listada nas linhas de 1 a 10 na coluna P-L. O número 99 na coluna P-S significa que a análise encontrou uma possível forma verbal que deverá ser confrontada com o dicionário de verbos (VDIC - quadro 3.7). O número 99 indica na coluna N-S que não foi identificada uma forma verbal. A coluna 'T' indica o tempo da forma verbal encontrada; a coluna 'M' indica o modo e a coluna 'P' indica a pessoa da formal verbal. A coluna 'C' (do inglês 'CUT') indica quantas letras devem ser cortadas de uma forma verbal flexionada pa

QUADRO 3.7

DICIONÁRIO DE VERBOS (VDIC)

3614	ESCURECER			
3615	ESFERRAR			
3616	ESFOLTAR			
3617	ESFREITAR			
3618	ESQUECER			
3619	ESTABELECER			
3620	ESTACIAR			
3621	ESTAR			
3622	ESTEJAR	ESTAR	PRSS	
3623	ESTENDER			
3624	ESTEVER	ESTAR	PRTI	
3625	ESTIMAR			
3626	ESTIVER	ESTAR	IMPS	
3627	ESTORVAR			
3628	ESTRADAR			
3629	ESTRANHAR			
3630	ESTREAR			
3631	ESTREMECER			
3632	ESTRUTURAR			
3633	ESTUDAR			
3634	EVASIAR			
3635	EVITAR			
3636	EXAGERAR			
3637	EXAMINAR			
3638	EXAURIR			
3639	EXCEDER			
3640	EXCLAMAR			
3641	EXCLUIR			
3642	EXECUTAR			
3643	EXERCER			
3644	EXIGIR			
3645	EXISTIR			
3646	EXPERIMENTAR			
3647	EXPLICAR			
3648	EXPLICITAR			
3649	EXPLORAR			
3650	EXPOR			
3651	EXPRESSAR	EXPRIMIR	PARP	
3652	EXPRIMIR			
3653	EXUMAR			
3654	FACER	FAZER		
3655	FACILITAR			
3656	FALAR			
3657	FALIR			
3658	FALTAR			
3659	FASCINAR			
3660	FAZER			
3661	FECHAR			
3662	FERVER			
3663	FESTEJAR			
3664	FICAR			
3665	FINALIZAR			
3666	FINGIR			
3667	FIZER	FAZER	PRTI	
3668	FOCAR			
3669	FOGUER	FOCAR IR		
3670	FOR			
3671	FORMAR			
3672	FORMAR-SE			
3673	FORNECER			
3674	FOTOGRAFAR			
3675	FRACASSAR			
3676	FR	SER		
3677	FREQUENTAR			
3678	FUGIR			
3679	FUJAR	FUGIR	PRSS	
3680	FUMAR			
3681	FUNCIONAR			
3682	GANHAR			
3683	GARANTIR			
3684	GASTAR			

ra se obter sua forma infinitiva. A coluna 'XX' indica os casos de indecisão na definição da vogal temática. Fato que é solucionado com o confronto entre a forma infinitiva encontrada por VTAB com as que estão armazenadas no VDIC. Este confronto é necessário porque o programa VTAB identificaria formas como MUNDO como sendo formas verbais e apresenta infinitivos como "MUNDAR", MUNDER, MUNDIR, MUNDOR. A coluna "KOMMENTAR" apresenta as desinências verbais que são eliminadas para se conseguir a forma infinitiva.

Em síntese, esta tabela identifica, a partir das desinências verbais, o modo, o tempo, o número e a pessoa da forma verbal submetida à sua análise.

Como já vimos, após a definição da forma verbal pela tabela VTAB, esta é confrontada com o dicionário de verbos (VDIC) que tem 780 verbetes do qual o quadro 3.7 é uma amostragem. Esta tabela teve a contribuição de todos os integrantes da equipe do LINGA em sua organização, sendo que a marcação lexical dos verbos foi elaborada pelo autor desta dissertação. A tabela está descrita em detalhe em 3.6.5

A etapa seguinte da rotina do programa LINGA compreende a análise das palavras pela rotina do dicionário de freqüência (FDIC) (V. excerto no quadro 3.8). Esta etapa analisa e define as palavras gramaticais de um texto. Dada a sua importância decisiva no programa, o FDIC vem descrito na seção 3.6.2 Abaixo, temos uma amostragem da tabela.

QUADRO 3.8

DICIONÁRIO DE FREQUÊNCIA (FDIC)

		F	T	M		T	M		T	M
		A		I	/		I	/		I	/
	*FWORD	C CNP	M	G	CNP	M	G	CNP	M	G
	A	309S		F	18			29S3		F
	A*	129S3		F						
	A4	118S3		F						
	A4GUELA	118S3		F						
	A4GUELAS	116		S						
	A4GUELE	118		S						
	A4GUELES	118		S						
	A4GUILO	118		S						
	A4S	118P		F						
	ABAIXO	126		L						
	ACASO	201S		M	26		M			
	ACIMA	126		L						
	ADIANTE	126		L						
	AFINAL	126		M						
	AGORA	126		T						
	AI	126		L						
	AI1	126		L						
	AINDA	126		T						
	AMBAS	211P3		F	12P3		F			
	AMBOS	211P3		M	12P3		M			
	ALE1M	126		M						
	ALEM	126		M						
	ALI	118		L						
	ALGO	211S3		M	12S3		M			
	ALGUE1M	211S3		M	12S3		M			
	ALGUEM	211S3		M	12S3		M			
	ALGUM	211S3		M	12S3		M			
	ALGUMA	211S3		F	12S3		F			
	ALGUMAS	211P3		F	12P3		F			
	ALGUNS	211P3		M	12P3		M			
	ALIAS	126		M						
	ANTE	118								
	ANTES	218		T	26		T			
	AO	118S		M						
	AOS	118P		M						
	APENAS	126								
	APESAR	120								
	APOIS	126		T						
	APOS	126		T						
	AQUELA	211S3		F	12S3		F			
	AQUELAS	211P3		F	12P3		F			
	AQUELE	211S3		M	12S3		M			
	AQUELES	211P3		M	12P3		M			
	AQUI	126		L						
	AQUILO	112S3		N						
	AS	309P		F	29P		F	18		
	AS*	129P		F						
	ASSIM	126		M						
	ATE	218			26		M			
	ATE1	218			26		M			
	ATRA1S	126		L						
	ATRAS	126		L						
	ATRAVE1S	118								
	ATRAVES	118								
	BEM	226		M	01S		M			
	BREVE	202S3		M	26		T			
	BREVES	102S3		M						
	CA	126		D						
	CA1	126		D						
	CADA	111S3		M						
	CASO	301S		M	25			03S1PRSI		
	CEDO	203S1PRSI		26			T			
	CEM	207P			08P					
	CERCA	301S3			19		L	26		L
	CINCO	207P			08P					
	CINQUENTA	207P			08P					

Está integrado ao programa, nos moldes do FDIC, um dicionário de freqüência da língua inglesa (E/FDIC) com a finalidade de identificar eventuais ocorrências de frases da língua inglesa que possam ocorrer dentro de um texto em português.

A tabela NDIC (dicionário de formas nominais v. excerto no quadro 3.9), descrita em maior profundidade na seção 3.6.4 desta dissertação, consta de substantivos e adjetivos que oferecem dificuldades, por sua homografia; adjetivos pré-nominais; prováveis não descritores; descritores em contexto específico e palavras que podem fazer parte de um descritor composto. Abaixo, temos uma amostragem da tabela que possui mais de 400 registros. A classificação dos verbetes que aparece nas colunas 31, 33, 35, 37 e 39 foi elaborada em conjunto por Simone Bastos e o autor desta dissertação.

A tabela da morfologia nominal - NTAB (quadro 3.10) identifica a forma das desinências nominais das palavras que escaparam da análise das tabelas anteriormente referidas, ou seja: ATAB, VTAB, VDIC, FDIC, E/FDIC, NDIC. Esta tabela está descrita em maior profundidade na seção 3.6.3 Temos a seguir uma amostragem desta tabela.

QUADRO 3.9

DICIONÁRIO DE FORMAS NOMINAIS (NDIC)

2949	NOUN/ADJ-LEXEM	NEG	ADJ1		NDES	PAP (22.9.83)
2950	ABAJUR					
2951	ABANDONO	*				
2952	ABERTA		1	1		1
2953	ABERTURA	*				
2954	ACO				1	1
2955	ACORDO	*			1	1
2956	ACRESCIMO					1
2957	ADICIONAL			1		
2958	ADORAVEL		1	1		
2959	ADVERTENCIA				1	1
2960	AGENTE				1	1
2961	AGRUPAMENTO					
2962	AJUDA	*			1	1
2963	ALTO		1	1		
2964	AMBITO			1		
2965	AMOSTRA				1	1
2966	AMPLO		1	1		
2967	ANALISE	*				
2968	ANIMAL				1	1
2969	ANTIGO		1		1	
2970	APLAUDIDO		1	1		
2971	APLICABILIDADE					1
2972	APLICADA					1
2973	APLICACAO			1		
2974	APLICAVEL					1
2975	APOIO	*				
2976	APROVADA			1		
2977	AQUISICAO				1	
2978	AREA				1	
2979	ASSISTENCIA				1	1
2980	ATRASO	*				
2981	ATENTO		1	1		
2982	ATO	*				
2983	ATUAL			1		
2984	AUMENTO	*				
2985	AUSENCIA			1		
2986	AUTOR				1	
2987	AUXILIO					1
2988	BAIRRO			1		
2989	BAIXA	*	1			1
2990	BAIXO		1	1		
2991	BANCO				1	1
2992	BASEADA			1		
2993	BASICA					1
2994	BASICO			1		
2995	BASTANTE		1	1		
2996	BELO		1	1		
2997	BENEFICIARIO				1	1
2998	BOM		1	1		
2999	BRILHANTE		1		1	
3000	BUSCA	*			1	1
3001	CACA	*			1	1
3002	CALMO		1	1		
3003	CAMINHO	*		1		
3004	CAMPO				1	1
3005	CAPITULO			1		
3006	CAPTURADA					1
3007	CARTA				1	1
3008	CARTEIRA					1
3009	CASA	*				
3010	CASO	*		1		

QUADRO **3.10**

MORFOLOGIA NACIONAL (NTAB)

4100	--- ---	
4200	1SES 1S	
4300	ESES E2S	
4400	A3ES A30	
4500	A3O5 A3O	
4600	E1IS EL	
4700	ESAS E2S	
4800	FZAS EZA	
4900	OIIS OL	
5000	O5ES A3O	
5100	ONAS AO	
5200	ORAS OR	
5300	TRIZ OOR	
5400	VEIS VEL	
5500	***	
5600	--- ---	
5700	AIS A1	
5800	A3O A3O	
5900	A3S A3O	
6000	AIS AL	
6100	EIS E1	
6200	E2S E2S	
6300	EIS IL	
6400	ESA E2S	
6500	IIS I1	
6600	LES L	
6700	NES N	
6800	NIS NIL	
6900	O1S O1	
7000	O2S O2	
7100	OES AO	
7200	OIS OL	
7300	ONA A3O	
7400	ORA OR	
7500	RES R	
7600	U1S U1	
7700	UIS UL	
7800	VIS VIL	
7900	ZES Z	
8000	ZIS ZIL	
8100	***	
8200	2S 2S	
8300	A1 A1	
8400	A2 A2	
8500	A3 AO	
8600	AS A	
8700	E1 E1	
8800	E2 E2	
8900	ES ES	
9000	ES E	
9100	I1 I1	
9200	IS I	
9300	NS N	
9400	O1 O1	
9500	O2 O2	
9600	O3 O3	
9700	OA AO	
9800	OS O	
9900	U1 U1	
10000	US U	

Após a etapa da análise morfológica (representada no fluxograma constante do quadro 3.4), que compreende a busca nos dicionários e tabelas ATAB, VTAB, VDIC, FDIC, E/FDIC, NDIC e NTAB, segue-se a terceira etapa da análise linguística referida no parágrafo inicial desta seção. Esta etapa da análise contextual define os traços configuracionais das categorias lexicais e gramaticais codificadas na rotina HTAB - tabela da análise contextual - (quadro 3.11). Dada a complexidade desta tabela, descrevemo-la minuciosamente em 3.6.6 3.6.6.1 e 3.6.6.2

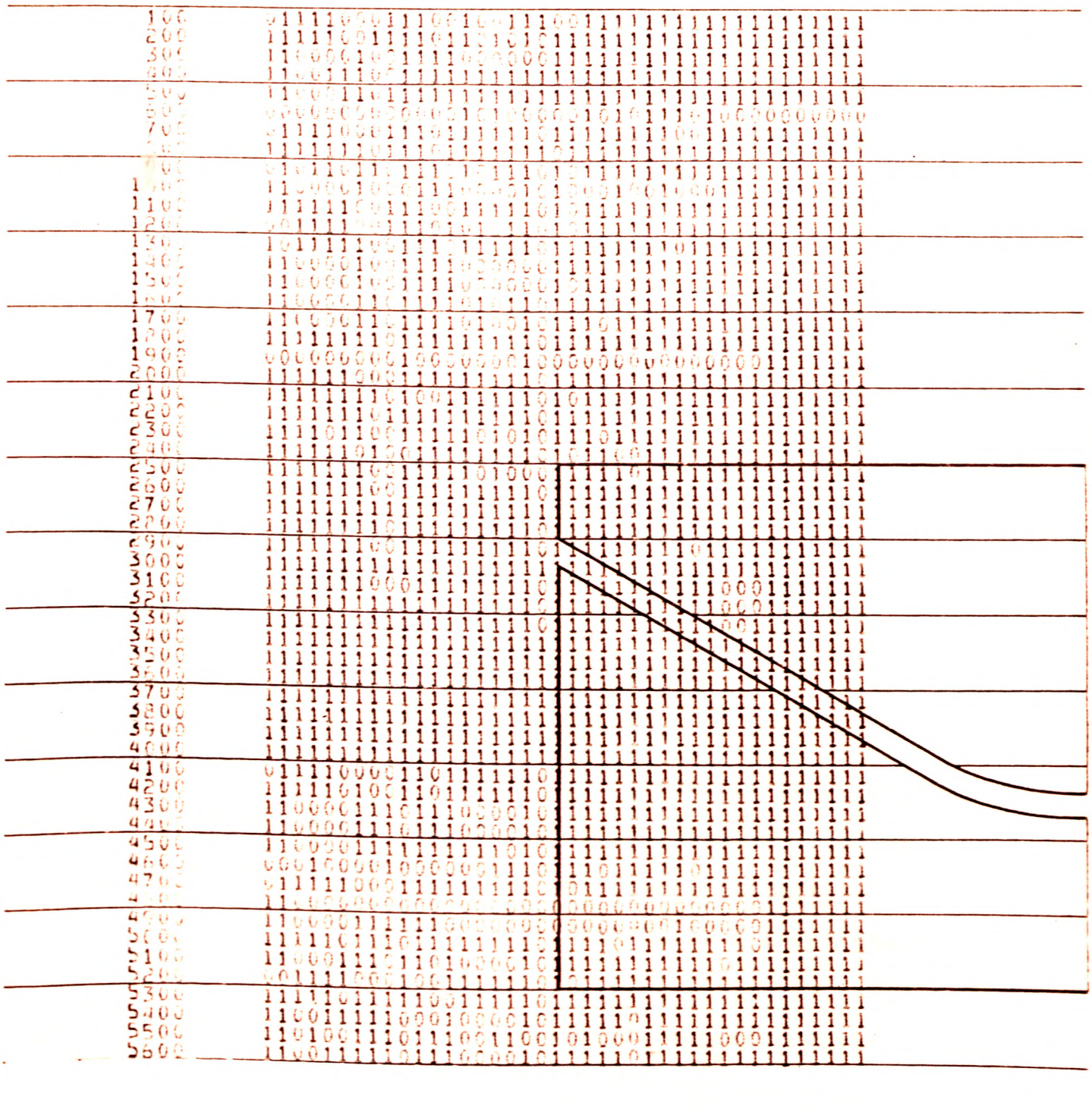
Feita a descrição do quadro 3.4 que é, em verdade, uma especificação da etapa da análise morfológica do fluxograma do LINGA, continuaremos com a descrição das etapas ainda não enfocadas do quadro 3.3.

A etapa de análise das homografias, em seqüência da etapa da análise contextual, objetiva a desambiguação das homografias ocorrentes entre o nome e o adjetivo; o nome e o verbo, etc. Esta etapa está descrita em maiores detalhes em 3.6.6 e 3.6.6.1.

Depois da definição das categorias gramaticais e lexicais, depois da análise contextual e depois da análise e desambiguação das homografias, o programa LINGA está apto a fornecer aos usuários do sistema uma análise automática da língua escrita, possibilitando, com isso, a sua utilização para os mais variados fins nos sistemas de informação, como, por exemplo, o fornecimento de listas de candidatos a descritores simples de um texto. Esta etapa está descrita no capítulo 5.

QUADRO 3.11

ANÁLISE CONTEXTUAL (HTAB)



3.6.1.2 Tabela das categorias morfossintáticas

Passaremos a seguir para a explicação do quadro 3.12, que nada mais é do que um quadro elucidativo das categorias morfossintáticas desenvolvidas no sistema LINGA. Consiste este quadro de (1) um número codificado de cada categoria morfossintática; (2) de uma sigla correspondente a cada categoria; (3) nominação da categoria e (4) um exemplo entre apóstrofes correspondente a cada categoria. Este quadro foi organizado pelo autor desta dissertação.

QUADRO 3.12

TABELA DAS CATEGORIAS MORFOSSINTÁTICAS

01 - NOM - núcleo de um sintagma nominal;

02 - ATT - adjetivo em função de adjunto adnominal;

03 - VRB - forma verbal flexionada;

04 - INF - infinitivo em função verbal;

05 - PAR - particípio em função verbal;

06 - PRD - adjetivo ou particípio em função predicativa;

07 - NNO - numeral como núcleo de um sintagma nominal;

08 - NAT - numeral em função de adjunto adnominal;

09 - DET - determinante (artigo definido e indefinido)

10 - REL - pronome relativo;

11 - QAT - categorias integrantes da supercategoria do adjetivo (cf. Lemle 1984: 150)

12 - NML - a supercategoria do adjetivo como núcleo do sintagma nominal;

13 - SPR - pronome do caso reto;

14 - AUX - forma flexionada de um verbo auxiliar;

15 - VBM - forma flexionada de um verbo modal;

16 - AUI - forma infinitiva de um verbo auxiliar;

17 - MDI - forma infinitiva de um verbo modal;

18 - PRP - preposição simples;

19 - ELP - elemento de uma locução preposicional;

20 - QNT - quantificador (na acepção da gramática gerativa);

21 - DEM - é a partícula 'demais' que Haller distinguiu de outras formas adverbiais;

22 - AVP - advérbio que modifica um advérbio, adjetivo ou nome;

23 - GER - gerúndio;

24 - CON - conjunção coordenativa

25 - SUB - conjunção subordinativa

26 - ADV - advérbio;

27 - - é um espaço que poderá ser preenchido por uma categoria ainda não desenvolvida ou por um desmembramento de alguma das categorias já desenvolvidas;

28 - NEG - advérbio de negação;

29 - OPR - pronome oblíquo proclítico (os pronomes mesoclíticos e enclíticos passarão a ser classificados como OPR quando as formas verbais hifenadas forem identificadas pela rotina VTAB;

30 - - é uma categoria vazia (como a categoria 27);

31 - PNT - ponto final de uma frase;

32 - COM - vírgula;

33 - SEP - elementos gráficos (parênteses, travessão, barra, etc.)

3.6.2 O dicionário de freqüência (FDIC)

Após a segmentação de um texto em frases e em palavras, as palavras são submetidas ao fluxograma de rotinas de análise morfológica do LINGA. Conforme podemos observar no quadro 3.4 as palavras são submetidas à rotina do dicionário de freqüência - FDIC (quadro 3.8) -, depois de sofrerem a análise das rotinas ATAB (quadro 3.5), VTAB (quadro 3.6) e VDIC (quadro 3.7). A disposição da rotina FDIC nessa ordem no fluxograma do LINGA (quadro 3.3) é conveniente, uma vez que, a esta altura da análise, os verbos ocorrentes no texto já estarão identificados por VTAB e VDIC. As palavras restantes da frase serão submetidas à análise do FDIC. Na rotina FDIC "são armazenadas as palavras mais freqüentes de uma língua que constituem mais ou menos 50% de um texto normal. (...) Estas palavras são ao mesmo tempo, as mais importantes da língua para determinar a estrutura (...) de uma frase" (Haller, 1983:107). São os instrumentos gramaticais com função relacional, segundo Vendryès (apud Freitas, 1981:31). A maioria das palavras arroladas neste arquivo pertencem às classes fechadas, (V. Pottier, 1978: 275-282 e Martinet, 1964). Estas palavras são os artigos, as preposições, as conjunções, os pronomes, os advérbios, conforme especificação de Mattoso Câmara (1977: 59-60). Conforme a classificação das categorias morfossintáticas do sistema LINGA, constam do FDIC as seguintes categorias: NNO, NAT, DET, REL, QAT, NML, SPR, PRP, ELP, QNT, DEM, AVP, CON, SUB, ADV, NEG e OPR. Os verbos anômalos, modais e auxiliares e suas flexões (AUX, VBM, AUI, MDI) cons-

tam também desta rotina porque não se enquadram no paradigma de conjugação dos verbos regulares e irregulares e, também, por fugirem da classificação específica feita pela rotina VTAB. Constam, ainda, desta rotina, palavras ambíguas devido à sua homografia. Por exemplo, a palavra 'caso' poderá ser uma conjunção subordinativa (SUE), um verbo (VRB) ou um substantivo (NOM). Dada a quantidade de categorias morfossintáticas analisadas pelo FDIC (21 ao todo), bem podemos avaliar a sua importância dentro do sistema.

Passemos, agora a uma descrição esmiuçada da estrutura do FDIC (ver quadro 3.8) no exemplo abaixo:

Classificações possíveis

	1ª				2ª			3ª			4ª
		T	M			T	M		T	M	
		J	/			I	/		I	/	
FWORD	FAC	CNP	M	G	CNP	M	G	CNP	M	G	CNP
AO	111S3		M								
AQUI	126		L								
CASO	301S3		M25					03S1PRSI			
DAQUELAS	118P3	5	F								
DOU	103S1PRSI										

- modo gênero ou outras informações semânticas
- Tempo verbal
- pessoa verbal
- número
- categoria morfossintática do LINGA analisada
- número de funções morfossintáticas possíveis.

As palavras constantes da rotina FDIC estão listadas em ordem alfabética. Na coluna FAC, do inglês 'FACTOR', identifica-se o número de categorias morfossintáticas que uma palavra pode ter. Por exemplo, a palavra 'caso' poderá ter três diferentes classificações, conforme já vimos acima. Na coluna C, os dígitos indicam a categoria morfossintática. Já vimos na explicação do quadro 3.12 que cada categoria desenvolvida no LINGA está codificada por um número. É este número que aparece na coluna C. A coluna N indica a flexão de número da palavra pelas letras S (singular) ou P (plural). A coluna P indica a pessoa da forma verbal. A coluna TIM indica o tempo das formas verbais flexionadas que estão codificados da seguinte forma:

PRS - Presente

PRT - Pretério perfeito

FUT - Futuro do presente

IMP - Imperfeito

CON - Futuro do pretério (condicional)

A coluna M/G indica o modo das formas verbais sob o código: I - indicativo; S - subjuntivo; P - particípio e G - gerúndio. Esta coluna indica também o gênero das palavras. Não há problema para as indicações de gênero ou modo nesta coluna porque as palavras aqui assinaladas ou são verbos (e então receberão indicação de modo) ou serão nomes (e então receberão a indicação de gênero). Esta coluna indica, também, a circunstância para as formas adverbiais.

A seqüência de colunas, a partir da coluna C, é repetida tantas vezes quantas for o número de categorias que a palavra poderá ter. Para melhor esclarecermos este assunto, tomemos como exemplo as palavras 'abaixo', 'acaso' e 'caso' constantes da amostragem do FDIC no quadro 3.6. A palavra 'abaixo' está classificada em apenas uma categoria, conforme especificação da coluna FAC. É um advérbio (ADV) codificado sob o número 26 na coluna C; na coluna M/G há a indicação da circunstância de lugar (L). A palavra 'acaso' está classificada em duas categorias, conforme indicação da coluna FAC; poderá ser um substantivo (NOM) codificado pelo número 01 na coluna C; o 'S' da coluna N (número) indica o singular e o 'M' da coluna M/G indica masculino; a outra classificação de 'acaso' poderá ser a de um advérbio (ADV) codificado sob o número 26 na coluna 26. A palavra 'caso' poderá ser classificada em três categorias morfossintáticas desenvolvidas no LINGA, conforme indicação da coluna FAC. A primeira classificação indica um nome (NOM) codificado pelo número 01 na coluna 'C'; a coluna 'N' indica 'S' (singular) e a coluna M/G indica M (masculino); a segunda classificação de 'caso' indica uma conjunção subordinada, codificada pelo número 25; a terceira classificação da palavra 'caso' indica um verbo (VRB) codificado pelo número 03 na coluna 'C' (categoria); a coluna N (número) indica S (singular); a coluna P (pessoa) indica 1 (primeira pessoa); a coluna TIM (tempo) indica PRS (presente) e a coluna M/G (modo) indica I (indicativo). Como se percebe pela descrição destes três exemplos, a tabela FDIC procura dar cobertura, com a maior quantidade de informações possíveis, às palavras nela cons-

tantes. A tabela FDIC foi organizada em seus elementos estruturais por Carlos A. de Oliveira. No entanto, durante a fase de testes e correções, todos os integrantes da equipe do LINGA tiveram participação efetiva.

A rotina E/FDIC (v. excerto no quadro 3.13) English Frequency Dictionary organizada inteiramente por Mary Lynn Kelso, obedece aos mesmos procedimentos do FDIC. Esta tabela está inserida na rotina do programa LINGA para, eventualmente, identificar e definir elementos gramaticais da língua inglesa que possam ocorrer em textos portugueses.

3.6.3 A morfologia nominal do LINGA

A análise da morfologia nominal do LINGA é processada a partir das desinências nominais que as palavras podem ter. A morfologia nominal no LINGA é processada após a análise da morfologia dos elementos gramaticais processada pela rotina FDIC.

