PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

EXTRAÇÃO AUTOMÁTICA DE CONCEITOS A PARTIR DE TEXTOS EM LÍNGUA PORTUGUESA

LUCELENE LOPES

Tese apresentada como requisito parcial à obtenção do grau de Doutor em Ciência da Computação na Pontifícia Universidade Católica do Rio Grande do Sul.

Orientadora: Renata Vieira

Porto Alegre
2012

L864e Lopes, Lucelene
Extração automática de conceitos a partir de textos em língua portuguesa / Lucelene Lopes. – Porto Alegre, 2012.
156 f.

Tese (Doutorado) – Fac. de Informática, PUCRS.
Orientador: Prof. Dr. Renata Vieira.

1. Informática. 2. Ontologia. 3. Processamento da Linguagem Natural. 4. Recuperação da Informação. I. Vieira, Renata. II. Título.

CDD 006.35

Ficha Catalográfica elaborada pelo
Setor de Tratamento da Informação da BC-PUCRS

Pontifícia Universidade Católica do Rio Grande do Sul
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

TERMO DE APRESENTAÇÃO DE TESE DE DOUTORADO

Tese intitulada "Extração Automática de Conceitos a partir de Textos em Língua Portuguesa", apresentada por Lucelene Lopes, como parte dos requisitos para obtenção do grau de Doutor em Ciência da Computação, Inteligência Computacional, aprovada em 26/01/2012 pela Comissão Examinadora:

Profa. Dra. Renata Vieira – Orientadora — PPGCC/PUCRS

Profa. Dra. María del Rosario Girardi Gutiérrez - — UFMA

Profa. Dra. Viviane Pereira Moreira - — UFRGS

Profa. Dra. Vera Lúcia Strube de Lima – — PPGCC/PUCRS

Homologada em 24/04/2012, conforme Ata No. 009 pela Comissão Coordenadora.

Prof. Dr. Paulo Henrique Lemelle Fernandes
Coordenador.

PUCRS
Campus Central
Av. Ipiranga, 6681 – P. 32 – sala 507 – CEP: 90619-900
Fone: (51) 3320-3611 – Fax (51) 3320-3621
E-mail: ppgcc@pucrs.br
www.pucrs.br/facin/pos

Ao meu esposo Paulo, às minhas filhas Karina Mylena, Maria Eduarda e a pequena Sophia (ainda em meu ventre)

... porque sem vocês eu não seria EU!

AGRADECIMENTOS

Obrigada a Laércio, Cida, Lucinéia, Lucélia, Leandro, Paulo, Karina Mylena, Maria Eduarda, Dirceu, José Lucio, Juliana, Matheus, Cota, Mauro, Grasiele, Jaqueline, Lucíola, Sílvia, Alessandra, Cíntia, Edson Emílio, Valmir, Nancy, Cláudia, Daniel, Kamila, Guilherme, Gabriel, Renata, Duncan, Fernando, Roger, Igor, Luís Otávio, Vinícius e Maria José

... porque "*O valor das coisas não está no tempo que elas duram, mas na intensidade com que acontecem. Por isso existem momentos inesquecíveis, coisas inexplicáveis e pessoas incomparáveis*" (Fernando Pessoa).

Agradeço também a minha orientadora, a todos do grupo PLN, aos membros da minha banca examinadora, à FACIN e ao CNPq.

RESUMO

Essa tese descreve um processo para extrair conceitos de textos em língua portuguesa. O processo proposto inicia com *corpora* de domínio linguisticamente anotados, e gera listas de conceitos dos domínios de cada *corpus*. Utiliza-se uma abordagem linguística, que baseia-se na identificação de sintagmas nominais e um conjunto de heurísticas que melhoram a qualidade da extração de candidatos a conceitos. Essa melhora é expressa por incrementos aproximadamente de 10% para mais de 60% nos valores de precisão e abrangência das listas de termos extraídas. Propõe-se um novo índice (*tf-dcf*) baseado na comparação com *corpora* contrastantes, para ordenar os termos candidatos a conceito extraídos de acordo com suas relevâncias para o *corpus* de domínio. Os resultados obtidos com esse novo índice são superiores aos resultados obtidos com índices propostos em trabalhos similares. Aplicam-se pontos de corte para identificar, dentre os termos candidatos classificados segundo sua relevância, quais serão considerados conceitos. O uso de uma abordagem híbrida para escolha de pontos de corte fornece valores adequados de medida F, trazendo qualidade ao processo de identificação de conceitos. Adicionalmente, propõem-se quatro aplicações para facilitar a compreensão, manipulação e visualização dos termos e conceitos extraídos. Essas aplicações tornam as contribuições dessa tese acessíveis a um maior número de pesquisadores e usuários da área de Processamento de Linguagem Natural. Todo o processo proposto é descrito em detalhe, e experimentos avaliam empiricamente cada passo. Além das contribuições científicas feitas com a proposta do processo, essa tese também apresenta listas de conceitos extraídos para cinco diferentes *corpora* de domínio, e o protótipo de uma ferramenta de software (EχATO$_{LP}$) que implementa todos os passos propostos.