Para que fosse possível converter as formas nominais flexionadas em gênero e número em sua forma dicionarizada, Carlos Alberto de Oliveira desenvolveu a rotina NTAB (quadro 3.10) estruturada em quadrigramas, trigramas e digramas. Por meio dos quadrigramas as palavras são analisadas a partir de suas quatro últimas letras. Encontrando-se em uma palavra a conformidade com um dos quadrigramas propostos no algoritmo NTAB, esta palavra é identificada como um nome ou adjetivo e sua forma flexionada é convertida à

QUADRO 3.13

DICIONÁRIO DE FREQUÊNCIA DO INGLÊS (E/FDIC)

	FWORD	FAC CNP M G CNP	T M I / INFORMACOES EXTRAS
400	***	*** *** * * ***	**********
500	A	109S	+18 "$5 A WEEK"
600	ABACK	126	
700	ABOARD	218.26	
800	ABOUT	318.22.26	22 ONLY WITH 06
900	ABOVE	302.18.26	
1000	ABREAST	126	
1100	ABROAD	126	
1200	ACROSS	218.26	
1300	AFAR	126	
1400	AFLOAT	206.26	
1500	AFOOT	206.26	
1600	AFTER	318.25.26	+02 "AFTER YEARS"
1700	AFTERWARDS	126	
1800	AGAIN	126	
1900	AGAINST	118	
2000	AGO	202.22	02 ALWAYS AFTER NOUN;
2100	AHEAD	206.26	22 ONLY W "LONG"
2200	ALBEIT	125	
2300	ALIKE	206.26	
2400	ALIVE	206.26	
2500	ALL	411.12.22.26	1
2600	ALMOST	222.26	
2700	ALOFT	206.26	
2800	ALONE	206.26	
2900	ALONG	218.26	
3000	ALONGSIDE	126	
3100	ALOOF	302.06.26	
3200	ALOUD	126	
3300	ALREADY	126	
3400	ALSO	126	
3500	ALTHOUGH	125	
3600	ALTOGETHER	126	+01 SLANG
3700	ALWAYS	126	
3800	AM	203S1FRS 14	
3900	AMID	118	
4000	AMIDST	118	
4100	AMISS	206.26	
4200	AMONG	118	
4300	AMONGST	118	
4400	AN	109S	+18 "$5 AN HOUR"
4500	AND	124	
4600	ANOTHER	211.12	
4700	ANY	311.12.22	
4800	ANYBODY	213.29	
4900	ANYHOW	225.26	
5000	ANYONE	213.29	
5100	ANYTHING	213.29	
5200	ANYWAY	225.26	
5300	ANYWHERE	225.26	
5400	APART	126	+02 AFTER NOUN "A MAN APART" (USAGE)
5500	ARE	203.14	
5600	AREN'T	303.14.28	
5700	AROUND	218.26	
5800	AS	310.22.25	22 ONLY WITH 25;
5900	ASHORE	126	+ "18 MODIFIER"
6000	ASIDE	201.18	
6100	ASKEW	126	
6200	ASUNDER	126	
6300	AT	118	
6400	AWAY	126	
6500	AWHILE	126	USAGE; "A WHILE"
6600	BACK	501.02.03.04.26	
6700	BACKWARD	202.26	WENT BAD"
6800	BACKWARDS	126	BAD ABOUT IT; FOOD
6900	BAD	301.02.26	26 USAGE; "HE FELT
7000	BARELY	222.26	+ QUASI-NEG
7100	BE	203.14	
7200	BECAUSE	219.25	
7300	BEEN	203.14	
7400	BEFORE	318.25.26	
7500	BEHALF	201.20	01 OBJ 18 ONLY
7600	BEHIND	218.26	
7700	BEING	201.23	
7800	BELOW	218.26	
7900	BENEATH	218.26	

semelhança com o que é encontrado no dicionário. O mesmo procedimento ocorre com os trigramas e os digramas. No trigrama, as palavras são convertidas a sua forma dicionarizada a partir de suas três letras finais, caso ainda não tenham sido convertidas pelo quadrigrama. No digrama, as palavras são convertidas a sua forma dicionarizada, a partir de suas duas letras finais, caso ainda não tenham sido definidas pelos quadrigramas ou trigramas. Qual a razão destes procedimentos? Vejamos a partir do exemplo: a palavra 'cantoras', caso fosse analisada apenas pelo digrama, seria convertida para a forma cantora e não para o singular masculino, como é sua forma dicionarizada. Por esta razão, todas as palavras que são analisadas por NTAB são submetidas aos quadrigramas, aos trigramas e aos digramas até que se encaixem em uma de suas formas. Assim, exemplificando mais, teremos:

a) Quadrigramas

veis -- vel	automóveis	-	automóvel
triz -- dor	imperatriz	-	imperador
onas -- ão	choronas	-	chorão, etc.

b) Trigramas

ais -- al	jornais	-	jornal
res -- er	poderes	-	poder
ões -- ão	leitões	-	leitão, etc.

c) Digrama

ns -- m	homens	-	homem
as -- a	facas	-	faca
os -- o	cocos	-	coco, etc.

No quadro 3.14 Tabela de palavras especiais, encontramos uma amostragem dos substantivos que oferecem dificuldades de análise pelos quadrigramas, trigramas ou digramas de NTAB. Se por acaso os substantivos paroxítonos terminados em 's' fossem analisados por NTAB, aconteceria que palavras como 'lápis', 'bônus', 'atlas', 'pires', etc. teriam sua forma convertida para 'lápi', 'bonu', 'atla', 'pire ', etc. Nesta tabela, da qual o quadro 3.14 é uma amostragem, encontramos também palavras proparoxítonas como 'óculus', 'ônibus' - que recebem o mesmo tratamento. Enfim, nesta tabela encontram-se todos os substantivos e adjetivos que oferecem dificuldades de serem analisados por NTAB. Como se percebe pela descrição de todos os procedimentos executados pelo sistema LINGA através de suas rotinas, toda a estrutura morfossintática da língua portuguesa é coberta. Assim, sendo, o sistema está apto para as mais variadas aplicações, dentre elas, a indexação automática, a correção automática de erros de ortografia, tradução automática. etc.

Se por acaso acontecer de uma forma não se enquadrar em nenhum dos procedimentos descritos, esta forma será considerada como estranha à estrutura da língua portuguesa e virá assinalada por EST na análise morfossintática. Assim mesmo, será considerada como uma forma nominal (NOM). Poderá ser uma sigla, uma palavra estrangeira incorporada ao nosso uso lingüístico, um nome próprio, etc. Exemplos: BNH, merchandising, royalty, Spohr, etc.

QUADRO 3.14

TABELA DE PALAVRAS ESPECIAIS

2669	ARVORES	ARVORE
2670	ATLAS	ATLAS
2671	ATOA	ATOA
2672	ATOAS	ATOA
2673	AZUNUS	AZUNUS
2674	AZA	AZ
2675	AZAS	AZ
2676	AZULS	AZULS
2677	AZULMENTES	AZUL MENTE
2678	AZULMENTES	AZULMENTE
2679	AZULTIJOLA	AZULTIJOLO
2680	AZULTINLA	AZULTIJOLO
2681	BRUA	BRUXA
2682	BRUAS	BRUXA
2683	BEUSA	BEIJOS
2684	BEUSAS	BEIJOS
2685	BIPEDA	BIPEDA
2686	BIPEDAS	BIPEDA
2687	BIMA3	BIMA3
2688	BIMA3S	BIMA3
2689	BMA	BMA
2690	BMAS	BMA
2691	CAPIS	CAPIS
2692	CAPIS	CAPIS
2693	CALES	CAL
2694	CISTA	CISTO
2695	CISTAS	CISTO
2696	CUA	CU
2697	CUAS	CU
2698	CICLLOS	CICLOS
2699	CCULOS	CCULOS
2700	CONIBUS	CONIBUS
2701	CONIBUS	CONIBUS
2702	CPIRES	CPIRES
2703	CELES	CELES
2704	SIMPLES	SIMPLES
2705	SOIS	SOL
2706	SOS	SOL
2707		

3.6.4 Dicionário de formas nominais (NDIC)

Da tabela do dicionário de formas nominais (NDIC) constam os substantivos e adjetivos que oferecem dificuldades de análise e classificação pelo programa de análise automática (LINGA). Na listagem, em ordem alfabética, encontram-se palavras homógrafas; adjetivos de ocorrência pré-nominal; prováveis não-descritores; descritores em contexto específico e palavras que podem fazer parte de um descritor composto. Este dicionário foi organizado pelo autor desta dissertação em colaboração com Simone Bastos que prestou as informações necessárias quanto à parte da biblioteconomia.

3.6.4.1 As homografias

As homografias vêm assinaladas por um asterisco (*) na coluna 31 do dicionário. São palavras que na análise de textos podem ser analisados como NOM, VRB ou ATT. Em verdade, esta listagem das homografias no dicionário NDIC não se faria necessária, uma vez que o próprio sistema LINGA já poderia analisar e classificar estas palavras. No entanto, por questão de economia e rapidez no processamento, é conveniente que essas palavras permaneçam na tabela.

3.6.4.2 Adjetivos pré-nominais

Na coluna 33 do dicionário NDIC, vêm assinalados os adjetivos (ATT) com o número (1) porque os mesmos podem ocorrer na posição pré-nominal. Na seção 4.2.1 desta dis

sertação apresentamos um estudo de Pazini (1978) e de Lobato (1979) sobre as condições que os adjetivos devem preencher para poderem ser deslocados para a posição pré-nominal. Os adjetivos são subcategorizados no léxico de diferentes modos, conforme veremos na seção 4.1, especificamente a partir do exemplo 14 daquela seção. Como se percebe, os adjetivos têm uma marcação lexical que envolve conseqüências significativas para a análise lingüística do LINGA. Por exemplo, na seqüência: "A bela casa", se o adjetivo 'bela' não fosse marcado lexicalmente no dicionário NDIC na coluna 33, o mesmo seria considerado um nome (NOM). Isto porque as regras distribucionais contidas na rotina HTAB (quadro 3.1), determinadas a partir dos traços configuracionais das categorias lexicais, estabelecem que após um determinante ocorre uma forma nominal. Ora, sabemos que um adjetivo pode ter as mesmas desinências de um nome, resultando daí uma confusão entre a classificação de um nome e um adjetivo para o LINGA. Esta confusão é desfeita pela marcação dos adjetivos no NDIC. Em 4.1 será analisado em profundidade este problema da distribuição das categorias no SN.

3.6.4.3 Coluna dos não-descritores

A extração de descritores de um texto pelo sistema LINGA compreende a seleção das formas nominais representativas através da rotina do dicionário NDIC. Este dicionário contém uma listagem de nomes e adjetivos classificados em três categorias para efeitos de indexação automática. A saber:

1) Prováveis não-descritores. Essas palavras, assinaladas na coluna 35, dificilmente sintetizam o conteúdo de um texto analisado. São palavras que têm uma significação vaga. Por isso, são descartadas para efeitos de indexação. Como exemplos, temos as palavras: tipo, coisa, vazio, etc.

2) Descritores em contexto específico. Estas palavras, assinaladas no NDIC na coluna 37, só serão descritores de um texto quando o mesmo se referir a um assunto bem específico. Por exemplo, a palavra 'causa' poderá ser descritor de um texto da área do direito, mas dificilmente será descritor da área de literatura; a palavra 'palavra' será descritor de um texto de lingüística, mas não de um texto de medicina, etc.

3) Candidatos a descritores compostos. Essas palavras assinaldas na coluna 39, sozinhas, dificilmente serão descritores de um texto, mas não poderão ser desprezadas para efeitos de indexação. Estando assinaladas, possibilitam que sejam melhor especificadas por um adjetivo restritivo . Por exemplo, em ajuda clínica/financeira/econômica, etc. percebemos que a palavra 'ajuda' é um termo muito vago em sua especificação, necessitando por isso de um adjetivo que precise seu significado.

O estudo da estrutura do SN da língua portuguesa (cf. Capítulo 4 desta dissertação) será de importância para a definição dos descritores de um texto, para a definição dos não-descritores, para a definição dos descritores em contexto específico e, sobretudo, para a definição da estrutura lingüística do descritores compostos.

Da coluna 41 de NDIC constarão os particípios irregulares (p. ex.: morto - matado) dos verbos. A análise verbal realizada pela rotina VTAB utilizará esta informação para completar a análise do verbo.

3.6.5 A morfologia verbal (VTAB)

A rotina VTAB (quadro 3.6), a partir das desinências verbais possíveis da língua portuguesa, realiza a análise morfológica das palavras ocorrentes em um texto e determina se estas pertencem ou não à categoria verbal. Encontrando-se uma forma verbal, a mesma é reduzida à sua forma infinitiva e então será confrontada com as mais de 770 formas verbais infinitivas que estão armazenadas no dicionário de verbos - VDIC (ver quadro 3.7). Havendo confronto entre a forma definida por VTAB com uma das formas armazenadas em VDIC, a mesma será considerada um verbo e então sofrerá os efeitos das demais análises do sistema LINGA. Não havendo o confronto entre VTAB e VDIC, a forma será considerada uma forma nominal para efeitos de análise sintática no sistema. Evidentemente, poderá suceder que a forma definida por VTAB seja de fato uma forma verbal e que ainda não esteja armazenada em VDIC. Neste caso deverá ser inserida na ordem alfabética a referida forma infinitiva do verbo em VDIC.

Pode acontecer que apareçam formas verbais infinitivas artificiais geradas pela análise morfológica dos verbos irregulares. Neste caso, a tabela VDIC converte esta

forma artificial para sua forma infinitiva correta. Por exemplo, as formas artificiais "estejar", "estever", "estiver" são convertidas para a forma infinitiva "estar". Estas formas artificiais estão listadas no VDIC porque a rotina VTAB define-as como formas infinitivas das formas flexionadas de 'estar', ou seja de: 'estejam', 'esteve', 'estiveram'. Neste caso, as formas infinitivas corretas aparecerão na coluna 21 do VDIC; nas colunas 41 a 43 aparecerá o tempo do verbo e na coluna 44 o modo da forma verbal irregular. Nas colunas 61 a 63 encontra-se a regência verbal de cada verbo codificada da seguinte forma, obedecendo aos princípios da marcação lexical de cada verbo.

Código - marcação lexical

1 - V [___] (verbo intransitivo)

2 - V [___SN] (verbo transitivo direto)

3 - V [___SP] (verbo transitivo indireto)

4 - V [___SN SP] (verbo bitransitivo)

5 - V [___Cop { SN / SAdj / SP }] (verbo de ligação)

6 - V [VM___] (verbo modal)

A ocorrência de mais de um código significa que a forma verbal terá mais de uma regência. A marcação lexical dos verbos constantes em VDIC foi elaborada pelo autor desta dissertação.

3.6.6 A desambiguação das homografias

No processo de comunicação em língua natural, o ouvinte recebe uma mensagem na qual os elementos e suas combinações podem ser susceptíveis de várias interpretações; a primeira orientação para o ouvinte é a identificação do domínio conceptual no qual se situa a mensagem recebida. Esse processo semasiológico, conforme Pottier (1975: 130-3), não se constitui em dificuldade para o falante/ouvinte, dada a sua capacidade de desfazer as polissemias identificando o contexto da mensagem.

No caso da decodificação da mensagem pela análise morfossintática do sistema LINGA, o processo é semelhante: quando as palavras são submetidas à rotina HTAB (quadro 3.11) as mesmas já vêm assinaladas pelas rotinas FDIC, VDIC, VTAB, NDIC e NTAB com as possíveis categorias lexicais que poderão assumir. Veja-se o exemplo da palavra 'caso' no quadro 3.8. Isto significa dizer que qualquer palavra ao ser submetida à rotina HTAB já estará definida em sua estrutura morfológica, mesmo sendo uma palavra de outro código lingüístico, ou uma sigla, ou uma palavra com erro ortográfico. Isto porque cada palavra sofre o processo da segmentação e da busca nos dicionários. É de importância ressaltar que pela rotina HTAB é definido o contexto sintático das palavras, obedecendo-se a seus traços configuracionais marcados no léxico.

No entanto,podem ocorrer, ainda, algumas ambigüidades provenientes da identidade de significantes que pertencem a categorias lexicais diferentes (cf. Pottier, 1975). É o caso dos substantivos deverbais que coincidem com uma forma flexionada do próprio verbo, sobretudo nas desinências - o ou - a(s), Como exemplos, citamos: luta (verbo)-luta (nome); gosto (verbo), gosto (nome); abandono (verbo), abandono (nome); etc. Nestes casos, o significante recobre dois sememas que têm semas em comum, e o contexto sintático, permite, em geral, escolher a palavra adequada (V. Pottier, 1975). No sistema LINGA quem estabelece este contexto sintático, a partir dos traços configuracionais da palavra, é a rotina HTAB. Por exemplo, as palavras: canto, acordo, apoio, busca(s), ajuda(s) serão marcadas lexicalmente como sendo nomes ou verbos conforme a existência, à sua esquerda, de determinantes ou elementos com função sintática de sujeito, respectivamente.

Lemle (1984:114) considerou três casos potencialmente duvidosos em termos de separação entre nomes e verbos. O primeiro diz respeito a palavras com a forma de infinitivos verbais, tais como: 'os deveres' e 'o parecer'. A solução aí adotada foi considerá-los como nomes e admitir que existe uma regra no léxico que relaciona verbos e nomes deverbais de formas básicas idênticas. O segundo caso considerado foi o dos infinitivos flexionados, precedidos pelo determinante 'o'. A análise proposta foi a de admitir que a sentença infinitiva é dominada por um módulo de sintagma nominal cujo núcleo está vazio e que deve ser preenchido in

terpretativamente pela palavra 'fato'. Por exemplo, a seqüência: 'O fazeres isto me constrange' deve ser interpretado como: 'O fato de fazeres isto me constrange'. Finalmente, a autora considerou seqüências como: "saiu para passear" ou "quero dormir", em que o infinitivo é analisado como um verbo não flexionado. No LINGA, os dois primeiros casos não oferecem dificuldades para a análise e definição, dado o fato de serem marcados pelo determinante. No terceiro caso, a forma infinitiva recebe a classificação INF (infinitivo em função verbal) pela tabela VTAB e portanto, também não oferece dificuldades para sua classificação.

Podem ocorrer polissemias onde não se perceba mais a relação entre os sememas, quaisquer que tenham sido suas relações históricas, por exemplo:

- planta (vegetal, plano de uma construção);
- canto (canção, borda);
- cabo (parte de um instrumento, acidente geográfico, patente militar).

Nestes casos, no LINGA, igualmente não haverá dificuldades, pois a rotina HTAB define as categorias apenas pelo contexto morfossintático em que se encontra cada palavra em uma frase. A definição apropriada da categoria lexical de cada palavra polissêmica é de suma importância no sistema LINGA, tanto para indexação automática quanto para a tradução automática, conforme Haller (1983).

Para uma análise sintática completa de uma frase, o sistema LINGA realiza uma busca dos vizinhos à esquer-

da e à direita de cada palavra a partir da marcação lexical inserida na rotina HTAB. Assim, as homografias de palavras que podem ocupar categorias diferentes são desfeitas, como é o caso das palavras que se seguem:

- cedo VRB/ADV
- cerca VRB/PRP/NOM
- como VRB/CON
- era NOM/AUX
- estado NOM/PAR
- nada VRB/PRD/NOM
- para VRB/PRP

As poucas homografias restantes são decididas com base nas probabilidades de categorias soltas, descritas adiante.

Nos sistemas de informação que trabalham com textos sem uma análise lingüística, como por exemplo, o sistema STAIRS da IBM, em operação no PRODASEN e na EMBRAPA, estas palavras policategoriais proporcionam resultados bastante reduzidos quanto à precisão porque oferecem uma recuperação de informações de muitos documentos que não pertencem ao âmbito da pergunta do usuário.

As categorias PAR, REL, MDL e OPR (V. quadro 3.12) podem ser eliminadas da análise de uma base de dados na etapa da análise morfológica (V. quadro 3.4) se forem homógrafas, mais especificamente no dicionário de freqüência (FDIC) (V. excerto no quadro 3.8. Estas categorias são as que denominamos acima de categorias soltas. Depois des-

ta etapa, a desambiguação da homografia nome/adjetivo (NOM/ ATT) é particularmente difícil na língua portuguesa, dado o fato de ambas terem as mesmas desinências de gênero e número. Neste caso, na etapa da análise morfológica, estas duas possibilidades de classificação são deixadas em aberto para que a rotina HTAB defina pela análise contextual qual das categorias é nome (NOM) e qual é adjetivo (ATT).

Outras possibilidades de homografias com as palavras instrumentais são tratadas pela HTAB, tais como:

PRP/DET	- a, as
PRP/NOM	- pelo, pelos, etc.
CON/SUB	- como
SUB/OPR	- se

Algumas palavras, porém, necessitam ser tratadas no LINGA de modo explícito, dado o fato de serem policategoriais; são palavras que ainda apresentam dificuldades na desambiguação das homografias. Vejam-se os exemplos:

que	- REL/AVP/SUB/OPR
vão	- NOM/ATT/VRB/VBM
caso	- NOM/VRB/SUB
são	- NOM/ATT/AUX

É de ressaltar que todas as possibilidades de ocorrência dessas palavras policategoriais são definidas no dicionário de freqüência (FDIC- vide quadro 3.8) que descrevemos na seção 3.6.2. Cabe à rotina HTAB a definição sintática da categoria e sua função dentro da frase.

Como se percebe pela descrição da lógica do sistema LINGA, sobretudo da rotina HTAB, a maioria das homografias ocorrentes na língua escrita são resolvidas. Porém, há a necessidade de se estabelecer critérios para desambiguar as palavras ainda restantes. Assim, quando ocorrer uma homografia ATT/NOM, será excluída a possibilidade ATT para não perder o candidato a descritor. No caso da indexação automática, no atual estágio do sistema LINGA, somente são listados os nomes (NOM) como candidatos a descritores. Para os descritores compostos, o sistema LINGA necessita das regras do SN, abordadas no capítulo 4 desta dissertação. Quando estas regras forem incorporadas ao sistema, ainda assim, será necessário distinguir ATT de NOM, uma vez que os ATTs com a marcação lexical [NOM],isto é, os adjetivos pré-nominais não são considerados para efeitos de indexação. Como já vimos, os adjetivos com esta marcação lexical estão arrolados na tabela NDIC e descritos em 3.6.4.2.

Caso ainda haja dúvidas quanto à definição destas duas categorias e caso o adjetivo não esteja marcado lexicalmente no dicionário de formas nominais - NDIC, como descrevemos em 3.6.4.2, prevalecerá a seguinte ordem na classificação: a primeira categoria será um nome a e segunda será um adjetivo. Este procedimento é de relevância para a definição lingüística dos descritores, uma vez que serão os nomes os candidatos preferenciais para a indexação.

A rotina HTAB desambigua, também, as homografias ocorrentes com a forma verbal (VRB) e as categorias seguintes:

- sua VRB/QAT
- como VRB/CON
- entre VRB/PRP, etc.

as categorias QAT, CON, PRP representam no sistema LINGA as palavras semiplenas e instrumentais (Vide Biderman, 1978: 251-4).

3.6.6.1 Estruturas da rotina HTAB

A rotina HTAB (quadro 3.11) é composta de 80 linhas, cada uma das quais com 40 colunas para registro da codificação da marcação lexical dos itens lexicais desenvolvidos no LINGA. As primeiras 40 linhas, de 01 a 40, representam a marcação lexical à esquerda de cada item lexical em relação aos demais. As linhas 41 a 80 representam a codificação da marcação lexical à direita de cada item lexical em relação aos demais. Com esse procedimento objetiva-se cobrir as coocorrências de cada item lexical. A rotina HTAB compõe-se, portanto, de duas matrizes 40 X 40, sendo que os elementos (itens lexicais) estão dispostos na mesma ordem, tanto em linha como em coluna, i. e., cada item lexical terá a mesma codificação em linha como em coluna. Por exemplo, a preposição (PRP) receberá sua codificação na linha 18 em relação aos demais itens lexicais e, na coluna 18, todos os itens lexicais serão marcados em relação à preposição. Mostramos no quadro 3.15 a marcação lexical da categoria nome (NOM) representada pelo

QUADRO 3.15

MARCAÇÃO LEXICAL DO NOME

Categoria	marcação lexical à esquerda - linha 01		marcação lexical à direita - linha 41	
	Código	Marcação	Código	Marcação
NOM	01	Ø _____	01	Ø _____
ATT	02	1 _____NOM	02	1 NOM____
VRb	03	1 _____NOM	03	1 NOM____
INF	04	1 _____NOM	04	1 NOM____
PAR	05	1 _____NOM	05	1 NOM____
PRD	06	Ø _____	06	Ø _____
NNO	07	Ø _____	07	Ø _____
NAT	08	Ø _____	08	Ø _____
DET	09	1 _____NOM	09	Ø _____
REL	10	1 _____NOM	10	1 NOM____
QAT	11	1 _____NOM	11	1 NOM____
NML	12	Ø _____	12	Ø _____
SPR	13	Ø _____	13	1 NOM____
AUX	14	1 _____NOM	14	1 NOM____
VBM	15	Ø _____	15	1 NOM____
AUI	16	Ø _____	16	1 NOM____
MDI	17	1 _____NOM	17	1 NOM____
PRP	18	1 _____NOM	18	1 NOM____
ELP	19	1 _____NOM	19	1 NOM____
QNT	20	Ø _____	20	Ø _____
DEM	21	Ø _____	21	1 NOM____
AVP	22	1 _____NOM	22	1 NOM____
GER	23	1 _____NOM	23	1 NOM____
CON	24	1 _____NOM	24	1 NOM____
SUB	25	1 _____NOM	25	1 NOM____
ADV	26	1 _____NOM	26	1 NOM____
---	27	1 _____NOM	27	1 NOM____
NEG	28	1 _____NOM	28	1 NOM____
OPR	29	1 _____NOM	29	1 NOM____
---	30	1 _____NOM	30	1 NOM____
PNT	31	1 _____NOM	31	1 NOM____
COM	32	1 _____NOM	32	1 NOM____
SEP	33	1 _____NOM	33	1 NOM____

código 01 (V. quadro 3.12). No quadro (3.15) encontra mos na linha 01 a marcação lexical à esquerda do nome em re lação às demais categorias. Na linha 41, encontramos a mar cação lexical à direita do nome em relação às demais catego rias. Para visualizarmos melhor estas duas linhas da roti- na HTAB (quadro 3.11) que representam a marcação lexical do nome, reproduzimo-las a seguir:

linha 01 - Ø 1111ØØØ111ØØ1ØØ111ØØ1111111111111111111111111

41 - Ø 1111ØØØØ11Ø111111Ø1111111111111111111111111111

onde 'Ø' representa a impossibilidade de ocorrência e '1' representa a possibilidade de ocorrência.

Observando o quadro 3.15, temos a especificação de cada categoria com seu respectivo código numerado de 01 a 33 e seu confronto com a categoria nome. Temos então a seguinte marcação à esquerda do nome: nome à esquerda de no me não ocorre, portanto a marcação será 'Ø'; o adjetivo (ATT) com nome (NOM) pode ocorrer, portanto tem a marca- ção simbolizada por '1' que corresponde a ATT [_____NOM]; em relação ao verbo flexionado (VRB) ocorre a seguinte marca- ção: VRB [____NOM] simbolizada por [1] e assim sucessiva- mente até serem confrontadas todas as categorias à esquerda do nome. Processo semelhante ocorre com a marcação lexical à direita de cada categoria que está codificada a partir do registro 41. Portanto, o registro 41 representa a marcação lexical à direita do nome, assim como o registro 01 repre- senta sua marcação à esquerda. Observando novamente o qua- dro 3.15, temos em detalhes esta marcação. Assim, nome (NOM) com nome à direita não ocorre e é simbolizada por '0';

nome com adjetivo terá a seguinte marcação: ATT [NOM_____] simbolizado por '1'; nome com verbo (VRB) terá a seguinte marcação: VRB [NOM_____] simbolizada por '1' e, assim, ocorre sucessivamente com todas as categorias em relação ao nome. Para a marcação lexical das demais categorias, observam-se os mesmos procedimentos descritos acima em relação a categoria nome. O resultado desta marcação aparece codificado na rotina HTAB que é a parte nuclear do sistema LINGA. A rotina HTAB representa para o LINGA o léxico tal como é proposto em Lobato (na seção 12.2.1, a sair). O autor da tese elaborou a marcação lexical da rotina HTAB, na qualidade de lingüista, sendo auxiliado por Simone Bastos e Mary Lynn Kelso.

3.6.6.2 Testes, erros e casos difíceis

A maioria dos textos que até o momento foram submetidos como testes à análise no LINGA (ver anexo II, textos A e B), tem a predominância da função referencial (cf. Jakobson, 1970: 123) e foram caracterizados em Vicini (1981: 23-9) como sendo objetivos, impessoais, sem a presença da função conativa, não redundantes, lógicos e de vocabulários específicos para cada ramo da ciência.

Um texto-piloto de ambigüidades, organizado por Andreewsky (1982) sob o título "Estação de caça 1981" foi também, submetido à análise. Percebeu-se pela análise destes textos que a grande maioria dos erros ocorrentes com os sistema LINGA remetem a situações alheias às suas possibili

dades de análise, tais como a definição da categoria de 'E' que é analisado como CON e AUX simplesmente porque o sistema de computação B-6700, da UnB, opera somente com caracteres maiúsculos, sem tratar os sinais de acentuação gráfica, fato que gera problemas para definição de palavras como 'PARA' que pode ser PRP, VRB ou NOM. Outro problema ocorre com a colocação inadequada dos sinais de pontuação, por parte de quem redige. Como estes sinais também são usados na definição de categorias gramaticais pelo algoritmo HTAB (quadro 3.11), sua má colocação incorrerá, necessariamente, em erros de análise. Outro problema são as abreviações. Como estas são marcadas por um ponto, o LINGA considerará este ponto como ponto final de frase, com evidentes prejuízos para a análise automática. A solução encontrada é inserir uma lista, a mais completa possível, das abreviações ocorrentes na língua escrita (ver seção 3.6.1 e quadro 3.5).

As formas verbais mesoclíticas e enclíticas ainda são tratadas como NOM pela tabela NTAB (quadro 3.10) dado o fato de não serem ainda definidas como verbos pelo algoritmo VTAB (quadro 3.6). Esse problema será solucionado com a inserção das formas pronominais (quadro 3.1

O processamento de texto no LINGA obedece ao fluxograma apresentado no quadro 3.3. A segmentação das palavras (quadro 3.2) define qualquer agrupamento de dígitos como possíveis palavras. A análise morfológica e a busca nos dicionários (ver quadros 3.4 e 3.16) são etapas se-

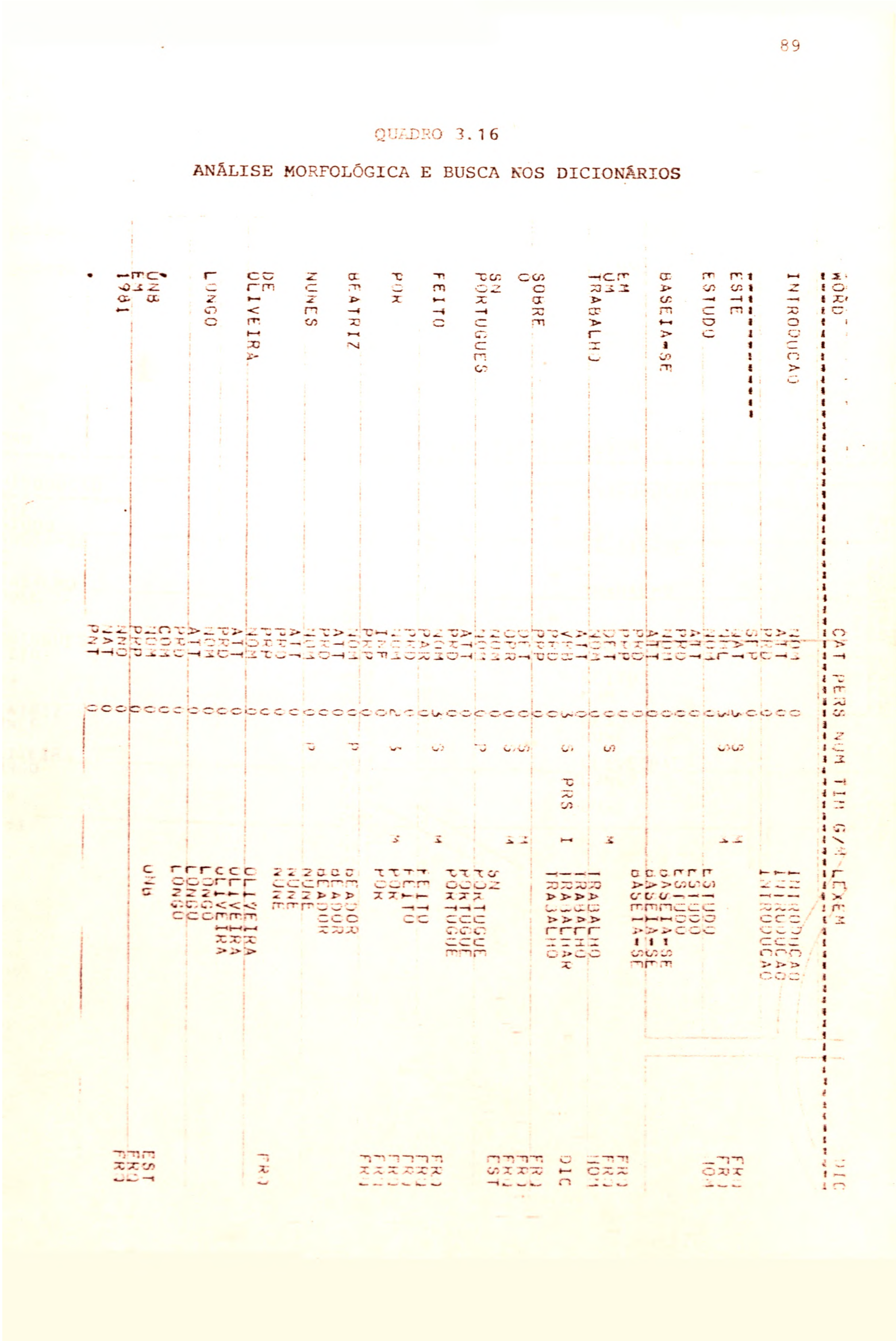
guintes à da segmentação. Nesta etapa, cada palavra recebe suas possíveis classificações. Assim, a palavra 'INTRODUÇÃO' poderá ser NOM, ATT ou PRD. Veja-se o erro ocorrente com a forma 'BASEIA-SE'. Em razão de ainda não ser submetida ao algoritmo das formas pronominais, esta forma ainda não é tratada como uma forma verbal. Outro caso curioso verifica-se no quadro 3.19 onde as formas 'PORTUGUE','BEADOR' e 'NUNE' são hipotéticas formas dicionarizadas de 'português', 'Beatriz' e 'Nunes'. Isto acontece porque estas formas são submetidas ao algoritmo NTAB (quadro 3.10) que converte a forma Beatriz pelo quadrigrama à forma BEADOR; português e Nunes são convertidos à forma PORTUGUE e NUNE pelo digrama de NTAB.

No quadro 3.16, em síntese, verificam-se as seguintes abreviações: FRQ indica que a palavra está definida no dicionário de freqüência (FDIC) (quadro 3.8); a abreviação HOM indica uma homografia, constante na tabela NDIC (quadro 3.9); DIC indica que a palavra está incluída no dicionário de verbos, VDIC (quadro 3.7); EST indica que a palavra analisada não pertence ao código lingüístico português ou está com erro ortográfico e não se enquadra no algoritmo NTAB. O autor desta dissertação elaborou um projeto para a correção de erros ortográficos a partir das rotinas do sistema LINGA (conferir, Scher (inédito)).

Após a análise morfológica e a busca nos dicionários (quadros 3.4 e 3.16), cada palavra da frase é submetida à análise contextual (quadro 3.17) preconizada pelo algo

QUADRO 3.16

ANÁLISE MORFOLÓGICA E BUSCA NOS DICIONÁRIOS



ritmo HTAB (quadro 3.11) que define a posição morfossintáti ca da palavra, considerando-se seus traços configuracionais e das palavras vizinhas. Assim, temos neste quadro, as palavras: FEITO que poderá ser NOM ou PAR; a palavra POR poderá ser INF ou PRP; e 1981 que poderá ser NAT ou NNO.

QUADRO 3.17

ANÁLISE CONTEXTUAL

WORD	CAT	PERS	NUM	TIM	G/M	LEXEM	DIC
INTRODUCAO	NOM	0				INTRODUCAO	
	SEP	0					
ESTE	DAT	3	S		M		PRO
ESTUDO	NOM	0				ESTUDO	NOM
BASEIA-SE	ATT	0				BASEIA-SE	
EM	PRP	0					PRO
UM	DET	0	S		M		PRO
TRABALHO	NOM	0				TRABALHO	NOM
SOBRE	PRP	0					PRO
O	DET	0	S		M		PRO
SN	NOM	0				SN	EST
PORTUGUES	ATT	0				PORTUGUE	
FEITO	NOM	3	S		M		PRO
	PAR	0				FEITO	PRO
POR	INF	0				POR	PRO
	PRP	0					PRO
BEATRIZ	NOM	0	P			BEADOR	
NUNES	ATT	0				NUNE	
DE	PRP	0					PRO
OLIVEIRA	NOM	0				OLIVEIRA	
LONGO	ATT	0				LONGO	
	CNJ	0					
UMA	NOM	0				UMA	EST
EM	PRP	0					PRO
1981	NNO	0					PRO
	NAT	0					
	PNT	0					

QUADRO 3.18

ANÁLISE DAS HOMOGRAFIAS

URD	CAT	PES	NUM	TIM	G/M	LEXEM	DIC
NTRODUCAO	NOM	0				INTRODUCAO	
	SEP	0					
ESTE	QAT	3	S		M		FR
STUDO	NOM	0				EST-UDO	NO
ASEIA-SE	ATT	0				BASEIA-SE	
M	PRP	0					FR
UM	DFT	0	S		M		FR
RABALHO	NOM	0				IRAB-ALHO	NO
OBRE	PRP	0					FR
	DET	0	S		M		FR
SN	NOM	0				SN	EST
PORTUGUES	ATT	0				PORTUGUE	
EITO	NOM	3	S		M		FR
OR	PRP	0					FR
BEATRIZ	NOM	0	P			BEA-DOR	
NUNES	ATT	0				NUNE	
DE	PRP	0					FR
LIVEIRA	NOM	0				OLIV-EIRA	
ONGO	ATT	0				LONGO	
	COM	0					
UNB	NOM	0				UNB	EST
1	PRP	0					FR
981	NUM	0					FR
	PNT	0					

Seguindo procedimentos que já descrevemos em 3.6.6 e 6.6.1 temos a análise das homografias (quadro 3.18) que decide a categoria morfossintática definitiva da palavra.

Com esta última etapa realizada, as palavras analisadas como NOM e ATT estarão à disposição num arquivo (quadro 3.19) como candidatos a descritores para a indexação, conforme descrevemos em 5.1.

QUADRO 3.19

CANDIDATOS A DESCRITORES

100	ALGEBRICO
200	ALUNO
300	AREA
400	AREA
500	AREA
600	ASSUNTO
700	ATIVIDADE
800	CAPACIDADE
900	CAPACIDADE
1000	CAPACIDADE
1100	CARGA
1200	CLASSE
1300	CONSTRUCAO
1400	CONTRIBUICAO
1500	CULTIVO
1600	DICOTOMIA
1700	DICOTOMIA
1800	DIDATICA
1900	DIRETA
2000	DISCUSSAO
2100	DIVORCIO
2200	DIVORCIO
2300	EMBASAMENTO
2400	ESCOLAR
2500	ESCOLAR
2600	ESSENCIA
2700	ESTEIO
2800	ESTRUTURA
2900	ESTRUTURA
3000	ESTRUTURA
3100	ESTRUTURA
3200	ESTRUTURA
3300	ESTRUTURA
3400	ESTUDANTE
3500	ESTUDO
3600	EXPLICACAO
3700	FASE
3800	FORMALISMO
3900	GERATIVA
4000	GRAMATICA
4100	GRAMATICA
4200	GRAMATICA
4300	GRAMATICAL
4400	HUMANO
4500	INFORMACAO
4600	INTERPRETACAO
4700	LEIGA
4800	LEITOR
4900	LEXICAL
5000	LINGUA
5100	LINGUA
5200	LINGUA
5300	LINGUA
5400	LINGUA
5500	LINGUAGEM
5600	LINGUISTICA
5700	LINGUISTICA
5800	LINGUISTICA
5900	LINGUISTICA
6000	LINGUISTICA
6100	LINGUISTICA
6200	LINGUISTICA
6300	LINGUISTICA
6400	LINGUISTICA
6500	LINGUISTICA
6600	LINGUISTICA
6700	LINGUISTICA
6800	LINGUISTICA
6900	LINGUISTICO
7000	LOGICA
7100	MATURIA
7200	MATERIAL
7300	MATERNA
7400	NIVEL
7500	NIVEL
7600	PAR
7700	PESSOA
7800	PESSOA
7900	POLO
8000	PONTE
8100	POSICAO
8200	POSICAO
8300	PRATICA
8400	PROFESSOR
8500	PROFESSOR
8600	PROFESSOR
8700	PROFESSOR
8800	PROFUNDA
8900	PURA
9000	QUESTIONAMENTO
9100	REAPROXIMACAO
9200	REGRA
9300	REPERTORIO
9400	REPRESENTACAO
9500	RESPEITO
9600	SEDE
9700	SEDE
9800	SEMANTICA
9900	SER
10000	SERIE
10100	SIGNIFICADO
10200	SINTAXE
10300	SUPERFICIAL
10400	SUPERFICIAL
10500	SUPERFICIAL
10600	TAREFA
10700	TAREFA
10800	TEORIA
10900	TEORIA
11000	TEORIA
11100	TEORIA
11200	TEORICA
11300	TEORICA
11400	TEORICO
11500	TEORICO
11600	TEXTO
11700	TEXTO
11800	TEXTO
11900	TEXTO
12000	TRABALHO
12100	TRABALHO
12200	TRABALHO
12300	TRABALHO
12400	TRABALHO
12500	TRABALHO
12600	TRANSFORMACAO
12700	VOCABULO

3.6.7 As categorias do léxico no LINGA

Segundo Lemle (1984: 96-101), o léxico do português pode ser descrito de maneira satisfatória com as seguintes categorias: nome (N), adjetivo (ADJ), determinante (DET), quantificador (QUANT), verbo (V), preposição (PREP), advérbio (ADV), complementador (COMP), conjunção (CONJ), antequessor (QU). São dez categorias ao todo. No sistema LINGA, como já vimos em 3.6.1.2, temos 33 categorias desenvolvidas. Na verdade, estas 33 categorias são desdobramentos das 10 categorias do léxico especificadas por Lemle, conforme poderemos observar pelo cotejo a seguir: (ver quadro 3.20 abaixo).

QUADRO 3.20
AS CATEGORIAS DO LÉXICO NO LINGA

Proposta de Lemle		Proposta do LINGA
Nome	N	NOM (01), NNO (07), NML (12) SPR (13), OPR (029)
Adjetivo	ADJ	ATT (02), NAT (08), QAT (11) PRD (06)
Determinante	DET	DET (09)
Quantificador	QUANT	ONT (20)
Verbo	V	VRB (03), INF (04), PAR (05) AUX (14), GER (23), VBM (15) AUI (16), MDI (17)
Preposição	PREP	PRP (18), ELP (19)
Advérbio	ADV	AVP (22), ADV (26), NEG (28) DEM (21)
Complementador	COMP	SUB (25)
Conjunção	CONJ	CON (24), SUB (25), COM (32)
Antequessor	QU	REL (10)

QUADRO 3.21

TEXTO A

001100901
O OBJETIVO DESTE TRABALHO E LANCAR UMA PONTE ENTRE A LINGUISTICA
TEORICA E O ENSINO ESCOLAR DA GRAMATICA. A CONTRIBUICAO LEVADA
PELA LINGUISTICA AS PRATICAS ESCOLARES E, NA MELHOR DAS HIPOTESES,
PUERIL, E, EM ALGUNS CASOS, PERNICIOSA. E, NO SENTIDO INVERSO, E
TAMBEM DESPREZIVEL O INVESTIMENTO LEVADO A PESQUISA LINGUISTICA
PARTIR DA LABUTA DIDATICA.
ESSE DIVORCIO QUE SE VERIFICA ENTRE AS DUAS AREAS SO PODE ADVIR DE
POSICOES FALSAS DE UMA OU DE AMBAS AS PARTES , POIS UMA TEORIA
QUE TEM COMO OBJETO A CAPACIDADE DE LINGUAGEM DO SER HUMANO NAO
PODERIA DEIXAR DE TER RELEVANCIA NA PARTE DA TAREFA EDUCACIONAL
QUE DIZ RESPEITO AO CULTIVO DA CAPACIDADE DE COMPREENDER E UTILIZAR
A LINGUA.
A QUE ATRIBUIR O DESENCONTRO? LIMITO A DISCUSSAO A AREA DA SINTAXE.
AVENTURO-ME A ATRIBUIR A CARGA MAIOR DO DESENCONTRO AO FOLO
LINGUISTICO DO PAR. E CONSIDERO QUE GRANDE PARTE DA EXPLICACAO PARA
ESSE DESENCONTRO RECAI SOBRE A DICOTOMIA QUE CONSTITUI A CONSTRUCAO
TEORICA BASILAR DE UMA CERTA FASE DA TEORIA DA GRAMATICA GERATIVA,
A CELEBRE E MAL COMPREENDIDA DICOTOMIA ENTRE ESTRUTURA PROFUNDA
E ESTRUTURA SUPERFICIAL.
E FACIL DE VER A LOGICA DA INCOMPATIBILIDADE ENTRE TAL TEORIA E A
ATIVIDADE PRATICA DO PROFESSOR DE LINGUA. SE A ESSENCIA DA TAREFA
DO PROFESSOR DE LINGUA MATERNA E A DE ESTIMULAR NO ESTUDANTE A
CAPACIDADE DE COMPREENDER E PRODUZIR TEXTOS, SEU OBJETO DE TRABALHO
TENDE A SER O TEXTO EM SI, EM SUA ESTRUTURA SUPERFICIAL. OS
VOCABULOS A INTRODUZIR AO REPERTORIO LEXICAL DO ALUNO, AS
ESTRUTURAS A DISSECAR, OS SIGNIFICADOS A DEPREENDER, TUDO ISSO
DEVE TER SEDE DIRETA NO MATERIAL LINGUISTICO CONCRETAMENTE
PRESENTE NOS TEXTOS UTILIZADOS NOS TRABALHOS EM CLASSE. CONTUDO,
UMA CERTA LINGUISTICA ALEGA NAO SER A ESTRUTURA LINGUISTICA
CONCRETAMENTE PRESENTE NO TEXTO A SEDE DA INTERPRETACAO SEMANTICA,
E SIM UM OUTRO NIVEL DE REPRESENTACAO, UM NIVEL ABSTRATO, BEM
DIVERSO DA ESTRUTURA SUPERFICIAL, E SO INDIRETAMENTE LIGADO A ELA
MEDIANTE UMA SERIE DE REGRAS GRAMATICAIS DENOMINADAS TRANSFORMACOES,
DESCRITAS POR MEIO DE UM FORMALISMO ALGEBRICO BIZARRO, ABOMINAVEL E
DESINTERESSANTE PARA A MAIORIA DAS PESSOAS. O DIVORCIO ERA INEVITAVEL.
NESTE ESTUDO TENTAREI APONTAR UM CAMINHO DE REAPROXIMACAO ENTRE AS
DUAS AREAS DE TRABALHO, A DA LINGUISTICA PURA E A DA LINGUISTICA
APLICADA AO ENSINO DA GRAMATICA.
O TRABALHO VISA SERVIR A LEITORES DIVERSAMENTE DOTADOS DE INFORMACOES
SOBRE LINGUISTICA: PROFESSORES DE LINGUISTICA QUE JA TEM SUAS PROPRIAS
POSICOES DENTRO DA TEORIA LINGUISTICA, PROFESSORES DE LINGUA QUE DESEJA
UM ENQUADRAMENTO TEORICO MAIS AMPLO PARA SEU TRABALHO, ESTUDIOSOS DE AREA
AFINS A LINGUISTICA QUE NECESSITAM DE UM ESBOCO TEORICO DA LINGUISTICA,
E, EVIDENTEMENTE, PESSOAS LEIGAS QUE SE INTERESSAM POR ASSUNTOS RELATIVOS
A LINGUA.

Como podemos observar neste quadro, à categoria do nome (N) correspondem no LINGA as seguintes categorias morfossintáticas: NOM (01), NNO (07), NML (12), SPR (13) e OPR (29). Ressalte-se que estas categorias não são nomes sob o aspecto morfológico, mas o são sob o aspecto sintático. À categoria do adjetivo (ADJ) correspondem no LINGA as seguintes categorias: ATT (02), NAT (08), QAT (11), PED (06). Ao determinante (DET) corresponde DET (09) do LINGA. Ao quantificador (QUANT) corresponde QNT (20). A verbo (V) correspondem VRB (03), INF (04), PAR (05), AUX (14), GER (23), VBM (15), AUI (16), MDI (17). À preposição (PREP) corresponde PRP (18) e ELP (19). Ao advérbio (ADV) corresponde AVP (22), ADV (26), NEG (28) e DEM (21). Ao complementador (COM) corresponde SUB (25). Às conjunções CONJ correspondem, no LINGA, CON (24), SUB (25) e COM (32). Ao antequessor (QU) corresponde REL (10).

Para definir as classes, Lemle caracteriza cada uma morfologicamente e aponta suas funções sintáticas e semânticas. No sistema LINGA algo semelhante sucede. Em primeiro lugar as palavras são definidas em seus aspectos morfológicos (ver quadro 3.4), para em seguida serem definidas em seus aspectos sintáticos.

A categoria nome (NOM) tem os traços morfológicos de gênero e número que no algoritmo NTAB (Quadro 3.10) definem a categoria nominal e reduzem a palavra à sua forma dicionarizada. A função sintática dos nomes é a de ocupar posição sintática do nome, como já vimos, as categorias NNO,

NML, SPR, OPR. Essas posições sintáticas, no sistema LINGA, são marcadas na rotina HTAB (quadro 3.11). A função semântica dos nomes, segundo Lemle, é a de referir. Daí concluir eu serem os candidatos naturais a descritores de um texto.

Os traços morfológicos do verbo expressos pela desinências de modo, pessoa e número, definidos por VTAB (quadro 3.6), reduzem a forma verbal flexionada à sua forma infinitiva, que, então, será confrontada com as formas armazenadas em VDIC (quadro 3.7). No LINGA as seguintes categorias ocupam a mesma função sintática: INF, PAR, AUX, VMB, AUI, MDI e GER. Como se percebe, são sub-classificações da forma verbal, o que proporciona no sistema LINGA maior precisão na análise, maior economia e rapidez no processamento dos dados.

Sob o aspecto morfológico, os adjetivos assemelham-se aos nomes quanto às desinências. Por esta razão, oferecem dificuldades para sua definição sob o aspecto sintático no sistema LINGA. Segundo Lemle, os adjetivos se adjungem sintaticamente ao nome e seu papel semântico é o de expressar uma qualidade do referente do nome. Razão pela qual, constituem-se em elementos preferenciais do descritor composto, conforme veremos em 4.1.1.2 a seguir. No LINGA, ocupam a mesma posição sintática do adjetivo as seguintes categorias: NAT, QAT e PRD.

Os determinantes e os quantificadores, segundo Lemle (1984: 150-1), não se distinguem, a nível sintático, dos adjetivos. Discutimos este aspecto em maior profundida

de em 4.1 adiante. No LINGA; todos os elementos destas duas categorias (DET e QNT) estão marcados no FDIC (quadro 3.8) que descrevemos em 3.6.2.

A categoria dos complementadores (COMP) e das conjunções (CONJ) referem-se às conjunções subordinativas (SUB) e às conjunções coordenativas (CON), que também estão marcadas no FDIC do LINGA.

Lemle criou o termo antequessor para referência ao pronome relativo e às palavras interrogativas, que formam a classe das denominadas "palavras QU". A categoria REL (10) do LINGA, marcada no FDIC, cobre essa classe. Essa categoria, assim como a dos complementadores e das conjunções, são importantes no sistema LINGA, porque são marcadores de limites de oração - que é a unidade máxima de análise automática do sistema.

Como se percebe pela análise das categorias do léxico proposta por Lemle (1984) e seu cotejo com as 33 categorias desenvolvidas no LINGA (quadro 3.12), a estrutura morfossintática da língua portuguesa é abrangida em sua amplitude pela análise automática do LINGA. As categorias gramaticais, constantes no FDIC, cobrem os sistemas fechados da língua; as categorias NOM e ATT são tratadas por NTAB e NDIC, enquanto a categoria verbal é tratada por VTAB e VDIC.

CAPÍTULO IV

O TRATAMENTO DO SINTAGMA NOMINAL DA LÍNGUA PORTUGUESA NO LINGA

Na introdução desta dissertação, procuramos deixar evidente que a LC contribui para que a análise lingüística seja executada de modo mais preciso e eficiente. Em outras palavras, a precisão e a formalização das regras lingüísticas são condições para que o computador se torne um efetivo auxiliar do analista. Neste trabalho escolhemos o SN português pelas razões aduzidas, a saber: 1) já existem inúmeros trabalhos a respeito na teoria gerativa; 2) o estudo do SN vem ao encontro de nosso objetivo prático, qual seja o de investigar como a teoria lingüística pode contribuir para a indexação automática; 3) a análise do SN nos permite discutir as implicações da teoria $\bar{X}$ e apresentá-la como uma abordagem possível para colimar o objetivo expresso no item 2 acima.

Neste capítulo procuramos mostrar não só como o sistema LINGA é uma formalização da estrutura lingüística do português, mas também como essa estruturação está ancorada em recentes postulações da teoria gerativa.