Título: EXTRAÇÃO AUTOMÁTICA DE CONCEITOS A PARTIR DE TEXTOS EM LÍNGUA PORTUGUESA

Palavras-chave: Processamento de linguagem natural; Extração automática de termos; Recuperação de informação; Ontologias.

ABSTRACT

This thesis describes a process to extract concepts from texts in portuguese language. The proposed process starts with linguistic annotated *corpora* from specific domains, and it generates lists of concepts for each *corpus*. The proposal of a linguistic oriented extraction procedure based on noun phrase detection, and a set of heuristics to improve the overall quality of concept candidate extraction is made. The improvement in precision and recall of extracted term list is from approximatively from 10% to more more than 60%. A new index (*tf-dcf*) based on contrastive *corpora* is proposed to sort the concept candidate terms according to the their relevance to their respective domain. The precision results achieved by this new index are superior to to the results achieved by indices proposed in similar works. Cut-off points are proposed in order to identify, among extracted concept candidate terms sorted according to their relevance, which of them will be considered concepts. A hybrid approach to choose cut-off points delivers reasonable F-measure values, and it brings quality to the concept identification process. Additionally, four applications are proposed in order to facilitate the comprehension, handling, and visualization of extracted terms and concepts. Such applications enlarge this thesis contributions available to a broader community of researchers and users of Natural Language Processing area. The proposed process is described in detail, and experiments empirically evaluate each process step. Besides the scientific contribution made with the process proposal, this thesis also delivers extracted concept lists for five different domain *corpora*, and the prototype of a software tool (EχATO$_{LP}$) implementing all steps of the proposed process.

Title: AUTOMATIC EXTRACTION OF CONCEPTS FROM TEXTS IN PORTUGUESE LANGUAGE

Keywords: Natural language processing; Automatic term extraction; Information retrieval; Ontologies.

Sumário

Lista de Figuras

Lista de Tabelas

1. INTRODUÇÃO

Processamento de Linguagem Natural (PLN) é a área de pesquisa que estuda o desenvolvimento de programas de computador que analisam, reconhecem ou geram textos em linguagens humanas, ou linguagens naturais. PLN é uma área com grandes desafios, devido à rica ambiguidade da linguagem natural, sendo isso um dos fatores que torna PLN diferente do processamento das linguagens formais que são definidas evitando a ambiguidade.

Segundo DuRoss Liddy [58], um dos objetivos usuais em PLN é a recuperação de informação a partir de textos, pois textos são, segundo Maedche e Staab [126], a forma mais abundante de informação disponível. Dentre os tipos de recuperação de informação em textos, a busca de termos em *corpora* (plural de *corpus*) de domínio é uma das principais aplicações de PLN.

Um *corpus* de domínio é um conjunto de textos sobre um domínio específico que pode ser utilizado para caracterizar esse domínio. Portanto, detectar termos relevantes em um *corpus* é uma forma adequada de identificar termos relevantes para o domínio descrito por esse *corpus*. Mais que isso, segundo Perini [152], a identificação de termos através da observação de *corpora* permite a observação dos padrões de uso da linguagem livre de preconceitos.

A extração de termos de *corpora* de domínio possui diversas aplicações, como por exemplo, a categorização de textos [114, 59, 33], a identificação de termos para mecanismos de busca [167, 149, 7, 164], e a identificação de conceitos [198, 217, 38, 71]. Cada uma dessas aplicações possui suas especificidades, mas em todas elas existe a necessidade de identificar termos que sejam de alguma forma relevantes ao domínio.

1.1 Motivação

Uma das iniciativas mais ambiciosas e necessárias da computação é o estabelecimento da Web Semântica [19]. Essa iniciativa se propõe a organizar, semanticamente, o acesso à extraordinária quantidade de dados disponíveis com o advento e expansão da Internet, na qual a maior dificuldade não é encontrar o que se procura, mas reconhecer o que foi encontrado. Um dos caminhos para essa organização é a representação do conhecimento através de ontologias [85].