Além disso, pretendemos mostrar como as regras postuladas pelo modelo clássico da teoria gerativa são redundantes com as informações do léxico, que no LINGA estão codificadas na rotina HTAB. Em outras palavras, o processo de um sistema de análise lingüística automática não pode prescindir de uma base teórica sólida que forneça e corrobore as categorias analíticas que serão empregadas. O simples conhecimento da língua, para o falante nativo, não seria suficiente para a elaboração de um tal sistema. É requerida, efetivamente, a familiaridade com os recursos analíticos da lingüística moderna. A análise do SN, nesse capítulo, é apresentada como ilustração de como os resultados da análise lingüística de uma língua natural são utilizados pela lingüística computacional em seu trabalho. Observe-se que, no caso em questão, valemo-nos do modelo da regência e ligação* da teoria gerativa para fornecer subsídios à identificação dos descritores simples e compostos. Observe-se, ainda, que a LC vale-se também de análises fonológicas, abordagens semânticas e pragmáticas, dependendo dos objetivos a alcançar

4.1 O sintagma nominal no LINGA

O tratamento do SN, assim como das demais categorias sintagmáticas (SV, SA, SP, SAdv, etc.), recebeu uma modificação radical na atual versão da teoria gerativo-transformacional. Na ótica do modelo padrão, propunha-se para cada categoria sintagmática uma regra de expansão (dita regra sintagmática ou regra categorial) que reescrevia o sím-

bolo sintagmático em questão em outros símbolos categoriais e/ou lexicais Por exemplo, SN se reescrevia nos elementos abaixo, entre outros:

(1) SN ⟶ (Art) (Poss) N (SP)

onde os parênteses indicam opcionalidade do elemento circundado. Na ótica da versão atual do modelo padrão estendido (muitas vezes denominado de teoria da regência e ligação), não mais se postulam regras específicas para a expansão de SN, SV, SA, SP ou SAdv. Essa mudança de perspectiva decorreu da constatação da redundância, no modelo padrão, entre, de um lado, as regras sintagmáticas e, de outro, as informações lexicais sobre subcategorização. Com efeito, as regras sintagmáticas não são mais do que generalizações a partir das informações sobre as subcategorizações. Por outro lado, as regras sintagmáticas não dispensam as subcategorizações no léxico. Por exemplo, a informação constante da expansão do SN, de que um nome pode ser precedido de artigo, não dispensa a subcategorização dos substantivos a respeito dessa coocorrência, pois nem todo substantivo aceita coocorrer com artigo. Por exemplo, o nome 'marte':*o marte.

No âmbito da teoria padrão, diferentes propostas foram feitas com relação ao SN português. Entre elas, podemos citar as de Pazini (1977, 1978), Lemle e Naro (1977) , Lemle (1978) e Lobato (1979). A proposta de Lemle (1984) também apresenta características de uma análise padrão, mas com uma particularidade: procura captar generalizações in-

tercategoriais, na linha da abordagem inicial da teoria $\bar{X}$.

Como a tendência atual, no âmbito da teoria padrão estendida, é eliminar as regras de expansão, que são redundantes com as informações do léxico, evidentemente, não se tem propostas específicas a respeito do SN em português nesta versão da teoria.

Transportando essa tendência da teoria padrão estendida para o sistema LINGA, percebemos que ela é mais simples do que uma proposta conforme o modelo padrão, em que se trataria o SN por um algoritmo específico, tendo esse algoritmo a forma da regra de expansão de SN. Percebemos, também que além de simplificar o sistema, essa tendência torna os resultados da análise operada pelo LINGA mais adequados aos dados da língua.

A estruturação do algoritmo com base na regra de expansão do modelo padrão tem pelo menos três desvantagens para a análise automática: 1) a geração de seqüências inaceitáveis; 2) a alta complexidade do algoritmo e 3) sua redundância com as informações contidas na rotina HTAB. Em primeiro lugar, os resultados finais deporiam contra sua implementação no sistema de análise automática (LINGA), porque esse algoritmo permitiria gerar seqüências inaceitáveis, como:

(2) a) * ambos terceiros amigos;

b) * algum vosso mesmo amigo, etc.

embora todos esses elementos pré-nominais estejam previstos na regra padrão de expansão do SN, que teria uma configuração parecida com (3) (cf. Lobato (a sair): §§ 3.5 e 12.2.1.1):

(3) $SN \rightarrow \left\{ \begin{matrix} (Quant)\ Art \\ (Indef) \end{matrix} \right\} (Poss)\ (Id)\ (Card)\ (Ord)\ (Delim)\ N\ (SA)^n\ (SP)^n\ (S)^n$

Em (3) os termos quantificador (todos, ambos), identificador (outro, mesmo) e delimitador (muito, poucos, diversos, etc.) não são rótulos da gramática tradicional. Os termos identificador (Id) e delimitador (Delim) são criações de Lobato. As chaves indicam exclusão dos elementos que delimitam; os parênteses em ((Quant) Art) indicam a obrigatoriedade de Art quando se escolhe Quant. O índice 'n' em SA, SP, S expressa a possibilidade de ocorrência n vezes da categoria sintagmática por ele marcado.

Em segundo lugar, o algoritmo seria complexo e os resultados de sua adoção não seriam adequados, como vimos em (2). Em terceiro lugar, como já apontamos e como veremos em mais detalhe em 4.3, a regra do SN é redundante com as informações do léxico, que no LINGA estão codificadas na rotina HTAB (quadro 3.11). Ora, se a rotina HTAB já cobre todas as particularidades do SN (como também das outras categorias sintagmáticas), não há necessidade nenhuma de se introduzir um algoritmo especial para sua análise.

Voltando a (3), ela cobriria SNs como:

(4) a) Todos aqueles meus outros valiosos livros.

b) Nossos três primeiros presidentes brasileiros militares revolucionários.

c) Nenhum livro relevante de linguística que seja atual.

d) Que você compreenda tudo facilmente (é óbvio).

Mas, como bem observa Lobato (a sair), ela é uma simplificação da regra que o modelo padrão teria de propor, pois ainda deixa de cobrir fatos como a possibilidade de alguns de seus elementos ocuparem diferentes posições dentro do SN, conforme ilustram os exemplos a seguir:

(5) a) O meu outro casaco
b) O outro meu casaco
c) O outro casaco meu

d) Os meus primeiros amigos
e) Os primeiros amigos meus
f) Os primeiros meus amigos

Como explicar os diferentes posicionamentos de certos elementos dentro do SN? Segundo a teoria padrão, pondo regras transformacionais de deslocamento. Sobre o português, as análises propostas se detiveram especificamente na variedade de posicionamentos dos elementos tradicionalmente denominados adjetivos.

Conforme Pazini (1978: 109), para que as transformações de deslocamento do adjetivo pudessem ocorrer, era

preciso que se preenchessem três condições:

1) o adjetivo ser marcado no léxico com o traço [+ gradação];

2) o adjetivo vir de uma sentença relativa não-restritiva;

3) o adjetivo não apresentar complemento [+ _____SP] no momento de se aplicar a regra transformacional de deslocamento.

Lobato (1979:1) estabelece uma distinção entre as palavras da categoria MERO e os adjetivos propriamente ditos. As palavras da categoria MERO: 1) só ocorrem em posição pré-nominal; 2) não aceitam a incidência de intensificadores (muito, extremamente); 3) não podem integrar construções comparativas. Os adjetivos propriamente ditos podem ocorrer antes e depois do núcleo nominal. Vejam-se os exemplos abaixo:

6) a) João é um homem simpático.

a') João é um homem que é simpático.

b) Foi um mero acidente.

b') * Foi um acidente que foi mero.

c) Luís é um simples artista.

c') * Luís é um muito simples artista.

d) * Luís é mais mero que do que seu colega.

Os adjetivos propriamente ditos podem receber a incidência de intensificadores que, em Pazini, são traduzidos pelo traço [± gradação] ou integrar construções comparativas, enquanto as palavras da classe MERO não podem.

Veja-se o exemplo (7):

7) a) João é um homem muito simpático.

b) * João é um homem muito mero.

No LINGA, as palavras da classe MERO são marcadas no NDIC (quadro 3.9) na coluna 33 que identifica os adjetivos na posição pré-nominal (ver seção 3.6.4.2).

Contreras (1979), por sua vez, questiona a derivação transformacional do adjetivo e alinha-se com os teóricos gerativistas que propõem a marcação,diretamente no léxico, da combinatória contextual dos membros dessa classe e sua geração diretamente na base como constituintes do SN. Os adjetivos atributivos seriam, então, classificados de acordo com a posição que ocupam antes e/ou depois do nome (N), tendo-se, então, diferentes subcategorizações expressas no léxico por meio das seguintes marcações: + [_____N]; +[N_____]; + [N____], + [_____N].

Para expressar a similaridade de restrições selecionais entre os adjetivos atributivos e predicativos, Contreras adota a concepção do léxico postulada por Hust em 1976 (apud Contreras) segundo a qual o verbete lexical é em forma de diagrama em ramificações, no qual o nódulo superior contém todos os traços comuns aos diferentes sentidos da palavras e as ramificações apresentam feixes de traços que caracterizam uma leitura particular da palavra (V. exemplo 10).

Para os adjetivos como 'antigo', 'novo', que têm diferentes interpretações, dependendo de sua posição em relação ao nome, cada traço de subcategorização é associado a um distinto feixe de traços semânticos.

Contreras apresenta, a título de exemplo, a entrada para o adjetivo 'antiguo' que, antes do nome, tem um significado de 'antigo' (former em inglês) e, depois do nome, significa 'velho', (old em inglês) sendo que, no priemiro sentido, pode ser associado a nomes animados e inanimados, como em (9a), mas no segundo sentido associa-se só a nomes inanimados, como em (9b):

(9) a) Vi a mi antiguo maestro.
Eu vi meu antigo professor.
'I saw my former teacher'.

b) Este es un edificio antiguo.
Este é um edifício antigo
'This is an old building'.

c) Este edificio es antiguo.
Este edifício é antigo.
'This building is old'.

d) Esa es mi antigua casa.
Esta é minha antiga casa.
'That is my former house'.

e) * Mi maestro es antiguo.
Meu professor é antigo.
'That is my old teacher'.

f) * Ese es mi maestro antiguo.
Esse é meu professor antigo.
'That is my old teacher'.

O verbete lexical para o adjetivo 'antiguo' será a seguinte:

(10) 'antiguo'

[+ Adjective]

[other common features]

+[_____N] 'former'

+[N_____] +[Cop _____] -[[+animate] X____] 'old'

In: Contreras, 1979: 19

Em (10), a ramificação à esquerda contém o traço de subcategorização e os traços semânticos que dão ao adjetivo o sentido de 'antigo' ('former'). Não havendo traços de seleção, o adjetivo poderá ocorrer com nomes animados e inamimados como em (9 a) e (9 d).

A ramificação à direita em (10), indica que o adjetivo 'antiguo' só pode ocorrer em posição pós-nominal +[N_____] ou no predicado +[Cop_____]. Seus traços semânticos dão-lhe a leitura de 'velho' ('old') e os traços selecionais -[[+ animate] X____] não o deixam ocorrer com nomes animados, como em (9 e) e (9 f).

Essa proposta de marcação dos adjetivos diretamente no léxico parece-nos ser a melhor por se tratar de uma solução simples que não exige transformações e a aplicação de mecanismos de exceção e, também, porque se enquadra perfeitamente ao sistema LINGA, dado que os adjetivos com marcação +[_____N] estão assinalados em NDIC na coluna 33

(quadro 3.9). A partir de NDIC, esses adjetivos são posicionalmente definidos por HTAB. Os adjetivos com a marcação + [N_____] e + [Cop_____] são definidos normalmente em sua posição sintática por HTAB, sem que seja necessário marcá-los no NDIC.

Lemle (1984), na ótica das primeiras versões da teoria $\bar{X}$, apresenta uma proposta que procura captar as similitudes intercategoriais de vários elementos que integram a regra do SN do modelo padrão. Segundo essa proposta, os elementos Poss, Indef, Card, Ord e Adj passariam a integrar uma supercategoria do adjetivo (ADJ) pois poderiam ser intercambiados com o adjetivo e serem considerados como tais. E, como adjetivos, poderiam coocorrer entre si.

Apesar de serem considerados como integrantes da supercategoria ADJ, manteriam seus traços específicos de [+Poss] , [+Indef], [+Card], [+Ord] e [+Aux], para serem diferenciados entre si. Deste modo, para referência à classe exclusiva dos pronomes possessivos seriam usados os traços $\begin{bmatrix}+\text{ADJ}\\+\text{Poss}\end{bmatrix}$; para os pronomes indefinidos, seriam usados os traços $\begin{bmatrix}+\text{ADJ}\\+\text{Indef}\end{bmatrix}$; para os numerais cardinais, seriam usados os traços $\begin{bmatrix}+\text{ADJ}\\+\text{Card}\end{bmatrix}$; para os numerais ordinais, seriam usados os traços $\begin{bmatrix}+\text{ADJ}\\+\text{Ord}\end{bmatrix}$ e para os tradicionais adjetivos, seriam usados os traços $\begin{bmatrix}+\text{ADJ}\\+\text{Adj}\end{bmatrix}$. Parece-nos, entretanto, que essa proposta não muda essencialmente a regra do SN do modelo pa

drão porque a redundância entre o componente categorial e o léxico persiste.

A regra do SN (V. Lemle, 1984: 150-1) com a proposta da supercategoria ADJ passaria a ter uma configuração parecida com:

(12) SN ⟶ ((Quant) Det) (ADJ)* N (ADJ)* (SPrep)* (Adj)* (S)*

onde os asteriscos indicam que os elementos por ele assinalados podem 'n' vezes.

Essa proposta procura, evidentemente, evitar o uso de transformações de deslocamento aplicadas a elementos integrantes do SN. Mas ela apresenta dificuldades tais de realização que a colocam em pé de igualdade com a proposta padrão. Por exemplo, os adjetivos propriamente ditos admitem a coocorrência entre si, como em (13):

(13) a) garotas esbeltas barragarcenses
b) garotas barragarcenses esbeltas

porém, nem todos os elementos da supercategoria ADJ podem coocorrer entre si, como (14):

(14) a) * vários muitos deputados
b) * muitos vários deputados

c) * certo cada aluno
d) * cada certo aluno

e) * outro mesmo caso
f) * mesmo outro caso

g) * diversos quatro amigos

h) * quatro diversos amigos, etc.

Como explicar essas incompatibilidades entre os elementos da supercategoria ADJ, a não ser por intermédio de indicações explícitas nos verbetes do léxico, conforme a proposta da teoria da regência e ligação? Mais uma vez iremos perceber o acerto de nossa proposta de marcação lexical desses elementos, assim como dos demais elementos do SN, na rotina HTAB que, como já vimos em 3.6.6.1, exerce o papel do léxico no LINGA. Ou seja, cada elemento da supercategoria ADJ já está marcado em relação aos demais itens lexicais em HTAB, obedecendo-se a seus traços configuracionais.

4.2 Interrogações sobre o SN no LINGA

Até o momento abordamos o SN na ótica do modelo da teoria da regência e ligação. Seus elementos estão detalhados em (3) e (12) neste capítulo. Uma pergunta merece ser feita: Como o sistema LINGA trata o SN da língua portuguesa, uma vez que ocorre no sistema lingüístico redundância entre o componente categorial e as informações no léxico na proposta padrão da regra do SN? Já vimos em 3.6.2, 3.6.3, 3.6.4, 3.6.5 e 3.6.6 que o sistema LINGA, através da análise morfológica e da busca nos dicionários (quadro 3.4) identifica todas as palavras de uma frase em sua estrutura morfológica. A partir dessa identificação, as posições sintáticas das palavras são indicadas pela rotina HTAB, obedecendo-se a suas configurações. Ora, havendo a identificação morfológica das palavras e suas posições sintáticas, ou melhor, havendo a

definição das coocorrências de cada item em relação aos demais, necessariamente toda a estrutura sintagmática do SN estará prevista, sem que haja necessidade de algoritmos específicos para sua definição. Para comprovar, vejamos os resultados da análise da frase abaixo (4) operados pelo LINGA sem auxílio de nenhum algoritmo especial para o SN:

QUADRO 4.1

```
DICTIONARY SEARCH AND MORPHOLOGY
WORD                CAT PERS NUM TIM G/M LEXEM                    DIC
--------------------------------------------------------------------
TODOS               ADJ   0    P       M                          FR
                    NML   0    P       M                          FR
AQUELES             ADJ   3    P       M                          FR
                    NML   3    P       M                          FR
MEUS                ADJ   1    P       M                          FR
OUTROS              ADJ   0    P       M                          FR
                    NML   0    P       M                          FR
DEZ                 NUM   0    P                                  FR
PRIMEIROS           ADJ   0    P                                  FR
                    NOM   0    P       M  PRIMEIRO
                    ATT   0              M  PRIMEIRO
                    PRD   0              M  PRIMEIRO
ESTRANHOS           NOM   0    P       M  ESTRANHO
                    ATT   0              M  ESTRANHO
                    PRD   0              M  ESTRANHO
POEMAS              NOM   0    P       M  POEMA
                    ATT   0              M  POEMA
                    PRD   0              M  POEMA
                    PNT   0

AFTER CONTEXT CHECK
DICTIONARY SEARCH AND MORPHOLOGY
WORD                CAT PERS NUM TIM G/M LEXEM                    DIC
--------------------------------------------------------------------
TODOS               NML   0    P       M                          FR
AQUELES             NML   3    P       M                          FR
MEUS                ADJ   1    P       M                          FR
OUTROS              ADJ   0    P       M                          FR
DEZ                 NUM   0    P                                  FR
PRIMEIROS           ATT   0              M  PRIMEIRO
ESTRANHOS           ATT   0              M  ESTRANHO
POEMAS              NOM   0    P       M  POEMA
                    ATT   0              M  POEMA
                    PNT   0

HOMOGRAPH ANALYSIS
WORD                CAT PERS NUM TIM G/M LEXEM                    DIC
--------------------------------------------------------------------
TODOS               NML   0    P       M                          FR
AQUELES             NML   3    P       M                          FR
MEUS                ADJ   1    P       M                          FR
OUTROS              ADJ   0    P       M                          FR
DEZ                 NUM   0    P                                  FR
PRIMEIROS           ATT   0              M  PRIM-EIRO
ESTRANHOS           ATT   0              M  ESTR-ANHO
POEMAS              NOM   0    P       M  PO-EMA
                    PNT   0
```

Neste quadro, na etapa Dictionary Search and Morphology (Análise morfológica e busca nos dicionários) são definidas as possíveis categorias de cada palavra por exemplo a palavra 'POEMAS' poderá ser um NOM, ATT ou PRD. Na etapa seguinte After context check (Análise contextual) as palavras são analisadas em suas configurações por HTAB para definir os seus vizinhos da esquerda e da direita, de modo que cada palavra tenha definida sua identidade morfológica. Dessa etapa restam as palavras homográficas As homografias são analisadas na etapa Homograph Analysis (Análise das homografias). Por exemplo, na etapa anterior, a palavra 'POEMAS' pode ainda ser definida como NOM, ATT em razão de sua desinência poder ser de ambas as categorias. Nesse caso, se a palavra não estiver marcada na coluna 33 de NDIC, esta será considerada como NOM. Assim, a análise morfossintática da frase estará completa e, no caso específico do SN, sua configuração terá sido abrangida sem auxílio de algoritmo específico do SN baseado na regra de expansão do modelo padrão.

Se o algoritmo do SN, proposto com base na regra de expansão do SN do modelo padrão, é redundante com os resultados apresentados pelo sistema LINGA, então não há a necessidade de implementá-lo no sistema. Veremos, a seguir, que a teoria da regência e ligação corrobora esta afirmação.

4.3 O modelo da regência e ligação: a projeção do SN

Como já dissemos na seção 4.1, no modelo padrão (V. Lobato, seção 12.2.1.1 (a sair)), o componente categorial era constituído por uma série de regras que expandiam categorias sintagmáticas (S, SN, SV, SA, SAdv, SP) em categorias lexicais (Art, N, V, etc.) ou sintagmáticas (SN, SV, etc.), e/ou certos morfemas gramaticais (como tempo e os elementos (i + -R), (te + -DO), (esta + - NDO). Em (3), da seção 4.1, vimos a regra de expansão de SN que especifica o esquema de subcategorização em que a categoria lexical N pode ocorrer. Por sua vez, o léxico no modelo padrão era um conjunto de verbetes lexicais, cada um com seus traços contextuis específicos (traços de subcategorização estrita e traços selecionais), entre outras informações.

Segundo Lobato (ibidem), "tornou-se evidente que há redundância, nesse modelo, entre o léxico e as regras sintagmáticas: as informações sobre subcategorizações estritas são traduxidas por ambos". Constatação semelhante estará ocorrendo com a nossa pretensão original de inserir no sistema LINGA em algoritmo específico do SN, baseado em sua regra de expansão: esse algoritmo seria redundante em relação às informações sobre os itens lexicais e gramaticais contidas na rotina HTAB do LINGA.

Diante da constatação da redundância do léxico e do componente categorial para expressão das subcategoriza-

ções estritas, os gerativistas modificaram a teoria nesse aspecto. Assim, as subcategorizações estritas passaram a ser expressas no léxico e o componente categorial não mais contém as regras de expansão do SN, SV, etc. Apenas contém os princípios da teoria $\bar{X}$.

Neste ponto, cabe uma digressão sobre a dicotomia gerativista entre gramática universal (GU) e gramática particular. Na GU estão as propriedades lingüísticas universais, inatas ao homem, e nas gramáticas particulares estão os fatos referentes a cada língua particular. Os princípios da teoria $\bar{X}$ são vistos, nesse enfoque, como integrando a GU.

No seu início, a convenção $\bar{X}$ continha esquemas universais de regras sintagmáticas, como em (15):

(15) a) $\bar{\bar{X}} \longrightarrow$ Espec $\bar{X}$
b) $\bar{X} \longrightarrow$ X Comp

onde Espec é símbolo de especificador (ou modificador) e Comp de Complemento e onde X é uma variável referente às categorias lexicais N, V, A e P. A variável X, sem barra, indica tratar-se do núcleo lexical. A variável X, com 2 barras, indica tratar-se do sintagma como um todo. A variável, com uma barra, indica tratar-se de nível intermediário, entre X e $\bar{X}$. As regras em (15) intentam captar a generalização de que as categorias lexicais N, V, A e P (referidas pela variável X) são, todas elas, seguidas de complementos e antecidas de especificadores ou modificadores. Em outras palavras, a previsão contida nas versões iniciais dessa

teoria era a de um isomorfismo entre os sintagmas que tinham por núcleo cada uma das referidas categorias lexicais. Assim, a previsão de colocação para os especificadores dessas categorias lexicais era expressa pela regra (15 a) e a previsão de colocação para os complementos dessas categorias lexicais era expressa pela regra (15 b).

No entanto, existem contra-argumentos à generalidade das regras em (15). No próprio português encontramos exemplos, como em (5) da seção 4.1, em que os especificadores ocorrem à direita do núcleo do SN. E por outro lado, nas línguas SOV os complementos têm sua posição mais neutra à esquerda do verbo.

Atualmente, a convenção $\bar{X}$ não contém mais as regras em (15), que eram regras de expansão (também denominadas regras de reescrita ou regras sintagmáticas). Na perspectiva atual, uma derivação pode-se iniciar a partir de qualquer unidade lexical, por exemplo: a palavra cantar ou a palavra construção. A partir da palavra escolhida, constrói-se a estrutura, fazendo-se projeções a partir dela, obedecidas as informações contidas no léxico. A convenção $\bar{X}$ se resume, nessa abordagem, em dois princípios: (1) o de que as categorias lexicais básicas N, V, A e P resultam da combinação dos traços sintáticos nominal ([+N]) e verbal ([+V]), e (2) o de que os demais sintagmas são obtidos por projeção dessas categorias básicas (V. Lobato, (a sair) seção 12.2.1.1)). A aplicação da notação em barras sobre essas categorias básicas resultará nas categorias sintagmáticas, projetadas a partir de X e que são, por isso mesmo, as

supercategorias de X, relacionadas entre si por meio da indicação abaixo que integra a GU:

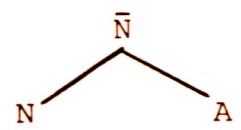
(16) $\bar{X}$ ... X ...

A indicação em (16) deve ser interpretada como uma informação de que as projeções se fazem a partir de X. Ao contrário de (15), ela não se refere à ordem em que se colocam os especificadores e os complementos. No entanto, para a análise automática, interessa a distinção entre elementos que ocorrem à esquerda e à direita de N, porque só os elementos que ocorrem exclusicamente à direita de N são possíveis descritores. Assim, as palavras do tipo MERO, vistas na seção 4.1, e os adjetivos que tanto podem ocorrer à esquerda quanto à direita de N, como bonito e feliz, jamais serão considerados descritores. Mas um adjetivo como computacional, que só ocorre depois de N, é um possível descritor. Dado isso, poder-se-ia pensar em conservar, então, a distinção entre os níveis de X, $\bar{X}$ e $\bar{\bar{X}}$, mas sem se comprometer com a afirmação de que $\bar{X}$ domina X e seus complementos e de que $\bar{\bar{X}}$ domina $\bar{X}$ e seus especificadores, e tampouco com a afirmação de que os complementos ocorrem à direita de X e que os especificadores ocorrem à esquerda de $\bar{X}$.

A razão para esse procedimento seria exatamente o fato de que, entre todos os adjetivos, só os que ocorrem exclusicamente à direita do núcleo do SN podem ser considerados como elementos de um descritor composto. Com a distinção entre X, $\bar{X}$ e $\bar{\bar{X}}$, dir-se-ia, então, que só os adjetivos sempre dominados por $\bar{N}$ são candidatos a descritores.

Em síntese, o princípio que se conservaria, da teoria $\bar{X}$, é o de que as categorias sintagmáticas são projeções de categorias lexicais. Para que isso fosse adequado, seria necessário que houvesse um paralelismo entre colocação à esquerda ou à direita de N e projeção em $\bar{N}$ e $\bar{N}$. Em outras palavras, seria preciso que os adjetivos que se projetam em $\bar{N}$ como em (17):

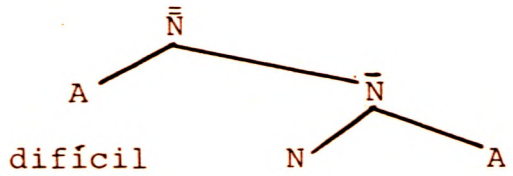
(17)



lingüística computacional

fossem sempre os que podem integrar um descritor composto. Quanto aos adjetivos que se projetam em $\bar{\bar{N}}$, que não podem integrar descritores, como o adjetivo <u>difícil</u>, teriam de se projetar sempre em $\bar{\bar{N}}$, como em (18):

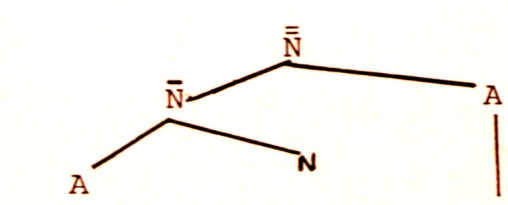
(18)



lingüística computacional

No entanto, isso não ocorre assim, porque o adjetivo <u>difícil</u> pode também, se projetar em $\bar{N}$, como em (19):

(19)



difícil lingüística computacional

Nesse mesmo exemplo, o adjetivo computacional se projeta em $\bar{N}$ e portanto, não poderia ser considerado como elemento de um descritor composto, enquanto o adjetivo difícil seria considerado como elemento de um descritor composto, pois se projeta em $\bar{N}$. Como percebemos, pelos exemplos (18) e (19), a manutenção da convenção $\bar{X}$ fica insustentável para tentar definir a estrutura dos descritores compostos como sendo uma projeção de $\bar{N}$. A solução que temos, então, é considerar apenas a distribuição dos adjetivos em relação ao nome. Ou seja, serão considerados como elementos de um descritor composto, apenas os adjetivos que ocorrerem exclusivamente à direita do núcleo do SN, sem atentar para a sua projeção em barras. Isto significa dizer que estamos usando apenas uma abordagem distribucionalista para os descritores compostos.

CAPÍTULO V

ELEMENTOS DO SINTAGMA NOMINAL PARA A INDEXAÇÃO AUTOMÁTICA

A análise do sintagma nominal realizada no capítulo IV nos oferece os subsídios lingüísticos que poderão contribuir para o aperfeiçoamento do processo da indexação automática na área da teoria da informação. Não foi nosso objetivo específico desenvolver um trabalho experimental nesse campo. Limitamo-nos a oferecer, com a conciliação que fazemos entre a análise lingüística do SN e as rotinas analíticas do LINGA, uma contribuição que poderá ser incorporada a técnicas já existentes de indexação automática. Dentre essas técnicas merece nosso destaque a técnica desenvolvida por Robredo (1983) denominada BIB/DIÁLOGO, onde através de um anti-dicionário de palavras vazias, denominado 'WORDFIXED', e de um dicionário de raízes de palavras, denominado 'WORDROOTS', realiza-se um processo de filtragem de palavras que permite uma seleção das que se apresentam como candidatas a descritores.

Abordamos em (3), seção 4.1 , a regra do SN do modelo padrão da teoria gerativa. Essa regra sofreu algumas reformulações em (12) seção 4.1 , com a proposição da supercategoria ADJ, feita por Lemle (1984).

No sistema LINGA as categorias que ocupam as posições ADJ estão codificadas no FDIC, com exceção dos adjetivos pré-nominais que estão assinalados na coluna 33 do NDIC. Esses adjetivos estão marcados no NDIC para evitar que a análise sintática do LINGA, operada pela rotina HTAB, os considere como nomes, o que comprometeria a correção da análise automática. A proposta de Pazini referente à anteposição do adjetivo ao nome trouxe uma importante contribuição para nosso propósito de marcar esse adjetivo em NDIC na coluna 33. Segundo Pazini (1978: 110) os adjetivos prénominais são não-restritivos e marcados com o traço [+ gradação] e, como tais, não se apresentam como necessários ao conteúdo lógico da sentença, pois refletem o fluido, o impreciso da linguagem, o pessoal e o subjetivo, por exemplo:

(18) Compramos boas ações na Bolsa do Rio.
Compramos ações boas na Bolsa do Rio.

Por isso, nós marcamos esses adjetivos na coluna 33 de NDIC e, em decorrência, serão desprezados como possíveis elementos de um descritor composto. Ao contrário, o adjetivo restritivo e marcado com traço [- gradação], segundo Pazini, delimita seu antecedente especificando com precisão uma classe de seres e não qualquer ser nomeado pelo nome; por isso, é imprescindível à compreensão lógica da setença, razão pela qual são candidatos naturais a comporem

com o nome um descritor composto. Na proposição de Contreras (1979), esses adjetivos pós-nominais apresentam a seguinte configuração + [N_____] e + [Cop_____]: Por exemplo:

(19) Compramos ações preferenciais na Bolsa do Rio.

Para a indexação automática as palavras das categorias Quant (quantificadores), Det (determinantes) e ADJ (Poss, Card, Ord e Indef) que são elementos pré-nominais do SN do modelo padrão não são considerados, pois são palavras instrumentais de escasso conteúdo nacional. Observe-se que esses elementos na teoria $\bar{X}$ estão marcados a nível de $\bar{\bar{X}}$, isto é, são especificadores de $\bar{X}$ ($\bar{N}$). Segundo Stockwell (1980: 94) essas categorias contêm informações semânticas e referenciais do nome. Integram, segundo Bidermann (1978: 251), a estrutura formal da língua, fazendo funcionar o sistema lingüístico, ao contrário das classes de palavras de significação externa (como o nome, verbo e adjetivo) que constituem os lexemas de conteúdo, formam os sistemas abertos do código lingüístico, têm referência extra-lingüística e são os candidatos naturais a descritores de um documento analisado.

Vimos na descrição do sistema LINGA, em 3.6.3, que as palavras gramaticais da língua estão codificadas na rotina FDIC, que para o processo de indexação automática funciona como um anti-dicionário, que elimina as palavras nele contidas. Além da rotina FDIC, a rotina NDIC também executa esse processo de eliminação de palavras do processo

da indexação automática. No caso do NDIC, as palavras de escasso conteúdo nocional estão assinalados na coluna 33.

Além das palavras excluídas por FDIC e NDIC do processo da indexação automática, o sistema LINGA também exclui desse processo as palavras definidas como verbos pelas rotinas VTAB e VDIC que, nesse caso, também funcionam como anti-dicionários. Como percebemos, esse processo de indexação automática é altamente seletivo. Dessa triagem operada pelas rotinas citadas restam somente os nomes e os adjetivos com a marcação +[N_____] e +[Cop____]. Estas duas categorias passam a ser consideradas pelo sistema LINGA como os candidatos a descritores de um documento analisado. Da referida triagem resulta uma lista de palavras (V. quadro 3.19) representativa do conteúdo do documento.

Conforme vimos em 4.3, consideramos como candidatos naturais a descritores a categoria básica do nome (N) e os adjetivos que ocorrem exclusivamente à sua direita. A projeção dos adjetivos em $\bar{N}$ não poderá ser considerada para definir os elementos de um descritor composto dado que os adjetivos que ocorrem exclusivamente à direita podem também se projetar em $\bar{\bar{N}}$ como vimos em (19). Os resultados definitivos da análise das homografias aparecem no quadro 3.18. Nessa etapa, todas as configurações dos itens lexicais já terão sido analisados. Dessa eta

pa, conforme podemos constatar no fluxograma do LINGA (V. quadro 3.3) Serão extraídos os descritores simples. Os descritores compostos também são extraídos da etapa da análise das homografias e terão a estrutura lingüística preferencial de NOM ATT. Portanto, sempre que suceder em seqüências com aquela estrutura, estas serão armazenadas em um banco específico de dados, como candidatos preferenciais a descritores de um documento analisado pelo LINGA.

5.1 A indexação automática no LINGA

A indexação automática, segundo Robredo (1978), é uma operação que identifica, através de programas de computador, termos ou expressões significativas dos documentos para descrever de forma condensada o seu conteúdo.

O processo de indexação automática baseia-se, segundo Robredo (1982) "na comparação de cada palavra do texto com uma relação de palavras vazias de significado, previamente estabelecidas, que conduz, por eliminação, a considerar as palavras restantes do texto como palavras significativas". Este processo pode identificar termos, pares de termos ou até frases significativas que expressam o conteúdo do documento e, pode-se dizer, que é semelhante ao processo de leitura-memorização. No sistema LINGA, conforme já vimos, estão arroladas na rotina FDIC e estão assinaladas na coluna 33 da rotina NDIC.

Não é objeto específico desta dissertação demonstrar todos os complicados métodos de freqüência ou análise estatística da ocorrência dos termos significativos e candidatos a descritores de um texto. Essa tarefa cabe especificamente ao âmbito da ciência da biblioteconomia. Sugerimos ao leitor conferir a tese de Simone Bastos, integrante da equipe do LINGA, sobre o tema: A análise comparativa entre a indexação automática e manual na literatura brasileira de ciência da informação. No entanto, cabe-nos estabelecer, como princípio para os procedimentos estatísticos, que os descritores mais adequados para a indexação (cf. Salton e Yang, 1973), são os que possuem freqüência média num documento. Os termos de alta e baixa freqüência são, respectivamente, de ocorrência rara e geral e possuem um baixo poder de discriminação.

Robredo (1982) estabeleceu três tipos de descritores que perfeitamente poderiam ser detectados a partir de procedimentos estatísticos inseridos no LINGA: 1) descritores de escopo - são termos de alta freqüência e baixa especialidade que caracterizam áreas de conhecimento; 2) descritores de facetas - são termos de média freqüência e especialidade que caracterizam sub-áreas de interesse e 3) descritores pontuais - são termos de baixa freqüência e alta especialidade, caracterizando um número limitado de documentos.

5.2 Contribuições do sistema LINGA ao processo de indexação automática

Essencialmente a contribuição do sistema LINGA ao processo de indexação automática é de ordem teórica e prática, na medida em que oferece análise lingüística, nos níveis morfológico e sintático das frases analisadas. A partir dessa análise, a aplicação prática do LINGA para o processo da indexação automática será mais completa, com resultados mais objetivos.

Poderíamos citar algumas contribuições:

- os candidatos a descritores serão resultado de uma análise linguística automática sem interferência de indexadores humanos;
- a linguagem de indexação poderá ser pré-coordenada com a análise da estrutura do SN que é denominada por $\bar{N}$, nas configurações de N previamente estabelecidas.
- a linguagem de indexação não apresentará estrutura hierárquica, pois não existe no sistema LINGA um mecanismo que previamente atribua hierarquia de valores a possíveis descritores;
- a linguagem será enumerativa, pois descritores aparecerão tantas vezes quantas ocorrerem em um documento, com a identificação da frase em que ocorrem.

CONCLUSÃO

Tivemos como objetivos em nosso trabalho:

- apresentar um resumo do estado da arte na LC em alguns centros da Europa, dos Estados Unidos e, especificamente, do Brasil;

- descrever sistematicamente todos os dicionários, tabelas, rotinas e procedimentos do sistema de análise automática da língua portuguesa, denominado LINGA;

- demonstrar como os resultados da análise lingüística teórica são apreendidos e aproveitados no desenvolvimento da LC que, num processo circular, contribui, por sua vez, para tornar a análise lingüística mais precisa e formalizada;

- discutir a estrutura do SN português, à luz de várias abordagens da teoria gerativa, enfatizando o fato de que o modelo padrão da regra de expansão do SN é redundante com as informações do léxico, que no LINGA estão codificadas na rotina HTAB. Mostrou-se, assim, ser desnecessária a implementação no sistema LINGA de um algoritmo do SN baseado no modelo padrão.

- demonstrar como as rotinas do LINGA, em especial a HTAB , desenvolvem uma análise consonante com a proposta da teoria $\bar{X}$ a respeito das projeções das categorias lexicais;
- demonstrar que os itens lexicais desenvolvidos no sistema LINGA estão marcados em relação aos demais na rotina HTAB, no nível de seus modificadores e especificadores , obedecendo-se suas configurações e suas possibilidades de coocorrências.

Como percebemos, o sistema LINGA proporciona um arcabouço teórico e tecnológico para a análise automática do sistema lingüístico português nos níveis morfológico e sintático. Isso faculta as mais diversas aplicações, como por exemplo: a tradução automática, a correção automática de erros ortográficos em um texto, a definição de estruturas morfossintáticas mais usadas por um autor em seus textos , contribuindo, assim, para a análise estilística, o ensino de línguas, etc. No caso específico deste trabalho, elegemos a indexação automática como uma das aplicações do sistema. Procuramos estabelecer os requisitos lingüísticos para a definição dos descritores simples e compostos, e compostos, e assumimos para o descritor simples a categoria lexical N (Nome) e para o descritor composto a configuração NOM + _____ATT conforme a terminologia do LINGA. Outras configurações de descritor composto, poderão ser pré-estabelecidas.

Serão considerados como elementos de um descritos composto, apenas os objetivos que ocorrerem exclusivamente à

direito do núcleo do SN, sem atentar para sua projeção em barras conforme estabelece a convenção $\bar{X}$. Isto significa que a convenção $\bar{X}$ não será usado para definir a estrutura do descritor composto. Usaremos, apenas, uma abordagem distribucionalista.

Aparentemente os resultados da análise lingüística automática de textos para a indexação automática são frutos de procedimentos relativamente simples; no entanto, quando consideramos a intrincada estrutura lingüística aliada à complexidade da computação que informam esses procedimentos e possibilitam esses resultados, bem teremos uma visão do que envolve um trabalho nessa área.

Não foi nosso objetivo específico desenvolver um trabalho empírico no processo da indexação automática e na teoria da informação. Limitamo-nos a oferecer, com a conciliação que fizemos entre a análise lingüística do SN e as rotinas analíticas do LINGA, uma contribuição que poderá ser incorporada a técnicas já existentes de indexação automática. Dessas técnicas merece nosso destaque a desenvolvida por Robredo (1983), denominada BIB/DIÁLOGO; que através de anti-dicionário de palavras vazias e de dicionário de raízes de palavras, alcança resultados altamente satisfatório no processo de indexação. O ideal seria que houvesse uma conjugação de esforços nessa área, juntando num sistema único os procedimentos de análise lingüística automática do sistema LINGA com os procedimentos desenvolvidos no sistema BIB/DIÁLOGO. E isso não deveria ser um problema maior, de vez que os dois sistemas estão incorporados ao Centro de Processamento de Dados da Universidade de Brasília.

GLOSSÁRIO

Algoritmo - Processo de resolução de um grupo de problemas semelhantes, em que se estipulam, com generalidade e sem restrições, regras formais para a obtenção do resultado, ou da solução do problema.

"es una serie finita de prescripciones potencialmente ejecutables expresadas en un lenguaje definido, que establece cómo ejecutar un cierto encadenamiento de operaciones para resolver todos los problemas de un tipo dado" (cf. Charles Corge apud Rosa s/d).

Análise automática - Procedimentos para definir a estrutura morfossintática de um sistema linguístico através de tabelas, rotinas e algoritmos para que a mesma possa ser submetida a processos de computação.

Argumento - Um SN que ocupa uma posição básica; denota por si mesmo em virtude de sua função referencial (V. Lobato (a sair)).

Biblioteconomia - Conjunto de conhecimentos referentes à organização e administração das bibliotecas.

Bit / bytes - É o número de alternativas binárias (do tipo sim/não) que permitem a escolha de um elemento entre 'n' elementos diferentes (V. Katz, 1975).

Busca superficial - Processo de recuperação de informações a partir de interrogações formuladas em língua natural. Essas interrogações são submetidas à análise automática para a definição das palavras-chave (descritores), após o que é possível efetuar a 'busca de precisão' que consiste em oferecer partes selecionadas de textos que tratam do problema definido na pergunta do usuário de um sistema de informação (cf. Fisher, 1981).

Caixa preta - (black box) Todo sistema cujo comportamento é desconhecido (V. Katz, 1975).

Categorias lexicais - São as categorias que englobam os morfemas lexicais, nocionais ou de significação externa que constituem o conjunto inacabado, aberto, do sistema lingüístico. São os nomes, adjetivos e verbos (cf. Guéron (1982); Pottier, (1975), (1978).

Categorias não-lexicais - São as categorias que agrupam os morfemas gramaticais. São os elementos de um conjunto finito, fechado. São os operadores, os determinantes, os complementadores, etc. São as palavras gramaticais, relacionais ou de significação interna (cf. Guéron (1982), Pottier (1975, 1978)).

Descritor - É a unidade lingüística (no caso específico é o substantivo comum) que veicula os semas essenciais do ser-referente. Apenas pelo substantivo, o descritor não pode particularizar. Necessita dos caracterizadores, que são os adjetivos atributivos, pois possibilitam a adequação do significado do substantivo ao ser-referente que o descritor relata ou recria, distinguindo-o por meio de alguma(s) característica(s) (cf. Silveira, 1980: 23). "Elementos del metalenguage con los que se describen contenidos. Sirven para describir lo común y lo diferente en el contenido de palavras. (Abrahan, 1981: 150)". São as palavras-chaves de um documento.

Entropia - É a função que define a quantidade de informação associada a uma dada mensagem (cf. Katz, 1975). "En la teoria de la información, la magnitud de medida desarrollada para el grado de incertidumbre en la transmissión de la información que está ligado a una señal; este grado de incertidumbre depende de la probabilidad de aparición de la señal. Para un número determinado de respuestas posibles se da el máximo de entropía cuando cada una de estas respuestas posibles tiene igual probabilidad de aparición (= la máxima incertidumbre acerca del contenido informativo que es de esperar); la entropía es débil (= la probabilidad de la predicción es mayor) quando una de las respues

tas posibles posde una probabilidade aparición mayor que todas las otras repuestas. La entropía aumenta con la incertidumbre de la respuesta que pueda darse" (cf. Abraham, 1981: 173).

Filtros lingüísticos - É o processo que elimina os elementos gramaticais (V.Categorias não-lexicais) da estrutura fixa da língua natural do processo da indexação automática para a definição dos descritores.

Gramática Universal - expressão usada para se referir ao estado mental inicial, à faculdade de linguagem da espécie humana que, segundo Chomsky (cf. Lobato, a sair), é uma estrutura cognitiva inata, humana e universal, e faz parte da herança genética de cada membro da espécie humana.

Indexação Automática - É uma operação que identifica, através de programas de computador, termos ou expressões significativas dos documentos para descrever de forma condensada o seu conteúdo (cf. Robredo, 1978). A indexação automática representa um documento usando critérios lingüísticos e/ou estatísticos para selecionar palavras ou frases significativas do texto (cf. Dillon e Gray, 1983).

Lematizar - Termo específico da computação que indica a "leitura" feita de dados pelo computador.

Lingüística Computacional (LC) - É o ramo da lingüística que tem por objetivo identificar, formalizar e testar as regras de um sistema lingüístico numa grande massa de dados (cf. Walker, 1981).

Matriz de precedência - É uma matriz que define o contexto sintático da palavra obedecendo à marcação lexical dos possíveis vizinhos, i.e. termos que precedem ou sucedem a categoria em análise. No sistema LINGA esta função é desempenhada pela rotina HTAB.

Teoria da ligação - É uma teoria lingüística cujo objetivo é a identificação do antecedente de um anafórico e de um pronominal, quando há tal antecedente. Essa teoria se resume em três princípios: 1) um anafórico (vestígio de SN e PRO) tem de estar ligado na sua categoria de regência 2) um pronominal (PRO - pronomes) tem de estar livre na sua categoria regente; 3) uma expressão R (variáveis ligadas por um operador e nomes) tem de estar livre (cf. Lobato, a sair).

Teoria da regência - Trata da relação entre o núcleo de uma construção e as categorias dependentes. É uma teoria, segundo Lobato (a sair), cujo objetivo é a formalização da noção de complementação. Essa teoria tem de determinar: 1) que elementos podem reger; 2) que elementos podem ser regidos e em

que condições; 3) qual a condição estrutural para se dar a regência.

Usuário - Cada um daqueles que usa ou desfruta alguma coisa coletiva, ligada a um serviço público ou particular. No caso específico de um sistema de informação, usuário vem a ser aquele que se serve de um serviço automatizado de informações, de uma biblioteca, por exemplo.

REFERÊNCIAS BIBLIOGRÁFICAS

ABRAHAM, Werner. Diccionário de Terminología Lingüística Actual. Madrid, Gredos, 1981.

ANDREEWSKY, A. & RUAS, Vitoriano. Indexação automática baseada em métodos lingüísticos e estatísticos e sua aplicabilidade à língua portuguesa. Rio de Janeiro, PUC/DI, 1982. 32p.

_____. The SPIRIT systems natural language retrieval of textual information. (manuscrito) maio, 1983.

APRESJAN, Ju. D. Idéias e Métodos da Lingüística Estrutural Contemporânea. São Paulo, Cultrix, 1980.

BAYLON, Christian & FABRE, Paul. Iniciação à Lingüística. Coimbra, Almedina, 1979.

BASTOS, Alex C. Programa COBOL. 3ª Ed. Rio de Janeiro, Livros Técnicos e Científicos, 1981.

BELY, N. et alii. Procedures d'analyse sémantique appliquées à la documentation scientifique. Paris, CNRS, 1970.

BIDERMAN, M.T. Lingüística Computacional: um desafio de trinta anos. Dados e Idéias. Rio de Janeiro, 2 (6): 29-39, jun./jul. 1977.

_____. Teoria Lingüística: lingüística quantitativa e computacional. Rio de Janeiro, Livros Técnicos e Científicos, 1978.

BOTT, M.F. Lingüística Computacional. in: Lyons John. Novos horizontes em Lingüística. São Paulo, Cultrix/ EDUSP, 1976. p. 206-218.

CABRAL, Leonor Scliar. Introdução à Lingüística. Porto Alegre, Globo, 1982.

CÂMARA Jr., Joaquim Mattoso. Dispersos. 2ª ed. Rio de Janeiro, Editora da Fundação Getúlio Vargas, 1975, pp. 3-14.

_____. Estruturas da Língua Portuguesa. 2ª ed. Petrópolis, Vozes, 1970.

_____. Problemas de Lingüística Descritiva. 10ª ed. Petrópolis, Vozes, 1981.

CARDOSO, Anibal P. Programação Estruturada em COBOL. Rio de Janeiro, Livros Técnicos e Científicos, 1981.

CESARINO, M.A. da N. & PINTO, M.C.M.F. Análise de Assuntos. Revista de Biblioteconomia de Brasília, 8 (1):32-45, Jan./Jun. 1980.

CHOMSKY, N. Aspectos da Teoria da Sintaxe. Coimbra, Armênio Amado, 1975.

_____. Remarks on nominalization - IN: JACOBS, R.A. & ROSENBAUM, P.S., (orgs.) Readings in English Transformational grammar. Waltham, Mass., Ginn, 1970. 184-221.

_____. Lectures on government and binding. The Pisa lectures. 2ª ed. rev. Dordrecht, Paris, 1981/1982b.

CONTRERAS, N. The case for base-generated atribute. Georgetown University Press, 1979.

DILLON, Martin & GRAY, Ann S. FASIT: A Fully automatic Syntactically Based Indexing System. Journal of the American Society for Information Science. 34(2): 99 - 108, Mar., 1983.

FISCHER, H.G. CONDOR. An integrated Data-base information retrieval system for structured and unstructured data. Siemens Forsch und entwickle-Berichte. Berlin. 10 (3): 179-87, 1981.

FREITAS, Horácio Rolim de. Princípios de Morfologia. 2ª ed. Rio de Janeiro, Presença, 1981.

HALLER, Johann. Análise automática de textos em sistemas de informação. Revista de Biblioteconomia de Brasília. Brasília 11 (1): 105-113, jan/jun., 1983.

_____. Processamento de textos em linguagem natural. Congresso Nacional de Informática, 15. Rio de Janeiro. Anais... Rio de Janeiro, out. 1982.

HUTCHINS, W.J. Languages of indexing and classification: a linguistic study of structures and functions. Stevenage, Peter Perenigrus, 1975. 145p.

JACKENDOFF, R. $\bar{X}$ - syntax: a study of phrase structure. Cambridge, Massa The MIT Press 1977.

JAKOBSON, Roman. Lingüística e comunicação. São Paulo, Cultrix, 1970.

KATZ, Chaim Samuel, DORIA, Francisco Antônio, LIMA, Luiz Costa. Dicionário Básico de Comunicação. 2ª ed. Rio de Janeiro, Paz e Terra, 1975.

LEMLE, M. A ordem dos adjetivos no sintagma nominal em inglês e português: implicações para a teoria gramatical. Encontro Nacional de Lingüística. Rio de Janeiro. Anais ... Rio de Janeiro, PUC, 1978.

_____. Análise Sintática. (Teoria Geral e Descrição do Português). São Paulo, Ática, 1984.

LOBATO, L.M.P. A estrutura do SN em português: derivação de seqüências nominais com adjetivos. Encontro Nacional de Lingüística, 6. Rio de Janeiro. Anais... Rio de Janeiro, 1979.

LOBATO, L.M.P. Sintaxe formal do português: da teoria padrão à teoria da regência e ligação. Belo Horizonte, Vigília, (a sair).

LONGO, B.N. de Oliveira. A estrutura do SN português. Brasília, UnB, 1981. (inédito).

LYONS, John. Lingua (gem) e lingüística - uma introdução. Rio de Janeiro, Zahar, 1982.

MACEDO, Walmirio. Elementos para uma estrutura da Língua Portuguesa. Rio de Janeiro, Presença. 1976.

MALMBERG, Bertil. As Novas Tendências da Lingüística. São Paulo, Nacional, 1974.

MARTINET, André. Elementos de Lingüística General. Madrid, Editorial Gredos, 1968. p. 148-9.

OLIVEIRA, Carlos Alberto de. Processamento de texto em Linguagem natural. (Projeto de Pesquisa LINGA II) manuscrito) jan/84.

_____. Programa de identificação automática de forma-base de substantivos portugueses simples. Encontro Nacional de Lingüística, 7. Rio de Janeiro, nov. 1982. Anais... . Rio de Janeiro, PUC, 1982.

PAIS, Cidmar Teodoro. Introdução à Fonologia. São Paulo, Global, 1981.

PAZINI, M.C.B. O Adjetivo - Um problema sintático. Florianópolis, Universidade Federal de Santa Catarina, 1978.

_____. A posição do adjetivo na locução nominal em português. Encontro Nacional de Lingüística, 3. Rio de Janeiro, out. 1977. Anais ... Rio de Janeiro, PUC, 1977.

PONTES, Eunice. Os determinantes em português. Tempo Brasileiro. 53/54 : 145-65, 1978.

POTTIER, Bernard et alii. Estruturas Lingüísticas do Português. 3ª ed. São Paulo, DIFEL, 1975.

_____. Lingüística Geral: Teoria e descrição . Rio de Janeiro, Presença, 1978.

GUÉRON, J. Les operateurs: Contribution a une théorie de traits syntaxiques. IN: GUÉRON, J. et alii, org. Grammaire Transformationelle: theorie et méthodologies. Paris, CNRS, 1982.

RECTOR, Mônica. A Linguagem da Informática. Perspectiva Universitária: informação sobre educação, cultura e esportes. Rio de Janeiro, 11 (185): 1 e 4, ago. 84.

ROBREDO, Jaime. Documentação de hoje e de amanhã. Brasília, ABDF, 1978.

ROBREDO, Jaime. A indexação automática de textos: o presente já entrou no futuro. Estudos avançados em Biblioteconomia e Ciência da Informação. Brasília, 1: 235-74, 1982.

_____. Otimização dos processos de indexação dos documentos e recuperação da informação mediante uso de instrumentos de controle terminológico. Ciência da Informação. Rio de Janeiro, 11 (1):3-18, 1982.

ROSA, Nicolás. Léxico de Lingüística. Buenos Aires, Centro e América Latina, s.d.

SALTON, G & YANG, C.S. On the specification of term values in automatic indexing. Journal of Documentation, 29 (4):351-72, dec. 1973.

STAUB, A. Herman Paul, F. de Saussure e K. Buhler na Lingüística moderna. Brasília, Thesaurus, 1981.

_____. As três gramáticas: uma unidade na trindade. in: Temas de lingüística aplicada. Brasília, Thesaurus, 1981. p. 11-31.

SCHER, Nelmo Roque. Projeto de correção semi-automática de erros de ortografia. 1983 (inédito).

SILVEIRA, Maria Cecília P. A gramática Portuguesa na pesquisa e no ensino nº 2. São Paulo, Cortez Editora, 1980.

VICINI, Alcides. Linguagem e contexto. Santa Rosa, Barcellos, 1981.

WALKER, Donald E. The organization and use of information: Contributions of Information Science, Computational Linguistics and Artificial Intelligence. Journal of the American Society for Information Science. 348-63. September 1981.

ZIMMERMANN H.E.A. JUDO: Juristische Dokumentanalyse Bericht 1977-1979. Regensburg, Universitaet Regensburg, 1980.