Ontologia, segundo Gruber [84], é uma forma de estruturar informações para representar conhecimento. No entanto, a representação desse conhecimento através de formalismos tratáveis por máquinas, como é o caso das ontologias, torna-se um grande desafio frente à quantidade enorme de dados textuais a estruturar. Portanto, é necessário automatizar o processo de construção de ontologias a partir de textos [44].

Retoma-se, então, a questão de extração de termos de *corpora*, visando a identificação dos conceitos que são, conforme será visto em detalhe nessa tese, os componentes fundamentais de ontologias [32, 56]. Nesse sentido, a motivação central do trabalho proposto é a dificuldade inerente à construção de ontologias, principalmente no que diz respeito à identificação dos elementos básicos, que são os conceitos e a sua expressão em termos linguísticos.

A esse fato soma-se que, em alguns cenários a utilização de ontologias em língua portuguesa se faz necessária. Alguns exemplos desses cenários são: a comunicação entre grupos de especialistas de domínio falantes do português, descrição de elementos culturais brasileiros, típicos de museu de cultura, ou em domínios que envolvem comunicação do especialista com o leigo como no caso de medicina e governo eletrônico.

1.2 Objetivo e Metodologia

O objetivo geral dessa tese é propor um processo para extrair automaticamente conceitos, ou seja, termos relevantes com valor conceitual [108], para um domínio caracterizado por um *corpus* em língua portuguesa, composto por textos representativos para este domínio.

Nesse sentido, para realizar o objetivo geral dessa tese é necessário alcançar os seguintes objetivos específicos:

- definir um método de extração de termos candidatos a conceitos a partir de um *corpus* anotado linguisticamente;
- definir um método de ordenar os termos extraídos segundo sua relevância;
- definir uma forma de identificar, dentre os termos extraídos, quais devem ser considerados conceitos do domínio;
- definir um conjunto de aplicações dos conceitos extraídos, que facilite a sua compreensão, manipulação e visualização.

1.2.1 Detalhamento do Processo Proposto

O processo desenvolvido nessa tese recebe como entrada um conjunto de *corpora* de domínio anotados linguisticamente e, após a aplicação do processo proposto, gera-se uma lista de conceitos e um conjunto de informações contextuais sobre esses conceitos. Em linhas gerais, esse processo pode ser dividido em quatro grandes etapas: extração de termos e contextos, ordenação de termos de acordo com sua relevância, identificação de conceitos e geração de recursos linguísticos (aplicações dos conceitos gerados). Essas quatro etapas e as informações sobre cada uma delas são descritas esquematicamente na Figura 1.1.

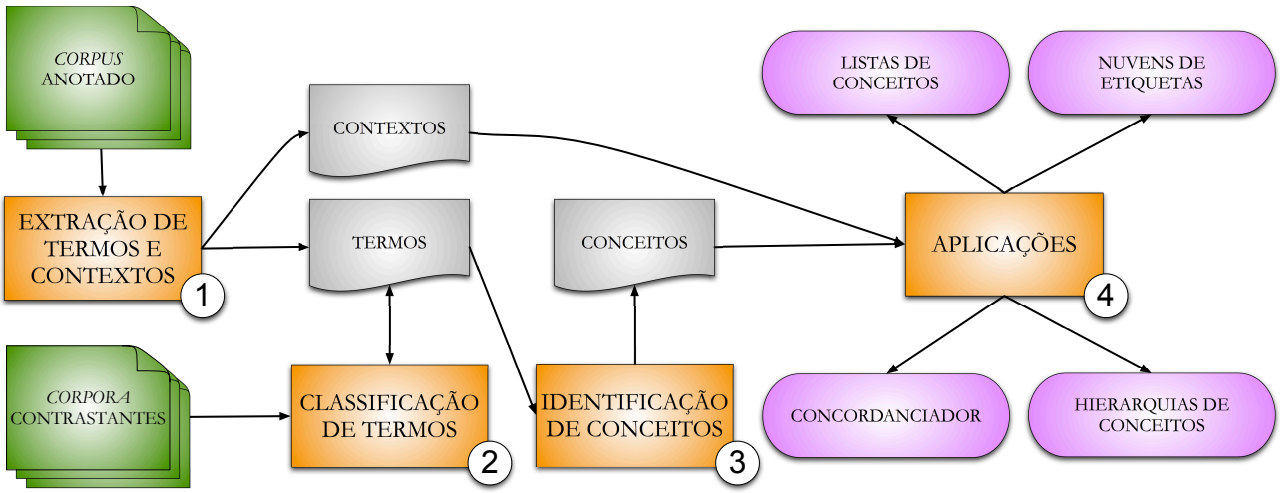


Figura 1.1: Processo geral de extração automática de conceitos.