ANEXOS

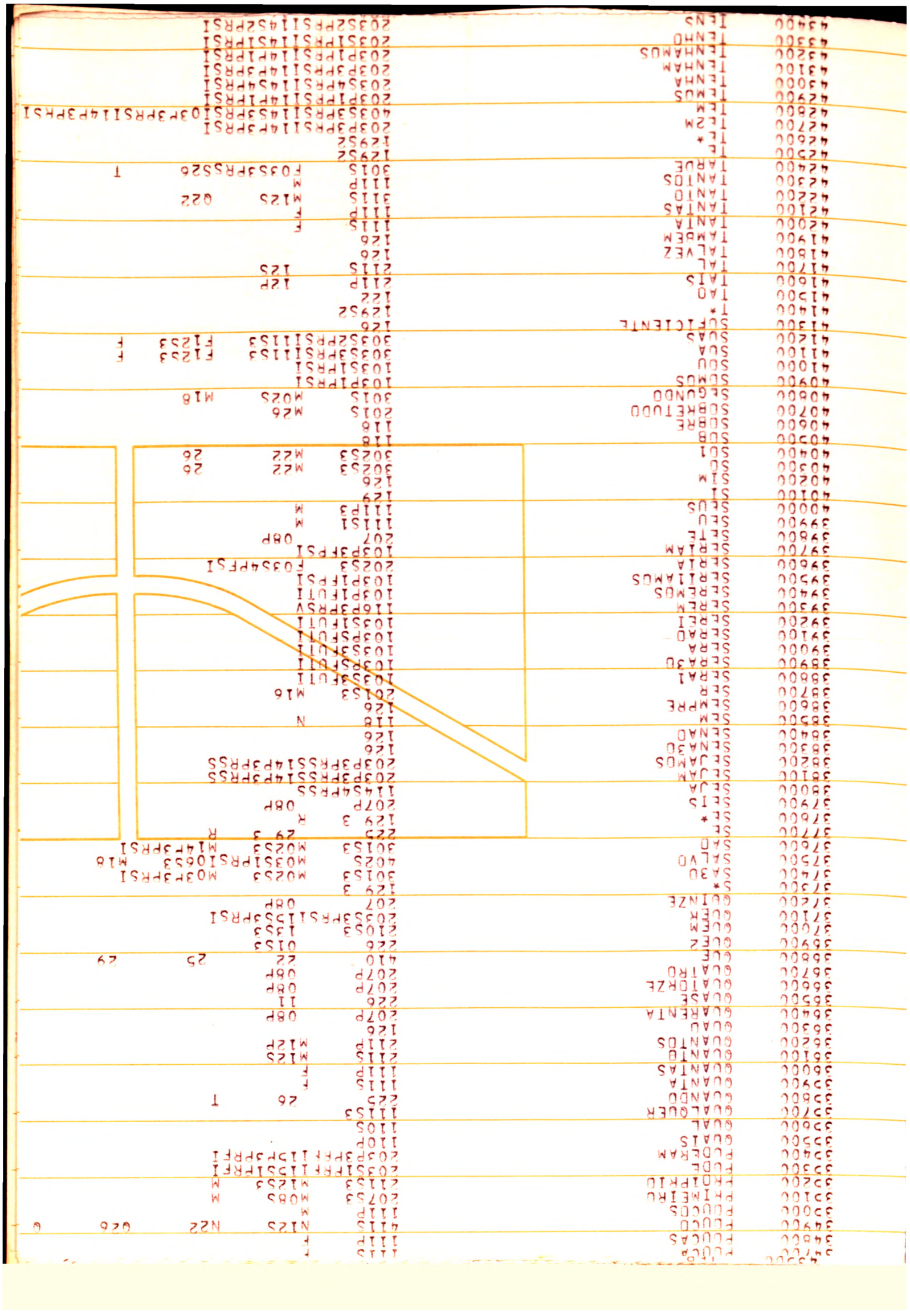
WORKFILE: PKR/FDIC (06/09/83)

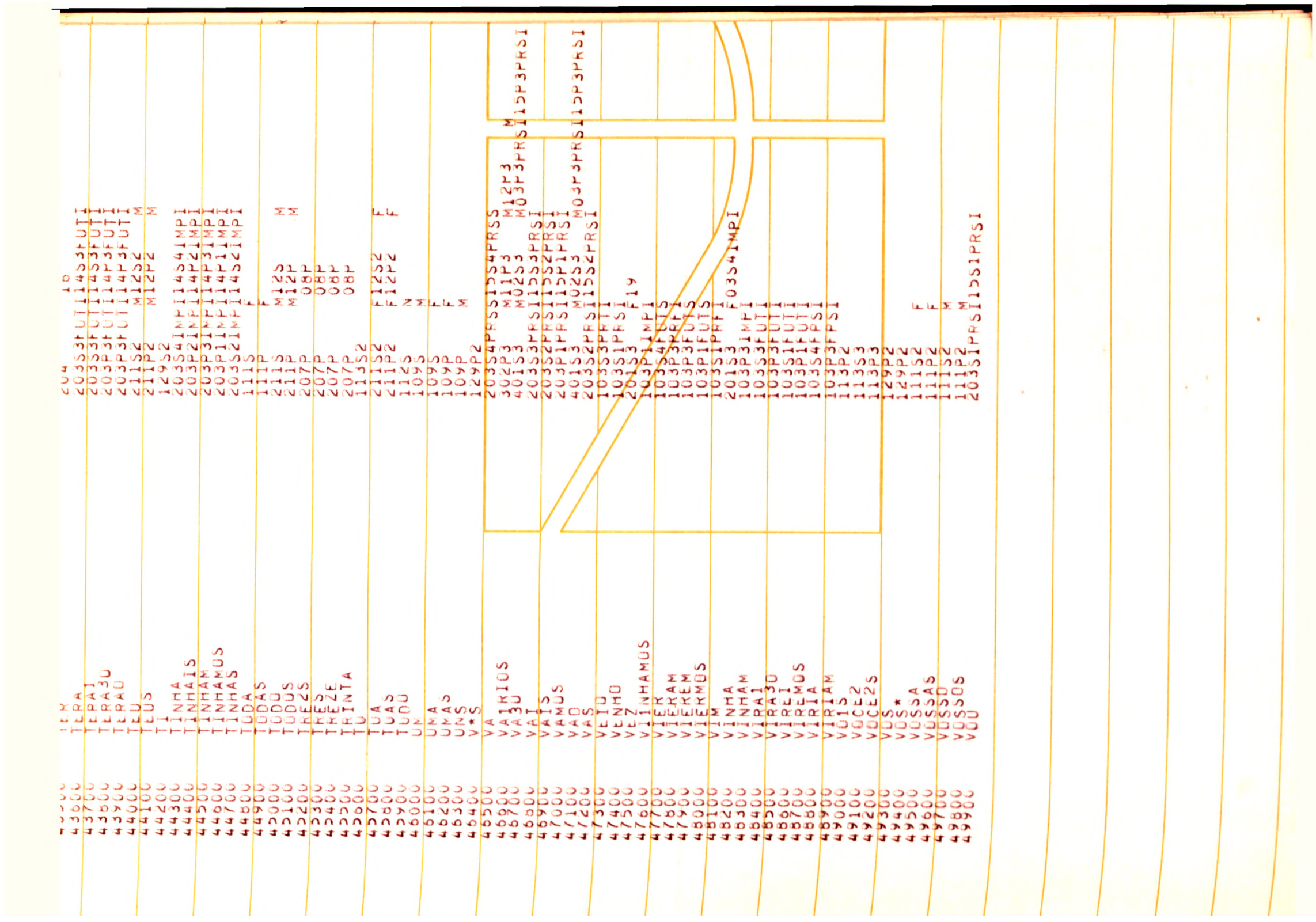
		F	T	M		T	M		T	M
100	*	A		I	/		I	/		I	/
200	*										
300	*FWORD	C	CNP	M	G	CNP	M	G	CNP	M	G
400	A	309S			F18			2953			F
500	A*	1295	3		F						
600	A4	1185	3		F						
700	A4QUELA	1185	3		S						
800	A4QUELAS	118			S						
900	A4QUELE	118			S						
1000	A4QUELES	118			S						
1100	A4QUILO	118			S						
1200	A4S	118P			F						
1300	ABAIXO	126			L						
1400	ACASO	201S			M26			M			
1500	ACIMA	126			L						
1600	ADIANTE	126			L						
1700	AFIM	206				19					
1800	AFINAL	126			M						
1900	AGORA	126			T						
2000	AI	126			L						
2100	AII	126			L						
2200	AINDA	126			T						
2300	AMBAS	211P3			F12P3			F			
2400	AMBOS	211P3			M12P3			M			
2500	ALEIM	126			M						
2600	ALEM	126			M						
2700	ALI	118			L						
2800	ALGO	211S3			N12S3			N			
2900	ALGUEIM	211S3			N12S3			N			
3000	ALGUEM	211S3			N12S3			N			
3100	ALGUM	211S3			M12S3			M			
3200	ALGUMA	211S3			F12S3			F			
3300	ALGUMAS	211P3			F12P3			F			
3400	ALGUNS	211P3			M12P3			M			
3500	ALIAS	126			M						
3600	ANTE	118									
3700	ANTES	218			T26			T			
3800	AO	118S			M						
3900	AOS	118P			M						
4000	APENAS	126									
4100	APESAR	120									
4200	APOIS	126			T						
4300	APOS	126			T						
4400	AQUELA	211S3			F12S3			F			
4500	AQUELAS	211P3			F12P3			F			
4600	AQUELE	211S3			M12S3			M			
4700	AQUELES	211P3			M12P3			M			
4800	AQUI	126			L						
4900	AQUILO	112S3			N						
5000	AS	309P			F29P			F18			
5100	AS*	129P			F						
5200	ASSIM	126			M						
5300	ATE	218			26			M			
5400	ATE1	218			26			M			
5500	ATRAIS	126			L						
5600	ATRAS	126			L						
5700	ATRAVEIS	118									
5800	ATRAVES	118									
5900	BEM	226			M01S			M			
6000	BREVE	202S3			N26			T			
6100	BREVES	102S3			N						
6200	CA	126			D						
6300	CA1	126			D						
6400	CADA	111S3			N						
6500	CASO	301S			M25			03S1PRSI			
6600	CEDO	203S1PRSI			26			T			
6700	CEM	207P			08P						
6800	CERCA	301S3			19			L26			L
6900	CINCO	207P			08P						
7000	CINQUENTA	207P			08P						
7100	COM	118									
7200	COMIGO	129S1									
7300	COMO	303S1PRSI			24			25			
7400	CONFORME	218			24						
7500	CONOSCO	129P1									
7600	CONSIGO	203S1PRSI			29S3						
7700	CONSOANTE	301S3			F18			25			
7800	CONTANTO	125									
7900	CONTIGO	129S2									
8000	CONTRA	118									
8100	CONTUDO	124									
8200	CONVOSCO	129P2									

```
```

17100	ESTEJAM	203P3PRSS14P3PRSS			
17200	ESTES	211P3	M12P3	M	
17300	ESTOU	203S1PRSI14S3PRSI			
17400	ETC	126			
17500	EU	113S1			
17600	FEITO	305	06	01S3	M
17700	FICAR	204	14		
17800	FO2SSEMOS	203P3IMPS14P3IMPS			
17900	FOI	203S3PRFI14S3PRFI			
18000	FOMOS	203P1PRFI14P1PRFI			
18100	FOR	303S1FUTS03S3FUTS14S1FUTS			
18200	FORA	303S1PQFI26		01S3	M
18300	FORAM	203P3PRFI14P3PRFI			
18400	FOREM	203P3FUTS14P3FUTS			
18500	FOSSE	403S1IMPS03S3IMPS14S1IMPS14S3IMPS			
18600	FOSSEM	203P3IMPS14P3IMPS			
18700	FUI	203S1PRFI14S1PRFI			
18800	HA	203S3PRSI24S3PRSI			
18900	HA1	203S3PRSI24S3PRSI			
19000	HA3O	203P3PRSI14P3PRSI			
19100	HAO	203P3PRSI14P3PRSI			
19200	HAVEMOS	203P1PRSI14P1PRSI			
19300	HAVENDO	123			
19400	HAVER	301	04	16	
19500	HAVERI1AMOS	303P1FPSI14P1FPSI			
19600	HAVERIA	203S4FPSI14S4FPSI			
19700	HAVERIAM	203P3FPSI14P3FPSI			
19800	HAVI1AMOS	203P1IMPI14P1IMPI			
19900	HAVIA	203S4IMPI14S4IMPI			
20000	HAVIAM	203P3IMPI14P3IMPI			
20100	HAVIDO	105			
20200	HEI	203S1PRSI14S1PRSI			
20300	HOJE	126	T		
20400	HOUVE	203S4PRFI14S4PRFI			
20500	HOUVEISSEMOS	203P3IMPS14P3IMPS			
20600	HOUVER	203S4FUTS14S4FUTS			
20700	HOUVERA	203S4PQPS14S4PQPS			
20800	HOUVERAM	203P3PQPS14P3PQPS			
20900	HOUVERMOS	203P1FUTS14P1FUTS			
21000	HOUVESSE	203S4IMPS14S4IMPS			
21100	HOUVESSEM	203P3IMPS14P3IMPS			
21200	HS	129P3	F		
21300	IDO	205	06		
21400	INCLUSIVE	218	26		
21500	INDO	123			
21600	IR	204	14		
21700	ISSO	112S3	N		
21800	ISTO	112S3	N		
21900	JA1	126	T		
22000	JA	126	T		
22100	JAH	126	T		
22200	JAMAIS	126	T		
22300	JUNTO	303S1PRSI06S3		M26	
22400	KS	129P3	M		
22500	LA	201S3	M26	L	
22600	LA1	201S3	M26	L	
22700	LEI	101S3	F		
22800	LH*	129S3	N		
22900	LH*S	129S3	N		
23000	LHE	129S3	N		
23100	LHE*	129S3	N		
23200	LHES	129P3	N		
23300	LHES*	129P3	N		
23400	LOGO	224	26	T	
23500	LONGE	126	L		
23600	M*	129S1	M		
23700	MAL	201S3	M26	M	
23800	MAS	301P3	M02P	M24	
23900	MAIS	122			
24000	ME	129S1			
24100	ME*	129S1			
24200	MEDIANTE	118			
24300	MEIO	301S3	M11S3	M26	
24400	MELHOR	211S3	N12S3	N	
24500	MENOS	122			
24600	MESMA	211S	F12S	F	
24700	MESMAS	211P	F12P	F	
24800	MESMO	311S	M12S	M26	
24900	MESMOS	211P	M12P	M	
25000	MEU	111S1	M12S1	M	
25100	MEUS	111P1	M12P1	M	
25200	MIM	129S1			
25300	MINHA	111S1	F12S1	F	
25400	MINHAS	111P1	F12P1	F	
25500	MUITA	111S	F		
25600	MUITAS	111P	F		
25700	MUITO	311S	M12S	M22	
25800	MUITOS	211P	M12P	M	

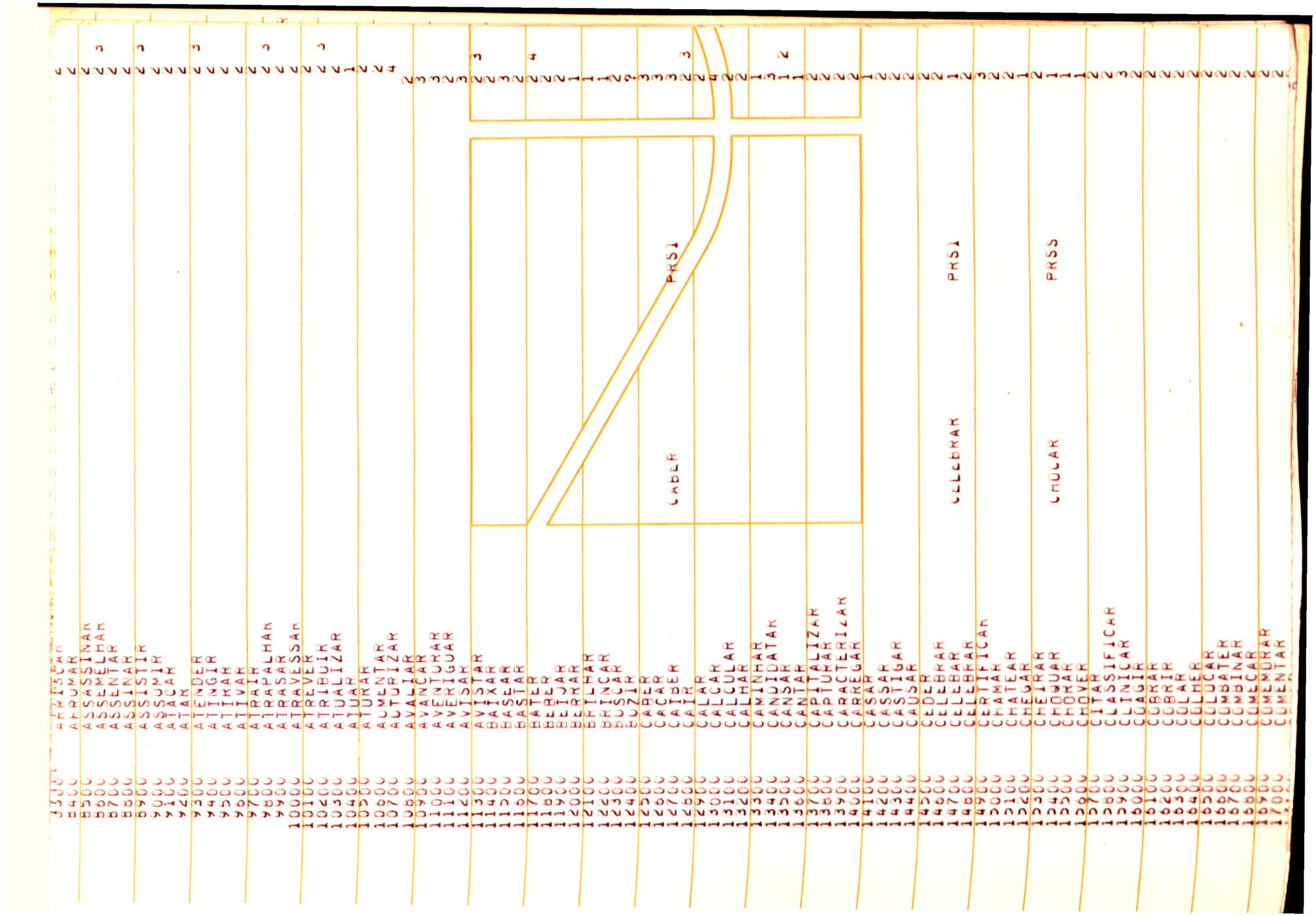
25900	N*S	129P1					
26000	NA	118S	F				
26100	NA3O	126	N				
26200	NAO	126	N				
26300	NADA	203S3PRSI12		N			
26400	NAQUELA	212S3	F18S3	F			
26500	NAQUELAS	212P3	F18P3	F			
26600	NAQUELE	212S3	M18S3	M			
26700	NAQUELES	212P3	M18P3	M			
26800	NAQUILO	212S3	N18S3	N			
26900	NAS	118P	F				
27000	NELA	129S	F				
27100	NELE	129S	M				
27200	NEM	124	N				
27300	NENHUM	212S	M18S	M			
27400	NENHUMA	212S	F18S	F			
27500	NESSA	212S3	F18S3	F			
27600	NESSAS	212P3	F18S3	F			
27700	NESSE	212S3	M18S3	M			
27800	NESSES	212P3	M18P3	M			
27900	NESTA	212S3	F18S3	F			
28000	NESTAS	212P3	F18P3	F			
28100	NESTE	212S3	M18S3	M			
28200	NESTES	212P3	M18P3	M			
28300	NISSO	211S3	N12S3	N			
28400	NISTO	211S3	N12S3	N			
28500	NINGUEM	112	N				
28600	NO	201S3	M18S	M			
28700	NO1S	201P3	M13P1				
28800	NOS	401P	M13P1	18P1	M29P1	M	
28900	NOS*	129P1	M				
29000	NOSSA	111S1	F				
29100	NOSSAS	111P1	F				
29200	NOSSO	111S1	M				
29300	NOSSOS	111P1	M				
29400	NOVE	207	08P				
29500	NUM	118S	M				
29600	NUMA	118S	F				
29700	NUMAS	118P	F				
29800	NUNCA	126	N				
29900	NUNS	118P	M				
30000	O	209S	M29S	M			
30100	O*	129S	M				
30200	OITO	207	08P				
30300	ONDE	210	L26	N			
30400	ONTEM	126	T				
30500	ONZE	207	08P				
30600	ORA	224	26				
30700	OS	209P	M29P	M			
30800	OS*	129P	M				
30900	OU	124					
31000	OUTRA	211S	F12S	F			
31100	OUTRAS	211P	F12P	F			
31200	OUTRO	211S	M12S	M			
31300	OUTROS	211P	M12P	M			
31400	PAR	201S	20				
31500	PARA	301S	03S3PRSI18				
31600	PELA	118S	F				
31700	PELAS	118P	F				
31800	PELO	218S	01S3	M			
31900	PELOS	218P	01P3	M			
32000	PERANTE	118					
32100	PERTO	126	L				
32200	PO2R	2013S	M04				
32300	PODE	203S4	I15S4	I			
32400	PODEM	203P3PRSI15P3PRSI					
32500	PODER	301S	M04	17			
32600	PODERA	203S3FUTI15S3FUTI					
32700	PODERAO	203P3FUTI15P3FUTI					
32800	PODERIA	203S4FPSI15S4FPSI					
32900	PODERIAM	203P3FPSI150PS1FPSI					
33000	PODIA	203S4IMPI15S4IMPI					
33100	PODIAM	203P3IMPI15P3IMPI					
33200	POIS	124					
33300	PONDO	123					
33400	POR	3013S	M04	18			
33500	PORA	103S3FUTI					
33600	PORA1	103S3FUTI					
33700	PORA3O	201S3	M03P3FUTI				
33800	PORAO	201S3	M03P3FUTI				
33900	POREIM	124					
34000	POREM	124					
34100	PORQUANTO	125					
34200	PORQUE	225	26				
34300	PORQUE2	226	01S3				
34400	PORTANTO	124					
34500	POSSO	203S1PFSI15S1PRSI					
34600	POSTO	201S3	M05				

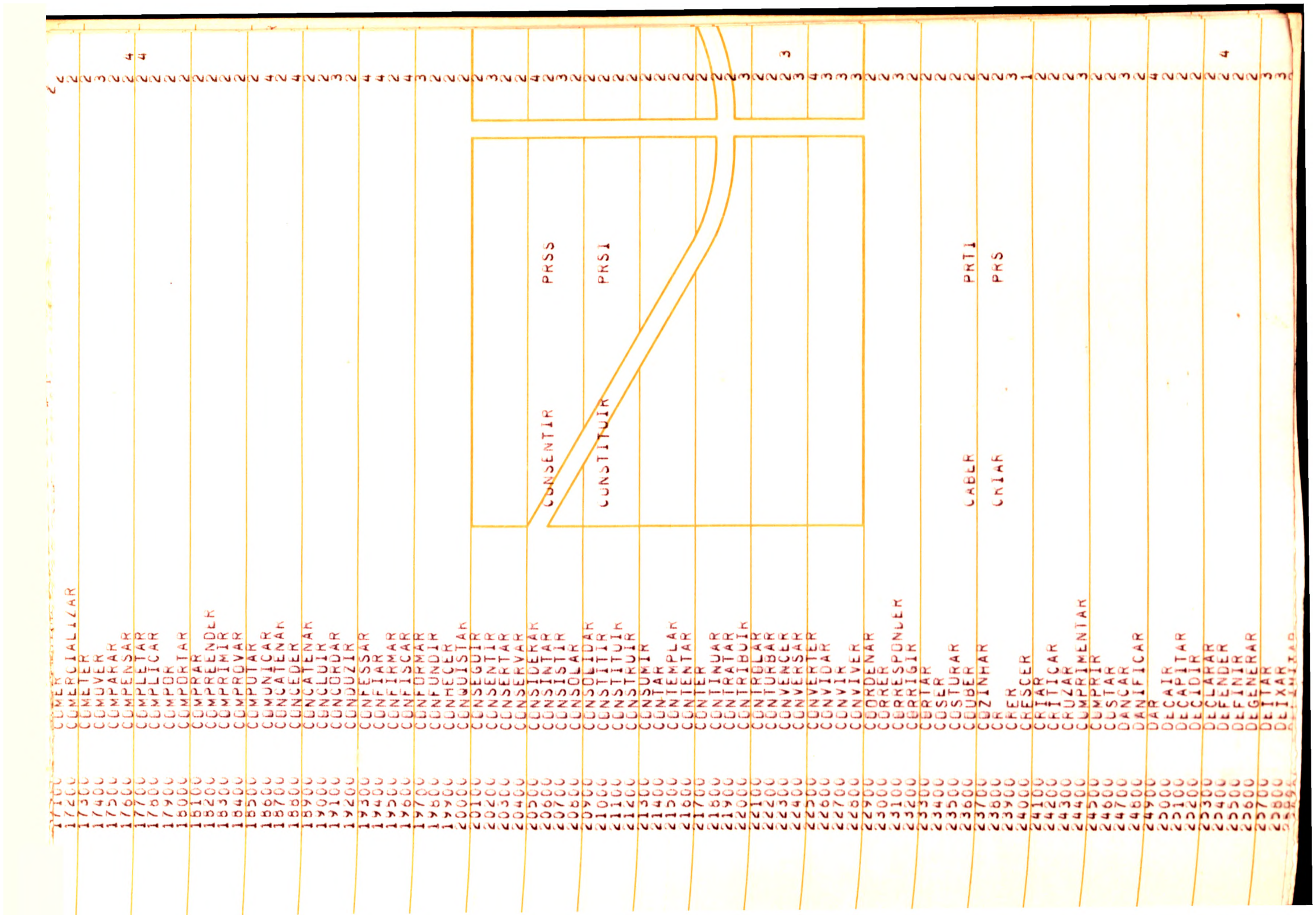




WORKFILE: PKR/VDIC (04/02/83)