A primeira etapa (1), extração de termos e contextos, descrita na Figura 1.1 corresponde a um processo linguístico onde recebe-se um *corpus* de domínio anotado e detectam-se os termos candidatos a conceitos desse domínio. Adicionalmente, informações referentes à forma como esses termos foram empregados no *corpus* (contextos), além dos números de ocorrências em que o termo foi encontrado em cada uma das suas situações de uso no *corpus*, são extraídos.

A segunda etapa (2), consiste em ordenar os termos segundo sua relevância através de um processo estatístico que leva em conta além do *corpus* de domínio, um conjunto de *corpora* usados como contraste ao domínio. Como resultado dessa etapa, cada termo recebe um valor

numérico que pode ser usado como índice de ordenação dos termos de acordo com sua relevância para o domínio.

A terceira etapa (3), identificação de conceitos, recebe a lista de termos ordenada segundo sua relevância e escolhe quantos destes termos devem ser considerados conceitos do domínio. Nesse sentido, essa etapa consiste em escolher e aplicar um ponto de corte à lista ordenada de termos.

A quarta etapa (4), consiste em utilizar os conceitos e seus respectivos contextos para gerar recursos linguísticos sofisticados. Dentre essas aplicações, apresenta-se nessa tese: listas de conceitos com informações contextuais; concordanciador para visualizar as frases onde cada um dos conceitos foi empregado no *corpus*; nuvens de conceitos (*tag clouds*) com informações visualmente estruturadas dos conceitos; e hierarquia de conceitos onde estruturam-se os conceitos segundo critérios linguísticos, ou seja, detectam-se relações taxonômicas entre os conceitos.

A metodologia empregada para definir e testar o processo descrito na Figura 1.1 consistiu da definição de cada uma das etapas, experimentação de técnicas pré-existentes, proposta de novas técnicas e, finalmente, a avaliação objetiva do resultado de cada etapa.

Especificamente, a avaliação de cada etapa foi feita com um *corpus* de domínio pre-existente (*corpus* de Pediatria) [49] para qual uma lista de termos relevantes (conceitos) foi disponibilizada [187]. Maiores detalhes sobre esse *corpus* e lista de referência serão apresentados na Seção 3.1.

Cabe salientar que além das avaliações feitas e apresentadas individualmente nos capítulos que descrevem, respectivamente, as etapas de extração, ordenação e identificação, outras avaliações externas a essa tese foram realizadas e são citadas na Seção 7.2 da conclusão.

Um ponto prático da metodologia de desenvolvimento dessa tese, é que todas as etapas do processo proposto foram implementadas em uma ferramenta de software, chamada EχATO$_{LP}$, Extrator Automático de Termos para Ontologias em Língua Portuguesa. Essa ferramenta [117, 118] se encontra ainda em estágio de protótipo, mas ela foi utilizada para a totalidade dos experimentos descritos nessa tese.

1.3 Partes desse Documento

Esse volume de tese é composto por cinco capítulos, além dessa introdução, uma conclusão e três anexos.

O Capítulo 2 apresenta conceitos básicos necessários à compreensão das contribuições feitas nos demais capítulos. Especificamente, apresenta-se um histórico da área de PLN situando o contexto desse trabalho e uma definição formal de ontologias. Nesse capítulo apresenta-se também o estado da arte em extração de termos que é o foco principal dos trabalhos dessa tese.

O Capítulo 3 contextualiza e apresenta o processo de extração de termos proposto. A contextualização é feita através de uma breve descrição dos *corpora* utilizados nessa tese, e o processo de anotação linguística utilizado, bem como um paralelo entre essa anotação e a gramática tradicional. O detalhamento do processo proposto descreve heurísticas com base linguística, que são aplicadas para qualificar o processo de extração. Em seguida, apresenta-se o produto final da etapa de extração de termos, sob a forma de uma lista de termos candidatos a conceitos com diversas informações contextuais de como os termos foram empregados nas suas diversas ocorrências.