100	INFINITIV			
200	ABANDONAR			2
300	ABORDAR			2
400	ABORRECER			2
500	ABRANGER			2
600	ABRIR			2
700	ABUSAR			2
800	ACABAR			2
900	ACALMAR			2
1000	ACENAR			2
1100	ACENDER			2
1200	ACERTAR			2
1300	ACHAR			2
1400	ACOMPANHAR			2
1500	ACONSELHAR			2 3
1600	ACONTECER			1
1700	ACORDAR			2 3
1800	ACREDITAR			3
1900	ACRESCENTAR			2 3
2000	ACUSAR			2 3
2100	ADERIR			3
2200	ADIANTAR			2
2300	ADIAR			2
2400	ADIANTAR			2
2500	ADIVINHAR			2
2600	ADMIRAR			2
2700	ADMITIR			2
2800	ADOECER			1
2900	ADORAR			2 3
3000	ADORMECER			2
3100	ADOTAR			2
3200	ADVIR			2
3300	AFASTAR			2
3400	AFETAR			2
3500	AFIRMAR			2
3600	AFIXAR			2
3700	AFLIGIR			2
3800	AGARRAR			2
3900	AGITAR			2
4000	AGRADAR			3
4100	AGRADECER			3
4200	AGRESCENTAR			2 3
4300	AGUARDAR			2 3
4400	AGUENTAR			2
4500	AJUDAR			2 3
4600	ALEGAR			2
4700	ALIMENTAR			2
4800	ALMOCAR			1
4900	ALTER	ALTERAR	PRSI	2
5000	ALTERAR			2
5100	ALUGAR			2
5200	AMAR			2 3
5300	AMEACAR			2 3
5400	AMPLIAR			2
5500	ANALISAR			2
5600	ANDAR			1 2
5700	ANEXAR			2 4
5800	ANIMAR			2
5900	ANOTAR			2
6000	ANTECIPAR			2
6100	ANUNCIAR			2
6200	APAGAR			2
6300	APANHAR			2
6400	APARECER			4
6500	APERTAR			2
6600	APOIAR			2 3
6700	APONTAR			2
6800	APOSENTAR			2
6900	APRENDER			2 3
7000	APRESENTAR			4
7100	APRESSAR			2
7200	APROFUNDAR			2
7300	APROVAR			2
7400	APROVEITAR			2
7500	AQUECER			2
7600	ARDER			2
7700	ARMAZENAR			2
7800	ARRANCAR			2
7900	ARRANJAR			2
8000	ARRASTAR			2
8100	ASPIRAR			2 3
8200	ARREFECER			2


PRS1
CABER
PRS1
CELEBRAR
PRSS
CHOCAR


PRSS
PRSI
CONSENTIR
CONSTITUIR
PRTI
PRS
CABER
CRIAR

26000	DELINEAR	
26100	DEMORAR	
26200	DENOMINAR	
26300	DEPENDER	
26400	DEPOR	
26500	DEPOSITAR	
26600	DEPREENDER	
26700	DERIVAR	
26800	DERRAMAR	
26900	DESAPARECER	
27000	DESAPERTAR	
27100	DESATAR	
27200	DESCANSAR	
27300	DESCARREGAR	
27400	DESCER	
27500	DESCOBRIR	
27600	DESCULPAR	
27700	DESEJAR	
27800	DESEMBRULHAR	
27900	DESEMPENHAR	
28000	DESENVOLVER	
28100	DESFAZER	
28200	DESISTIR	
28300	DESLIGAR	
28400	DESMENTIR	
28500	DESORIENTAR	
28600	DESPACHAR	
28700	DESPEDIR	
28800	DESPIR	
28900	DESTINAR	
29000	DESTRUIR	
29100	DESVALORIZAR	
29200	DETER	
29300	DETERMINAR	
29400	DEVOLVER	
29500	DICIONARIZAR	
29600	DIFUNDIR	
29700	DIR	DIZER FUTI
29800	DIRIGIR	
29900	DIMINUIR	
30000	DISCRIMINAR	
30100	DISCUTIR	
30200	DISPENSAR	
30300	DISPOR	
30400	DISPUTAR	
30500	DISSECAR	
30600	DISTINGUIR	
30700	DISSOLVER	
30800	DISTRAIR	
30900	DISTRIBUIR	
31000	DIVERSIFICAR	
31100	DIVERTIR	
31200	DIVIDIR	
31300	DIVULGAR	
31400	DIZER	
31500	DOBRAR	
31600	DOER	
31700	DORMIR	
31800	DRAMATIZAR	
31900	DURAR	
32000	DUVIDAR	
32100	ELABORAR	
32200	ELEGER	
32300	ELEVAR	
32400	ELIMINAR	
32500	EDUCAR	
32600	EMBRULHAR	
32700	EMPREGAR	
32800	EMPURRAR	
32900	ENCAMINHAR	
33000	ENCARECER	
33100	ENCARREGAR	
33200	ENCERRAR	
33300	ENCHER	
33400	ENCHUGAR	
33500	ENCOMENDAR	
33600	ENCONTRAR	
33700	ENCOSTAR	
33800	ENFEITAR	
33900	ENFIAR	
34000	ENGANAR	
34100	ENGOLIR	
34200	ENGULIR	
34300	ENQUADRAR	
34400	ENSINAR	
34500	ENTENDER	
34600	ENTURNAR	

34700	ENIRAR			2
34800	ENTREGAR			2
34900	ENTRETER			2
35000	ENTRISTECER			2
35100	ENTUPIR			2
35200	ENVELHECER			2
35300	ENVIAR			2
35400	EQUILIBRAR			4
35500	ERRAR			1
35600	ESCAPAR			2
35700	ESCLARECER			2
35800	ESCOLHER			2
35900	ESCONDER			2
36000	ESCREVER			2
36100	ESCUSAR			2
36200	ESFORÇAR			2
36300	ESFREGAR			2
36400	ESGOTAR			2
36500	ESPALHAR			2
36600	ESPANCAR			2
36700	ESPECIFICAR			2
36800	ESPERAR			2
36900	ESPETAR			2
37000	ESPREITAR			2
37100	ESQUECER			2
37200	ESTABELECER			2
37300	ESTAGIAR			2
37400	ESTAR			1
37500	ESTEJAR	ESTAR	PRSS	5
37600	ESTENDER			2
37700	ESTEVER	ESTAR	PRTI	5
37800	ESTIMAR			2
37900	ESTIMULAR			3
38000	ESTIVER	ESTAR	IMPS	5
38100	ESTORVAR			2
38200	ESTRAGAR			2
38300	ESTRANHAR			2
38400	ESTREAR			1
38500	ESTREMECER			2
38600	ESTRUTURAR			2
38700	ESTUDAR			2
38800	ESVASIAR			2
38900	EVITAR			2
39000	EXAGERAR			2
39100	EXAMINAR			2
39200	EXAURIR			2
39300	EXCEDER			3
39400	EXCLAMAR			1
39500	EXCLUIR			2
39600	EXECUTAR			2
39700	EXERCER			2
39800	EXIGIR			2
39900	EXISTIR			1
40000	EXPERIMENTAR			2
40100	EXPLICAR			2
40200	EXPLICITAR			2
40300	EXPLODIR			2
40400	EXPLORAR			2
40500	EXPLORAR			2
40600	EXPOR			2
40700	EXPRESSAR	EXPRIMIR	PARP	2
40800	EXPRIMIR			2
40900	EXTRAIR			2
41000	EXUMAR			4
41100	FACER	FAZER		2
41200	FACILITAR			2
41300	FALAR			1
41400	FALIR			1
41500	FALTAR			3
41600	FASCINAR			3
41700	FAZER			2
41800	FECHAR			2
41900	FERIR			2
42000	FERVER			2
42100	FESTEJAR			2
42200	FICAR			1
42300	FINANCIAR			2
42400	FINGIR			2
42500	FIZER	FAZER	PRTI	2
42600	FOCAR			2
42700	FOQUER	FOCAR		2
42800	FOR	IR		1
42900	FORCAR			2
43000	FORMAR			2
43100	FORMAR-SE			3
43200	FORNECER			2
43300	FOTOGRAFAR			2
43400	FRACASSAR			2

43500	FR	SER		3
43600	FREQUENTAR			2
43700	FUGIR			1
43800	FUJAR	FUGIR	PRSS	1
43900	FUMAR			2
44000	FUNCIONAR			1
44100	GANHAR			2
44200	GARANTIR			2
44300	GASTAR			2
44400	GEMER			3
44500	GERAR			2
44600	GERIR			2
44700	GOSTAR			3
44800	GOVERNAR			3
44900	GOZAR			3
45000	GRAVAR			2
45100	GRITAR			1
45200	GUARDAR			2
45300	GUIAR			2
45400	HABITAR			3
45500	HARMONIZAR			2
45600	HAVER			1
45700	HERDAR			2
45800	HESITAR			3
45900	HOSPITALIZAR			1
46000	IMAGINAR			2
46100	IMITAR			2
46200	IMPEDIR			2
46300	IMPOR			2
46400	IMPORTAR			2
46500	IMPRIMIR			2
46600	INCLINAR			2
46700	INCLUIR			2
46800	INCOMODAR			3
46900	INCOMBIR			3
47000	INDICAR			3
47100	INDIGNAR			3
47200	INDIGNAR-SE			3
47300	INFLAMAR			2
47400	INFORMAR			2
47500	INICER	INICIAR	PRSI	4
47600	INICIAR			2
47700	INJETAR			2
47800	INSISTIR			3
47900	INSTALAR			2
48000	INSTITUIR			2
48100	INTERESSAR			2
48200	INTERPRETAR			2
48300	INTERROGAR			2
48400	INTERROMPER			3
48500	INTRODUZIR			2
48600	INUNDAR			2
48700	INVENTAR			2
48800	IRRITAR			3
48900	IRROMPER			3
49000	JANTAR			1
49100	JOGAR			2
49200	JULGAR			2
49300	JUNTAR			2
49400	JURAR			2
49500	JUSTIFICAR			2
49600	LABUTAR			2
49700	LAMENTAR			2
49800	LANCAR			2
49900	LARGAR			2
50000	LAVAR			2
50100	LEMBRAR			2
50200	LER			2
50300	LEVANTAR			2
50400	LEVAR			2
50500	LIBERTAR			2
50600	LIDERAR			2
50700	LIDERAR			2
50800	LIMITAR			2
50900	LIXAR			2
51000	LIGAR			2
51100	LIMPAR			2
51200	LISTAR			2
51300	LIVRAR			2
51400	LOCALIZAR			2
51500	LOUVAR			2
51600	LUTAR			2
51700	MAGOAR			3
51800	MANDAR			2
51900	MANIFESTAR			2
52000	MANTER			2
52100	MARCAR			2
52200	MATAR			2

52300	MEDIR	2		
52400	MELHORAR	2		
52500	MENCIONAR	2		
52600	MENTIR	3		
52700	MERECER	2		
52800	METER	2		
52900	MEXER	2		
53000	MISTURAR	2		
53100	MODIFICAR	2		
53200	MOLHAR	2		
53300	MORAR	1		
53400	MORRER	1		
53500	MOSTRAR	2		
53600	MOTIVAR	2		
53700	MUDAR	2		
53800	MULTAR	2		
53900	NADAR	1		
54000	NASCER	1		
54100	NECESSITAR	2	3	
54200	NEGAR	2		
54300	NOMEAR	2	3	
54400	NOTAR	2		
54500	NUMERAR	2		
54600	OBEDECER	3		
54700	OBJETIVAR	3		
54800	OBRIGAR	3		
54900	OBSERVAR	2		
55000	OBTER	2		
55100	OCORRER	2		
55200	OCUPAR	2		
55300	OFENDER	3		
55400	OFERECER	4		
55500	ONERAR	2		
55600	OPOR	3		
55700	ORGANIZAR	2		
55800	ORGULHAR	3		
55900	PAGAR	2		
56000	PARAR	2		
56100	PARECER	5		
56200	PARTICIPAR	3		
56300	PARTIR	1	2	
56400	PASSAR	2		
56500	PASSEAR	1		
56600	PEDIR	2		
56700	PEGAR	2		
56800	PENDURAR	2		
56900	PENSAR	2		
57000	PENTEAR	2		
57100	PERCEBER	2		
57200	PERDER	2		
57300	PERDOAR	2		
57400	PERGUNTAR	2		
57500	PERMANECER	1		
57600	PERMITIR	3		
57700	PERSEGUIR	2		
57800	PERSUADIR	2		
57900	PERTENCER	3		
58000	PERTURBAR	2		
58100	PESAR	2		
58200	PESCAR	2		
58300	PESQUISAR	2		
58400	PISCAR	2		
58500	PINTAR	2		
58600	PISAR	1	2	
58700	PLANTAR	2		
58800	PODER	1		
58900	POSITIVAR	2		
59000	POSSIBILITAR	2		
59100	POSSUIR	2		
59200	POUPAR	2		
59300	PRATICAR	2		
59400	PRECEDER	3		
59500	PRECISAR	2		
59600	PREDOMINAR	2		
59700	PREFERIR	2		
59800	PREJUDICAR	2		
59900	PREGAR	2		
60000	PRENDER	2		
60100	PREOCUPAR	3		
60200	PREPARAR	2		
60300	PRESERVAR	2		
60400	PRESTAR	2		
60500	PRESSUPOR	2		
60600	PRETENDER	2	4	6
60700	PREVENIR	3		
60800	PROCEDER	1		
60900	PROCURAR	2		
61000	PROCLAMAR	2		
61100	PRODUZIR	2		

61100	PRODUZIR			2	
61200	PROGRAMAR			2	
61300	PROGREDIR			1	
61400	PROIBIR			2	
61500	PROMETER			1	4 6
61600	PROMOVER			2	
61700	PRONUNCIAR			2	
61800	PROPOR			2	
61900	PROPORCIONAR			2	
62000	PROSSEGUIR			2	
62100	PROVAR			2	
62200	PROVER			3	
62300	PROVOCAR			2	
62400	PUBLICAR			2	
62500	PUXAR			2	
62600	PURGAR			2	
62700	QUEBRAR			2	
62800	QUEIRER	QUERER		2	
62900	QUEIRAR	QUERER		2	
63000	QUERER			2	
63100	RACIOCINAR			1	
63200	RACHAR			2	
63300	RALHAR			1	
63400	RAPTAR			2	
63500	RASGAR			2	
63600	REAGIR			2	
63700	REALIZAR			2	
63800	REBENTAR			2	
63900	RECAIR			3	
64000	RECEAR			2	
64100	RECEBER			2	4
64200	RECLAMAR			2	
64300	RECLASSIFICAR			3	
64400	RECOMENDAR			3	
64500	RECONCILIAR			3	
64600	RECONDUZIR			2	
64700	RECONHECER			2	
64800	RECONSTRUIR			2	
64900	RECORDAR			2	
65000	RECRIAR			2	
65100	RECUPERAR			2	
65200	RECUSAR			2	
65300	REDIGIR			3	
65400	REDUZIR			2	
65500	REESCREVER			2	
65600	REFERIR			3	
65700	REFINAR			2	
65800	REFILMAR			2	
65900	REFERIR			3	
66000	REFLETIR			3	
66100	REFORÇAR			2	
66200	REFUTAR			2	
66300	REGEITAR			2	
66400	REGREDIR			3	
66500	REGRESSAR			3	
66600	REGULAMENTAR			2	
66700	REGULAR			2	
66800	RELACIONAR			2	
66900	REMETER			2	
67000	RENDER			2	
67100	RENOVAR			2	
67200	REPARAR			2	4
67300	REPERCUTIR			2	
67400	REPETIR			2	
67500	REPRESENTAR			2	
67600	RESERVAR			1	
67700	RESIDIR			3	
67800	RESISTIR			2	
67900	RESOLVER			2	
68000	RESPEITAR			2	
68100	RESPIRAR			3	4
68200	RESPONDER			2	
68300	RESTABELECER			2	
68400	RESTRINGIR			3	
68500	RESULTAR			3	
68600	RETIRAR			3	
68700	RETORNAR			2	
68800	REUNIR			1	
68900	RIR			3	
69000	RODEAR			2	
69100	ROLAR			2	
69200	ROMPER			2	4
69300	ROUBAR			1	
69400	SABER			2	
69500	SACUDIR			1	
69600	SAIR			2	
69700	SALTAR			3	
69800	SALVAR				

69900	SATISFAZER		3
70000	SECAR		2
70100	SEGUIR		2
70200	SEGURAR		2
70300	SEJAR	SER 5	
70400	SELECIONAR		2 4
70500	SENTAR		1
70600	SENTIR		2
70700	SEPARAR		2
70800	SER		5
70900	SERVIR		3
71000	SIGNIFICAR		2
71100	SOAR		2
71200	SOBREVIVER		1
71300	SOFRER		3
71400	SOLICITAR		2
71500	SOLTAR		2
71600	SONHAR		1
71700	SORRIR		1
71800	SUAR		1 2
71900	SUBIR		1
72000	SUBSTITUIR		2 3
72100	SUGAR		2
72200	SUGERIR		2 4
72300	SUJAR		2
72400	SUPERAR		3
72500	SUPOR		2
72600	SUPRIMIR		2
72700	SURGIR		1
72800	SURPREENDER		3
72900	SUSPENDER		2
73000	SUSTAR		2
73100	SUSTENTAR		2
73200	TELEFONAR		3
73300	TENCIONAR		2 6
73400	TENDER		3
73500	TENTAR		2 6
73600	TER		2
73700	TESTAR		2
73800	TIRAR		2
73900	TOCAR		2
74000	TOLERAR		2
74100	TOMAR		2
74200	TORCER		3
74300	TORNAR		2 6
74400	TOCIR		1
74500	TRABALHAR		1
74600	TRADUZIR		2
74700	TRANSBORDAR		1
74800	TRANSFORMAR		2
74900	TRANSMITIR		2
75000	TRANSPORTAR		2
75100	TRATAR		2
75200	TRAZER		2
75300	TREMER		1
75400	TROCAR		2
75500	TROUXER	TRAZER PRT	2
75600	ULTRAPASSAR		2
75700	USAR		2
75800	UTILIZAR		2
75900	VALER		3
76000	VARIAR		1
76100	VEDAR		2
76200	VENCER		2
76300	VENDER		2
76400	VENHAR	VIR PRSS	1
76500	VER		2
76600	VERIFICAR		2
76700	VESTIR		2 4
76800	VIAJAR		1
76900	VINGAR		2
77000	VIR		1
77100	VIRAR		2
77200	VISAR		2
77300	VISITAR		2
77400	VIVER		1
77500	VOAR		1
77600	VOLTAR		3
77700	VOTAR		3
77800	ZANGAR		3
77900			

CODIGO DA REGENCIA VERBAL : 1 = VERBO INTRANSITIVO
2 = VERBO TRANSITIVO DIRETO
3 = VERBO TRANSITIVO INDIRETO
4 = VERBO TRANS. DIRETO E IND.
5 = VERBO DE LIGACAO
6 = VERBO MODAL

WORKFILE: PKR/NDIC (09/17/85)

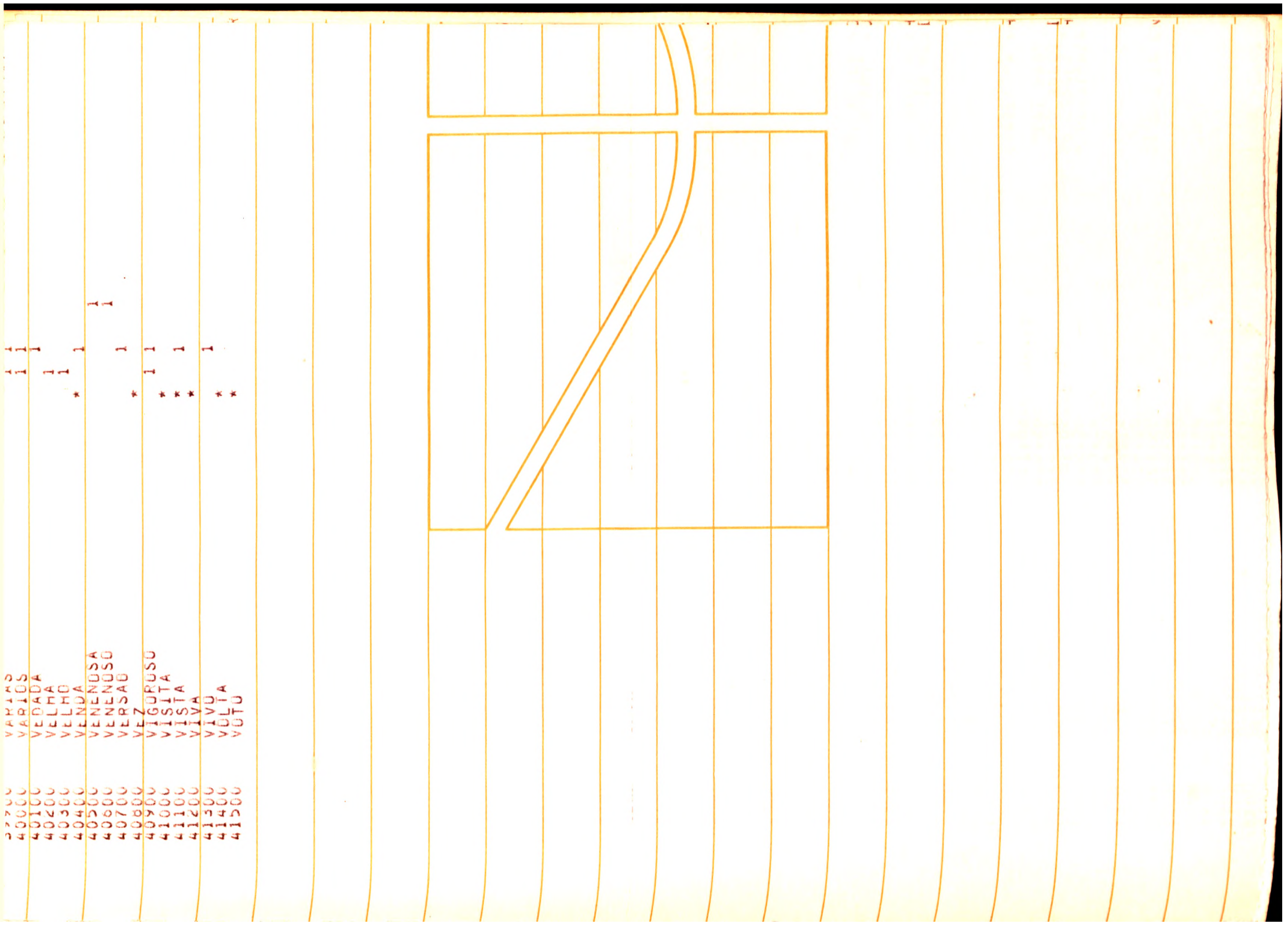
	NOUN/ADJ-LEXEM	HOM	ADJ1	NOES	PAR (22.9.83)
100	ABAJOUR				
200	ABANDONO	*			
300	ABERTA		1	1	1
400	ABERTURA	*			
500	ABOMINAVEL		1		
510	ABSTRATO		1		
520	ACO			1	1
600	ACORDO	*		1	1
700	ACRESCIMO				1
800	ADICIONAL			1	
900	ADORAVEL		1	1	
1000	ADVERTENCIA			1	1
1100	AGENTE			1	1
1200	AGRUPAMENTO				
1300	AJUDA	*		1	1
1400	ALTO		1	1	
1500	AMBITO			1	
1600	AMOSTRA				1 1
1700	AMPLO		1	1	
1800	ANALISE	*			
1900	ANIMAL			1	1
2000	ANTIGO		1	1	
2100	APLAUDIDO		1	1	
2200	APLICABILIDADE				1
2300	APLICADA	*	1	1	1
2320	APLICADO	*	1	1	1
2330	APLICACAO			1	
2500	APLICAVEL				1
2600	APOIO	*			
2700	APROVADA			1	
2800	AQUISICAO			1	
2900	AREA			1	
3000	ASSISTENCIA			1	1
3100	ATRASO	*			
3200	ATENTO		1	1	
3300	ATO	*			
3400	ATUAL			1	
3500	AUMENTO	*			
3600	AUSENCIA			1	
3700	AUTOR				1
3800	AUXILIO				1
3900	BAIRRO			1	
4000	BAIXA	*	1		1
4100	BAIXO		1	1	
4200	BANCO				1 1
4300	BASEADA			1	
4400	BASICA				1
4500	BASICO			1	
4600	BASTANTE		1	1	
4700	BELA		1	1	
4800	BELO		1	1	
4900	BENEFICIARIO				1 1
5000	BIZARRO		1		
5010	BOM		1	1	
5100	BRILHANTE		1		1
5200	BUSCA	*			1 1
5300	CACA	*			1 1
5400	CALMO		1	1	
5500	CAMINHO	*		1	
5600	CAMPO				1 1
5700	CAPITULO			1	
5800	CAPTURADA				1
5900	CARTA				1 1
6000	CARTEIRA				1
6100	CASA	*			
6200	CASO	*		1	
6300	CAUSA	*			1
6400	CENSURA	*			
6500	CERTA		1		
6550	CERTEZA			1	
6600	CERTO	*	1	1	
6700	CHOQUE	*			
6800	CINZENTO		1		1 1
6900	COBERTURA		1		
7000	COBRA	*			
7100	COISA			1	
7200	COMPLETA	*		1	
7300	COMPLETO	*	1	1	
7400	COMPLEXO				1
7500	COMPRA	*			
7600	COMPREENDIDA	*	1		
7620	COMPREENDIDO	*	1		

7650	COMPREENDIDO	*	1				
7700	COMUM			1			
7800	CONDICAO				1	1	
7900	CONFERENCIA			1			
8000	CONTRARIO			1			
8100	CONJUNTO					1	
8200	CONQUISTA	*					
8300	CONTA	*			1	1	
8400	CONTEXTO			1			
8500	CONTO	*			1	1	
8600	CONTRATO	*				1	
8700	CONTROLE				1	1	
8800	CORRECAO				1	1	
8900	CORRESPONDENTE					1	
9000	CORRENTE		1		1	1	
9100	CORRETA			1			
9200	CORRETO	*		1			
9300	COTIDIANO		1	1			
9400	CRIA	*			1	1	
9500	CRITICA	*					
9600	CRUEL		1	1			
9700	CUMPRIMENTO	*			1		
9800	CURIOSO		1	1			
9900	CUSTO	*	1				
10000	DATILOGRAFIA				1		
10100	DECORRENCIA			1			
10120	DENOMINADA	*	1				
10130	DENOMINADO	*	1				
10200	DENTRO			1			
10300	DEFESA	*					
10400	DEPENDENCIA					1	
10500	DEPOSITO	*					
10600	DESCANSO	*					
10700	DESEJO	*					
10800	DESEMPENHO	*					
10810	DESENCONTRO	*	1				
10820	DESINTERESSANTE		1				
10830	DESPREZIVEL		1				
10900	DESTEMIDO		1	1			
11000	DETERMINADO		1				
11100	DIFICIL		1	1			
11200	DIRETRIZ					1	
11300	DISCIPLINA				1	1	
11400	DISPENSA	*					
11500	DISPOSICAO					1	
11600	DISPOSTO					1	
11700	DISPUTA	*					
11800	DISTINTO		1	1			
11900	DITO						1
12000	DOCE		1				
12001	DOTADA		1				
12002	DOTADO	*	1				
12100	EFEITO					1	
12101	EFICAZ		1				
12200	ELEGANTE		1				
12300	ELEMENTO					1	
12400	EMOCIONANTE		1				
12500	ENCONTRO	*					
12600	ENDERECO			1			
12700	ENTIDADE					1	
12800	EMPREGO	*					
12900	ENORME		1	1			
13000	ENTRADA			1			
13100	ENTREGA	*					
13200	EQUIVALENCIA				1	1	
13300	ERRO	*					
13400	ERRADO			1			
13500	ESCALA				1	1	
13600	ESCANDALOSO		1	1			
13700	ESCOLHA	*					
13800	ESPECIAL		1	1			
13900	ESPERA	*					
14000	ESPINHOSO		1			1	
14100	ESPONTANEO		1	1			
14200	ESQUECIDO		1	1			
14300	ESTA	*		1			
14400	ESTACAO				1	1	
14500	ESTADO					1	
14600	ESTRANHA		1				
14700	ESTRANHO		1				
14800	ESTRATO				1	1	
14900	ESTRUTURA	*			1	1	
14990	ESTUDIOSA		1				
14991	ESTUDIOSO		1				
15000	ESTUDO	*					
15100	ETERNA		1	1			
15200	EVENTUAL			1			
15300	EXATO		1	1			

15400	EXAGERO	*					
15500	EXECUCAO			1			
15600	EXERCICIO					1	
15700	EXPERIENTE		1	1			
15800	EXPLENDOROSO		1	1			
15900	EXPRESSIVA		1	1			
16000	EXTENSO		1	1			
16100	EXTRAORDINARIO		1	1			
16200	EXUBERANTE		1	1			
16300	FACIL		1	1			
16400	FACA	*					
16500	FALA	*					
16600	FALAR	*					
16660	FALSA		1				
16700	FALTA	*					
16800	FAVOR			1			
16900	FEIO		1				
17000	FEITO	*					1
17100	FELIZ		1				
17200	FIM		1	1			
17300	FINAL		1				
17400	FINALIDADE		1				
17500	FINANCEIRO					1	
17600	FINITO		1		1	1	
17660	FIRME		1				
17700	FIXACAO			1			
17800	FLEXIBILIDADE			1			
17900	FOCO	*			1	1	
18000	FONTE				1	1	
18100	FORMA	*			1	1	
18200	FORTE	*			1	1	
18300	FRACASSO	*					
18400	FRAGIL		1	1			
18500	FREQUENTE	*	1	1			
18600	FREQUENTES	*	1		1	1	
18700	FRONTEIRA				1	1	
18800	FUNCAO				1	1	
18900	FUNDO					1	
19000	FUGA	*					
19100	FUTURO		1	1			
19200	GARANTIA	*	1			1	
19300	GOSTO	*		1			
19400	GOVERNO	*					
19500	GRANDE			1			
19600	GUARDA	*			1	1	
19700	HOMEM						
19800	HORA			1			
19900	HORROROSO		1	1			
20000	IMENSO		1	1			
20100	IMPORTANCIA			1			
20200	IMPORTANTE		1	1			
20300	INCENDIARIO		1		1		
20400	INCISIVO		1			1	
20500	INDESTRUTIVEL		1		1		
20510	INEVITAVEL		1				
20600	INFELIZ		1	1			
20700	INFLUENCIA	*					
20800	INFINITO		1	1			
20900	INICIO	*					
21000	INQUIETO		1	1			
21100	INSTITUICAO				1	1	
21200	INSTRUCAO				1	1	
21300	INTERESSE	*					
21400	INTEIRO			1			
21500	INTERESSANTE			1			
21600	INTRODUZIDO			1			
21700	INSTRUMENTO				1	1	
21800	INUTIL		1	1			
21900	IRREQUIETO		1	1			
22000	ITEM			1			
22100	JOVIAL		1	1			
22200	JULHO			1			
22300	JURISDICIONAL					1	
22400	JUNTO			1			
22500	LADO			1			
22600	LINDO		1	1			
22700	LINGUA				1	1	
22800	LINHA			1			
22900	LISTA	*		1			
23000	LIVRE		1			1	
23100	LIXO	*					
23200	LOCAL			1			
23300	LUTA	*					
23400	MA		1	1			
23500	MAIOR		1	1			
23501	MAIORIA		1				
23600	MARAVILHOSO		1	1			
23700	MATA	*			1	1	

23801	MATERNA				1	1
23802	MA		1	1		
23900	MAU		1	1		
24000	MAXIMO			1		
24100	MEIA				1	1
24200	MELHORAMENTO					1
24300	MENCAO	*				
24400	MENOR		1	1		
24500	METODO				1	1
24600	MINIMO				1	
24700	MISTURA	*				
24800	MODIFICACAO				1	
24900	MONOGRAFIA				1	
25000	MORTO				1	
25100	MOSTRA	*				
25200	MEMBRO				1	1
25300	MERO		1	1		
25400	MIMOSO		1	1		
25500	MISTERIOSA		1			
25600	MISTERIOSO		1			
25700	MODELO				1	1
25800	MODESTO		1	1		
25900	MODO					1
26000	MULHER				1	
26100	MULTA	*				
26200	NATUREZA			1		
26300	NECESSARIO			1		
26400	NORMA				1	1
26500	NORTE			1		
26600	NORMAL			1		
26700	NOME			1		
26800	NOTAVEL		1	1		
26900	NOVA		1	1		
27000	NOVO		1	1		
27100	NUCLEO			1		1
27200	NUMERO	*			1	1
27300	OBJETIVO	*		1		
27400	OBJETO			1		
27500	OBRIGATORIO					1
27600	OESTE			1		
27700	ONEROSO		1	1		
27800	OPERACAO				1	1
27900	OPEROSO		1			
28000	ORGAO				1	1
28100	OSTENSIVO		1	1		
28120	PAR		1			
28200	PARA					
28300	PALAVRA				1	1
28400	PARCELA				1	1
28500	PARTE	*		1		
28600	PARTICULAR		1			
28700	PASSAGEM					
28800	PASSIVO		1			
28900	PASSO	*		1		
29000	PEDACO			1		
29100	PELO			1		
29200	PEQUENO		1	1		
29300	PERFEITO		1	1		
29400	PERGUNTA	*		1		
29410	PERNICIOSA		1			
29411	PERNICIOSO		1			
29500	PESO	*			1	
29600	PESQUISA	*		1		
29700	PLANO				1	
29800	POBRE		1	1		
29900	PONTO			1		
30000	POSSIVEL		1	1		
30100	POSTO	*				
30200	POUCA			1		
30300	PRATICA	*			1	
30400	PRAZO			1		
30500	PRECISA	*		1		
30600	PRECISO	*				
30700	PRE-FABRICADA					
30800	PREFERENCIA			1		
30900	PRESENTE		1	1		
31000	PRESTIGIOSO		1	1		
31100	PRETENDENTE			1		
31200	PRIMEIRA		1	1		
31300	PRIMEIRO		1	1		
31400	PROBLEMA			1		
31500	PROCEDIMENTO			1		
31600	PROCESSO				1	1
31700	PROCURA	*			1	1
31711	PROFUNDA		1			
31712	PROFUNDO		1			
31800	PROGRAMA	*			1	1

31900	PRONTO			1			
32000	PROPRIA		1	1		1	
32100	PROPRIO		1	1		1	
32200	PRORROGACAO					1	1
32300	PROVERBIAL		1		1		
32400	PROVIDENCIA			1			
32500	PROXIMA			1			
32600	PROXIMO			1			
32700	PUBLICO	*					
32800	QUADRO				1	1	
32900	QUALIDADE			1			
33000	QUENTE		1	1			
33100	RARO		1	1			
33200	RAZAO			1			
33230	RECAI		1				
33300	REFERENCIA				1	1	
33400	REFERIDA			1			
33500	REGIAO					1	
33600	REGULAMENTO	*					
33700	RELATIVO			1			
33770	RELEVANCIA		1				
33800	RELEVANTE		1				
33900	RELIGIOSO		1	1			
34000	RENDA	*					
34100	RENOMADO		1	1			
34200	REPETIDO			1			
34300	REQUISITO			1			
34400	RESERVA	*					
34500	RESOLUCAO			1			
34600	RESPEITO	*			1	1	
34610	RESPONSAVEL		1				
34700	RESPOSTA			1			
34800	RESTO	*		1			
34900	RESULTADO	*		1			
35000	RETORNO	*		1			
35100	RETUMBANTE		1				
35110	REVOLUCIONARIO		1				
35200	SALDO	*		1			
35300	SANTA			1			
35400	SALVO			1			
35500	SANTO		1		1		
35600	SAO			1			
35700	SAUDOSO		1			1	
35800	SEGUIDA			1			
35900	SEGUINTE		1	1			
36000	SEGUNDO		1	1			
36100	SEGURO	*			1	1	
36200	SELECAO		1				
36300	SENTIDO	*		1			
36400	SER				1		
36500	SIMBOLO				1	1	
36600	SIMPLES		1	1			
36700	SINUOSO		1	1			
36800	SISTEMA					1	
36900	SOFISTICADO			1			
37000	SOL				1	1	
37100	SOLIDARIO			1			
37200	SOLITARIO		1	1			
37300	SOMBRIO		1	1			
37400	SUAVE		1				
37500	SUMA	*	1	1			
37600	SUPERINTENDENCIA					1	
37700	SUPERIOR					1	
37800	SURRADO		1	1			
37900	TABELA				1	1	
38000	TALENTOSO		1				
38100	TEMA			1			
38200	TEMPO			1			
38300	TERCEIRA		1				
38400	TERCEIRO		1				
38500	TERMO			1			
38600	TIPICO		1	1			
38700	TIPO			1			
38800	TODA			1			
38810	TOTAL		1				
38900	TRABALHO	*			1	1	
39000	TRANSFERENCIA					1	
39100	TRANSPORTE	*					
39200	TRISTE		1	1			
39300	UNICO		1	1			
39400	UNIDADE			1			
39500	ULTIMA		1	1			
39600	ULTIMO		1	1			
39700	USO	*		1			
39710	UTILIZADA		1				
39711	UTILIZADO	*	1				
39770	VALIOSO		1				
39800	VANTAGEM			1			



WORKFILE: FKR/ENGLISH/FDIC (09/17/85)

	FWORD	FAC CNP M G CNP	INFORMACOES EXTRAS
100		T M	
200		I /	
300	FWORD	FAC CNP M G CNP	INFORMACOES EXTRAS
400	***	*** *** * * ***	**********
500	A	109S	+18 "85 A WI
600	ABACK	126	
700	ABOARD	218.26	
800	ABOUT	318.22.26	22 ONLY WITH
900	ABOVE	302.18.26	
1000	ABREAST	126	
1100	ABROAD	126	
1200	ACROSS	218.26	
1300	AFAR	126	
1400	AFLOAT	206.26	
1500	AFOOT	206.26	
1600	AFTER	318.25.26	+02 "AFTER
1700	AFTERWARDS	126	
1800	AGAIN	126	
1900	AGAINST	118	
2000	AGO	202.22	02 ALWAYS AF
2100	AHEAD	206.26	22 ONLY W
2200	ALBEIT	125	
2300	ALIKE	206.26	
2400	ALIVE	206.26	
2500	ALL	411.12.22.26	1
2600	ALMOST	222.26	
2700	ALOFT	206.26	
2800	ALONE	206.26	
2900	ALONG	218.26	
3000	ALONGSIDE	126	
3100	ALOOF	302.00.26	
3200	ALOUD	126	
3300	ALREADY	126	
3400	ALSO	126	
3500	ALTHOUGH	125	
3600	ALTOGETHER	126	+01 SLANG
3700	ALWAYS	126	
3800	AM	203S1PRS 14	
3900	AMID	118	
4000	AMIDST	118	
4100	AMISS	206.26	
4200	AMONG	118	
4300	AMONGST	118	
4400	AN	109S	+18 "85 AN H
4500	AND	124	
4600	ANOTHER	211.12	
4700	ANY	311.12.22	
4800	ANYBODY	213.29	
4900	ANYHOW	225.26	
5000	ANYONE	213.29	
5100	ANYTHING	213.29	
5200	ANYWAY	225.26	
5300	ANYWHERE	225.26	
5400	APART	126	+02 AFTER NO
5500	ARE	203.14	APART" CU
5600	AREN'T	303.14.26	
5700	AROUND	218.26	
5800	AS	310.22.25	22 ONLY WITH
5900	ASHORE	126	+ "18 MODI
6000	ASIDE	201.18	
6100	ASKEW	126	
6200	ASUNDER	126	
6300	AT	118	
6400	AWAY	126	
6500	AWHILE	126	USAGE; "A WH
6600	BACK	501.02.03.04.26	
6700	BACKWARD	202.26	WENT BAD"
6800	BACKWARDS	126	BAD ABOUT I
6900	BAD	301.02.26	26 USAGE; "H
7000	BARELY	222.26	+ QUASI-NEG
7100	BE	203.14	
7200	BECAUSE	219.25	
7300	BEEN	203.14	
7400	BEFORE	318.25.26	
7500	BEHALF	201.20	01 OBJ 18 ON
7600	BEHIND	218.26	
7700	BEING	201.23	
7800	BELOW	218.26	
7900	BENEATH	218.26	
8000	BESIDE	118	
8100	BESIDES	218.26	
8200	BEST	401.02.22.26	+ 03 SLANG

8500	BEYOND	301.16.26	
8600	BIG	202.26	
8700	BOTH	411.12.24.26	
8800	BUT	416.24.25.26	+ 10? NEG
8900	BY	218.26	
9000	CAN	301.03.14	I
9100	CANNOT	214.26	I
9200	CAN'T	214.26	I
9300	CERTAIN	206.11	+ "FOR CERTA
9400	CLEAN	402.03.22.26	
9500	CLEAR	302.03.26	
9600	CLOSE	401.02.03.26	I
9700	CONCERNING	218.23	
9800	CONSIDERING	218.23	
9900	COULD	114	
10000	COULDN'T	214.26	
10100	COUNTER	401.02.03.26	
10200	DARE	401.03.14.15	ANOMALOUS FI
10300	DAREN'T	314.15.26	ANOMALOUS FI
10400	DEAR	301.02.26	
10500	DEEP	401.02.22.26	
10600	DESPITE	118	
10700	DID	203.14	
10800	DIDN'T	214.26	
10900	DIRECT	302.03.26	
11000	DO	203.14	
11100	DOES	301.03.14	
11200	DOESN'T	214.26	
11300	DOING	123	
11400	DONE	203.14	
11500	DON'T	214.26	
11600	DOUBLE	401.02.03.26	
11700	DOWN	601.02.03.06.18.26	
11800	DOWNHILL	202.26	
11900	DOWNRIGHT	202.22	
12000	DOWNWARD	202.26	
12100	DOWNWARDS	126	
12200	DUE	401.02.19.22	22 "DUE EAST
12300	DURING	118	
12400	EACH	211.12	
12500	EARLY	202.26	
12600	EAST	301.02.26	
12700	EIGHT	207.08	
12800	EITHER	411.12.24.26	24 WITH "OR"
12900	ELSE	126	
13000	ELSEWHERE	126	
13100	ENOUGH	301.02.26	
13200	EVEN	302.03.26	
13300	EVER	126	+ 22 "EVER S
13400	EVERY	111	MODIFIES I
13500	EVERYBODY	213.29	
13600	EVERYONE	213.29	
13700	EVERYTHING	213.29	
13800	EVERYWHERE	126	
14000	EXCEPT	303.18.25	I AFTER "NOT" ETC. ONLY EXCE
14100	EXCEPTING	320.23.25	
14300	FAIR	301.02.26	
14400	FAR	301.02.26	
14500	FARTHER	401.02.03.26	
14600	FARTHEST	301.02.26	
14700	FAST	401.02.03.26	
14800	FEW	211.12	
14900	FIFTH	207.08	
15000	FINE	401.02.03.26	
15100	FIRM	401.02.03.26	
15200	FIRST	311.12.26	
15300	FIVE	207.08	
15400	FLAT	401.02.03.26	
15500	FOR	318.24.25	
15600	FOREMOST	202.26	
15700	FOREVER	126	
15800	FORWARD	401.02.03.26	
15900	FORWARDS	203.26	
16000	FORMER	211.12	
16100	FORTH	126	
16300	FOUL	401.02.03.26	
16400	FOUR	207.08	
16500	FOURTH	207.08	
16600	FREE	302.03.26	
16700	FULL	202.26	
16800	FURTHER	302.03.26	
16900	FURTHERMORE	126	
17000	FURTHEST	202.26	
17100	GET	303.14.15	+ 01 ANIMAL
17200	GOOD	201.02	IRR ATT CF
17300	GOT	303.14.15	I

17400	GLITTEN	303.14.15	ONLY USA
17500	HAD	303.14.15	
17600	HADN'T	403.14.15.28	
17700	HALF	311.12.22	
17800	HALFWAY	126	
17900	HARD	202.26	+ 01
18000	HARDLY	222.26	+ QUASI-NEG
18100	HAS	303.14.15	
18200	HASN'T	303.14.28	I
18300	HAVE	303.14.15	+ 01
18400	HAVEN'T	303.14.28	I
18500	HE	1135S M	
18600	HEAVY	202.26	
18700	HENCE	225.26	
18800	HER	211.29S3 F	
18900	HERE	126	+ OBJ 18 ONL
19000	HEREAFTER	126	
19100	HEREWITH	126	
19200	HERS	213.29	
19300	HERSELF	1	REFLEXIVE
19400	HIGH	202.26	+ 01 OBJ 18
19500	HIGHLY	222.26	AND SLANG
19600	HIM	129S3 M	
19700	HIMSELF	1	REFLEXIVE
19800	HIS	211.12	
19900	HOME	401.02.03.26	
20000	HOMELY	102	
20100	HOW	322.23.26	
20200	HOWEVER	422.24.25.26	
20300	I	113S1	
20400	IF	125	
20500	ILL	301.02.26	
20600	IN	216.26	+ 02 "IN PA
20700	INASMUCH	126	USUALLY
20800	INDEED	222.26	
20900	INDOORS	126	
21000	INSIDE	401.02.18.26	
21100	INSTEAD	219.26	
21200	INTO	116	
21300	IS	203.14	
21400	ISN'T	303.14.28	
21500	IT	213.29	
21600	ITS	111	
21700	ITSELF	1	REFLEXIVE
21800	JUST	302.22.26	
21900	KINDLY	202.26	
22000	LARGE	202.26	+ 01 OBJ "A
22100	LAST	403.11.12.26	+ 01 OBJ "U
22200	LATE	202.26	
22300	LATER	202.26	
22400	LATEST	202.26	
22500	LATTER	211.12	
22600	LEAST	311.12.22	
22700	LEFT	401.02.03.26	
22800	LEISURELY	202.26	
22900	LESS	511.12.18.22.26	
23000	LEST	125	
23100	LET	301.03.14	
23200	LIGHT	401.02.03.26	
23300	LIKE	601.02.03.18.25.26	
23400	LIKELY	202.26	
23500	LIKEWISE	224.26	
23600	LITTLE	301.02.26	
23700	LONG	302.03.26	
23800	LOST	302.03.26	
23900	LOTS	201.26	
24000	LOUD	202.26	
24100	LOW	401.02.03.26	
24200	MADE	302.03.15	
24300	MAKE	301.03.15	
24400	MAKES	301.03.15	
24500	MAKING	203.23	
24600	MANY	311.12.22	
24700	MAY	201.14	
24800	MAYBE	126	
24900	ME	129	
25000	MEANTIME	126	
25100	MEANWHILE	126	
25200	MIGHT	201.14	
25300	MIGHTN'T	214.28	
25400	MINE	401.03.13.29	
25500	MINUS	202.18	
25600	MORE	411.12.22.26	
25700	MOREOVER	126	
25800	MOST	411.12.22.26	
25900	MUCH	411.12.22.26	
26000	MUST	201.14	
26100	MUSTN'T	214.28	

26200	MY	111	
26300	MYSELF	1	REFLEXIVE
26400	NEAR	402.03.18.26	
26500	NEARBY	126	+ 02
26600	NEARLY	222.26	
26700	NEED	401.03.14.15	ANOMALOUS FI
26800	NEEDN'T	215.28	
26900	NEITHER	511.12.24.26.28	
27000	NEVER	226.28	
27100	NEVERTHELESS	224.26	
27200	NEW	202.22	22 NORMALLY
27300	NEXT	211.12	
27400	NINE	207.08	
27500	NINTH	207.08	
27600	NO	311.22.28	
27700	NOBODY	313.28.29	
27800	NONE	413.22.28.29	
27900	NOR	224.28	
28000	NORTH	301.02.26	
28100	NOT	226.28	
28200	NOTHING	313.28.29	
28300	NOTWITHSTANDING	218.26	
28400	NOW	225.26	+ 01 OBJ 16
28500	NOWHERE	226.28	
28600	OF	118	
28700	OFF	302.18.26	
28800	OFTEN	126	
28900	ON	218.26	+ 02 USUALLY
29000	ONCE	501.02.22.25.26	1
29100	ONE	407.08.11.12	
29200	ONESELF	1	REFLEXIVE
29300	ONLY	502.22.24.25.26	
29400	ONTO	118	
29500	ONWARD	202.26	
29600	ONWARDS	126	
29700	OPPOSITE	401.02.18.26	
29800	OR	124	
29900	OTHER	311.12.26	
30000	OTHERS	213.29	
30100	OTHERWISE	126	
30200	OUGHT	114	
30300	OUR	111	
30400	OURS	213.29	
30500	OURSELVES	1	REFLEXIVE
30600	OUT	401.02.19.26	
30700	OUTRIGHT	202.26	
30800	OUTSIDE	401.02.18.26	
30900	OUTWARD	202.26	
31000	OUTWARDS	126	
31100	OVER	218.26	
31200	OVERHEAD	301.02.26	
31300	OVERLAND	202.26	
31400	OVERLEAF	126	
31500	OVERNIGHT	202.26	
31600	OVERSEAS	202.26	
31700	OVERTIME	201.26	
31800	OWN	303.11.12	
31900	PARALLEL	401.02.03.26	
32000	PART	401.03.22.26	
32100	PAST	401.02.03.26	
32200	PER	118	
32300	PERHAPS	126	
32400	PLENTY	301.11.22	1
32500	PLUMB	401.02.03.26	
32600	PLUMP	401.02.03.26	
32700	PLUS	301.02.18	
32800	POOR	201.02	+ 26 USAGE
32900	POST	301.03.26	
33000	PRESENT	401.02.03.26	(26 OR 06?)
33100	PRETTY	202.22	
33200	PRIOR	301.02.19	
33300	QUICK	301.02.26	
33400	QUITE	222.26	
33500	RATHER	222.26	
33600	REALLY	222.26	
33700	RIGHT	401.02.03.26	
33800	ROUND	501.02.03.18.26	
33900	SAME	311.12.26	11 AND 12 W
34000	SAVE	401.02.18.25	1
34100	SCARCELY	222.26	+ QUASI-NEG
34200	SECOND	403.11.12.26	
34300	SELDOM	126	
34400	SEVEN	207.08	
34500	SEVERAL	211.12	
34600	SHALL	114	
34700	SHE	113	
34800	SHORT	401.02.03.26	
34900	SHOULD	114	

35000	SHOULDN'T	214.26	
35100	SIDEWAYS	202.26	
35200	SINCE	318.25.26	
35300	SIX	207.08	
35400	SLOW	302.03.26	
35500	SO	322.25.26	+29 SOME VE
35600	SOFT	202.26	
35700	SOON	126	
35800	SOME	211.12	
35900	SOMEBODY	213.29	
36000	SOMEHOW	126	
36100	SOMEONE	213.29	
36200	SOMETHING	213.29	
36300	SOMETIMES	126	
36400	SOMEWHAT	222.26	
36500	SOMEWHERE	126	
36600	SOUTH	301.02.26	R
36700	STILL	601.02.03.22.24.26	
36800	STRAIGHT	301.02.26	
36900	SUBJECT	401.02.03.19	
37000	SUCH	211.12	BEFORE 09 "
37100	SURE	202.26	? "FOR SURE'
37200	TALL	202.26	
37300	TANDEM	301.02.26	
37400	TEN	207.08	
37500	THAN	218.25	
37600	THANKS	203.19	
37700	THAT	510.11.12.22.25	
37800	THE	109	
37900	THEIR	111	
38000	THEIRS	213.29	
38100	THEM	129	
38200	THEMSELVES	1	REFLEXIVE
38300	THEN	402.24.25.26	+ OBJ 18 ON
38400	THERE	126	
38500	THEREABOUT	226.?	"2 LITERS OR
38600	THEREABOUTS	226.?	THEREABOUT(S
38700	THEREAFTER	225.26	
38800	THEREAT	126	
38900	THEREBY	126	
39000	THEREFORE	224.26	
39100	THESE	211.12	
39200	THEY	113	
39300	THIRD	211.12	
39400	THIRTY	207.08	
39500	THIS	211.12	
39600	THOSE	211.12	
39700	THOUGH	225.26	
39800	THREE	207.08	
39900	THROUGH	218.26	
40000	THROUGHOUT	218.26	
40100	THUS	224.26	
40200	TILL	401.03.18.25	
40300	TO	318.26.?	? INDICATOR
40400	TODAY	201.26	
40500	TOGETHER	126	
40600	TOMORROW	201.26	
40700	TONIGHT	201.26	
40800	TOO	222.26	
40900	TOWARD	118	
41000	TOWARDS	118	
41100	TRUE	202.26	+ 03
41200	TWENTY	207.08	
41300	TWICE	126	
41400	TWO	207.08	
41500	UNDER	302.18.26	
41600	UNDERGROUND	301.02.26	
41700	UNDERHAND	202.26	
41800	UNDERNEATH	401.02.18.26	
41900	UNLESS	125	
42000	UNTIL	218.25	
42100	UNTO	118	
42200	UP	601.02.03.06.18.26	
42300	UPHILL	202.26	
42400	UPON	118	
42500	UPRIGHT	301.02.26	
42600	UPWARD	102	
42700	UPWARDS	126	
42800	US	129	
42900	USED	302.03.15	
43000	VARIOUS	211.12	
43100	VERSUS	118	
43200	VERY	202.22	
43300	VIA	118	
43400	WAS	303.14.15	
43500	WASN'T	403.14.15.28	
43600	WE	113	
43700	WELL	401.02.03.26	

43500	WERE	203.14	
43600	WEREN'T	303.14.26	
43700	WHAT	410.11.12.25	
43800	WHATEVER	210.11	
43900	WHATSOEVER	126	
44000	WHEN	310.25.26	
44100	WHENEVER	210.25	
44200	WHERE	310.25.26	
44300	WHEREAS	125	
44400	WHEREBY	225.26	
44500	WHEREFORE	225.26	
44600	WHEREVER	125	
44700	WHETHER	125	
44800	WHICH	410.11.12.25	
44900	WHICHEVER	310.11.25	
45000	WHILE	301.05.25	+ OBJ 18 FOR
45100	WHO	310.13.29	29 USAGE
45200	WHOEVER	310.13.29	29 USAGE
45300	WHOLE	311.12.22	
45400	WHOM	210.29	
45500	WHOMEVER	210.29	
45600	WHOSE	210.11	
45700	WHY	225.26	
45800	WIDE	301.02.26	
45900	WILD	301.02.26	
46000	WILL	301.05.14	
46100	WITH	118	
46200	WITHIN	218.26	+ OBJ 18
46300	WITHOUT	218.26	+ OBJ 18
46400	WON'T	214.26	!
46500	WORSE	301.02.26	
46600	WORST	301.02.26	
46700	WOULD	114	
46800	WRONG	302.03.26	
46900	YES	201.26	+ AFFIRMATIV
47000	YESTERDAY	201.26	PARTICLE
47100	YET	224.26	
47200	YOU	213.29	
47300	YOUR	111	
47400	YOURS	306.13.29	
47500	YOURSELF	1	REFLEXIVE
47600	YOURSELVES	1	REFLEXIVE
47700	ZERO	207.08	

EXPLICACAO DE SIGLAS

. 7 ESPACOS EM BRANCO
! PALAVRA POSSIVELMENTE COM PROBLEMAS
+ CATEGORIA NAO INCLUIDA EM FAC; OU
POSSIVEL CATEGORIA NOVA