O Capítulo 4 apresenta a proposta e os testes de um novo índice numérico para indicar a relevância de cada um dos termos extraídos no seu respectivo *corpus* de domínio. Ao contrário do processo de extração de termos, que possui forte base linguística, o processo de ordenação de termos descrito nesse capítulo possui forte base estatística. O índice de relevância proposto é comparado a outras iniciativas semelhantes presentes na literatura e o produto final deste processo é a lista de termos extraídos, porém, ordenados segundo o índice de relevância

O Capítulo 5 apresenta o processo de identificação, ou escolha, dentre os termos extraídos, de quais deles serão considerados conceitos. Nesse capítulo, analisa-se pontos de corte aplicados aos termos ordenados, para que seja possível identificar quais termos serão considerados conceitos.

O Capítulo 6 descreve quatro aplicações dos conceitos extraídos do *corpus* de domínio. Essas aplicações são processos automáticos que partem dos conceitos, e seus respectivos contextos, para gerar recursos linguísticos sobre o domínio. Especificamente, as aplicações exemplificadas nessa tese são: listas de termos e conceitos; concordanciador; nuvens de conceitos; e hierarquias de conceitos. Cabe salientar que essas aplicações não são as únicas possíveis, nem a sua formalização prática nessa tese pretende ir além de exemplos de utilização dos conceitos extraídos.

A conclusão dessa tese resume o trabalho desenvolvido salientando as contribuições científicas e tecnológicas obtidas. Igualmente, a conclusão cita os recursos linguísticos criados durante esse doutorado e sugere trabalhos futuros a essa tese. Os anexos dessa tese apresentam listas de termos de referência (anexo A), listas de conceitos extraídos dos *corpora* utilizados nessa tese (anexo B) e uma lista de etiquetas semânticas utilizadas pelo *parser* (anexo C).

2. REFERENCIAL TEÓRICO E ESTADO DA ARTE

Esse capítulo situa as contribuições científicas dessa tese dentro da área de Processamento de Linguagem Natural (PLN). Para tanto, define-se genericamente essa área através de um breve histórico (Seção 2.1). Após, apresenta-se uma definição formal de ontologias e hierarquias de conceitos para permitir localizar claramente onde se insere o objetivo central dessa tese, que é a extração automática de conceitos (Seção 2.2). Por fim, apresenta-se os problemas específicos de criação de ontologias, com ênfase na extração de termos e suas métricas usuais de qualidade (Seção 2.3).

2.1 PLN, Processamento de *Corpus* e Web Semântica

Historicamente[1], a área de PLN começou com tentativas de tradução automática na segunda metade da década de 1940 [199]. Esses trabalhos iniciais estavam relacionados com esforços prévios de quebra de códigos durante a Segunda Guerra Mundial. De um ponto de vista teórico, esses trabalhos iniciais em tradução automática estavam baseados na criptografia e teoria da informação [176]. Em 1957, Chomsky desenvolveu trabalhos relevantes sobre o tema. Um trabalho particularmente relevante dessa época é o livro *Syntactic Structures* [39] que introduziu a gramática gerativa. A partir desse trabalho, ficou mais claro como a área de linguística poderia auxiliar a área de tradução automática.

Nessa época houve também a inclusão de outras aplicações de PLN, especialmente a do reconhecimento da fala (*speech recognition*). Com isso, houve a primeira grande divergência, que, de certa forma, permanece até hoje, pois parte da comunidade optou pelo uso de linguística teórica e parte optou por métodos estatísticos. Infelizmente, cada uma dessas partes rechaçava os métodos da outra parte, prejudicando a integração dessas duas abordagens. Esse período marca também o advento da Teoria Sintática da Linguagem [40] e dos Algoritmos de *Parsing* [3]. Esses avanços foram muito importantes para a área, ainda que na época tenham sido recebidos com um entusiasmo excessivo, gerando a expectativa de que, em poucos anos, tradutores automáticos perfeitos estariam disponíveis. Essa expectativa se mostrou indevida tanto pelos conhecimentos linguísticos e computacionais da época, quanto por uma impossibilidade teórica da tarefa de tradução automática perfeita [13].

Consequência disso ou não, em 1966 o comitê assessor para processamento automático da língua (ALPAC) da Academia Americana de Ciência recomendou que a área de tradução automática não recebesse mais financiamento governamental, pois a tradução automática estava muito aquém dos conhecimentos científicos da época. Em contraste com essa decisão, vários avanços teóricos e práticos foram feitos nos anos seguintes. Entre eles, pode ser citado o trabalho teórico de Chomsky que introduziu o modelo computacional de competência linguística [40], que resultou nas gramáticas gerativas transformacionais. Diversos trabalhos subsequentes [92, 93] tentaram aproximar esses conceitos de modelos computacionalmente tratáveis.

[1] A descrição histórica da área de PLN apresentada nessa seção é um resumo do capítulo "Processamento de linguagem natural e o tratamento computacional de linguagens científicas" originalmente publicado em 2010 no livro "Linguagens Especializadas em *Corpora* - modos de dizer e interface de pesquisa" [121].

A partir desse período, houve uma multiplicação dos estudos sobre PLN com o estabelecimento de diversas subáreas que vêm sendo pesquisadas até hoje. Essas áreas se dedicam a assuntos tão variados quanto categorização de textos e extração de informações, passando pelos tradicionais temas de tradução automática e sistemas de diálogo. Os trabalhos desenvolvidos nesse período podem, segundo Jurafsky e Martin [94], ser agrupados em quatro grupos de acordo com os paradigmas utilizados: os métodos estocásticos, os métodos baseados em lógica, os métodos de entendimento de linguagem natural, e os métodos de modelagem de discurso. Os trabalhos do grupo de métodos estocásticos são baseados em abordagens estatísticas e frequentemente utilizam formalismos com os modelos ocultos de Markov (HMM - *Hidden Markov Models*). Esses métodos estão na base de diversos trabalhos de reconhecimento e síntese de fala [161]. Esses trabalhos estão na origem dos atuais trabalhos em que métodos estatísticos são empregados para diversas aplicações de PLN [217, 189, 56].

Os trabalhos baseados em lógica começaram com *Q-systems* e gramáticas metamórficas [46] que foram os precursores da linguagem Prolog [47] e das gramáticas de cláusulas definidas (*DCG - Definite Clause Grammar*) [151]. Dessa mesma época datam também as iniciativas de gramáticas funcionais na sua versão inicial [98] e na versão léxica [29].

Os trabalhos baseados em entendimento da linguagem natural seguiram na vertente do entendimento do discurso. De um ponto de vista teórico são típicos desses trabalhos aqueles sobre Gramáticas de Caso [66], Redes Semânticas [160], Teoria de Dependência Conceitual [173], Redes de Transição Aumentada [209] e Semântica de Preferência [205]. De um ponto de vista puramente prático, esse período viu o aparecimento de diversos programas que faziam uso intensivo de PLN. Esse foi o caso dos sistemas de diálogo ELIZA [200] e PARRY [45], mas também os sistemas de reconhecimento de fala SHRDLU [206], LUNAR [210], LIFER/LADDER [89] e PLANES [197].

Em seguida, as iniciativas centradas na modelagem do discurso focaram suas atenções em questões semânticas. Trabalhos significativos dessa época como o trabalho de Grosz [82] visavam diálogos funcionais (diálogos que especificam uma tarefa a ser executada). Os trabalhos subsequentes de Grosz e Sidner [83] definem uma teoria de partição do discurso baseado em relações entre a estrutura da tarefa a executar e a estrutura do diálogo que descreve essa tarefa. Nessa mesma época foi desenvolvida por Mann e Thompson a Teoria de Estrutura Retórica [129] que associa uma estrutura hierárquica para o discurso com o intuito de geração automática de texto. Outros trabalhos desse período também foram dedicados à geração de linguagem natural, como é o caso dos geradores de resposta TEXT [136] e MUMMBLE [135], que usam predicados retóricos para produzir descrições declarativas na forma de parágrafos.

2.1.1 Processamento de *Corpus*

Desde o início da década de 1990, o crescimento da internet e a profusão de textos disponíveis direcionaram os esforços do PLN para o tratamento de textos, mais do que para o discurso falado. Nessa época iniciaram as pesquisas sobre *corpora* anotados sintaticamente, ou seja, conjuntos de textos sobre um domínio de conhecimento, em que cada uma das suas palavras são identificadas segundo sua função sintática. Vários desses trabalhos foram desenvolvidos para a língua inglesa, utilizando três corpora bastante populares: *Brown corpus* [106], *Lancaster-Olso-Bergen corpus* [75] e *Penn Treebank* [132]. Por ocasião do uso de *corpora*, diversos trabalhos de pesquisa baseados em conceitos linguísticos utilizados em conjunto com abordagens estatísticas, possibilitaram resultados práticos mais robustos [131]. Essa reconciliação entre métodos linguísticos e estatísticos, que se percebe atualmente, desfaz a divisão de abordagens feita na área desde do final da década de 1950.

A grande quantidade de informação a ser tratada que impulsionou a reconciliação dos métodos estatísticos e linguísticos teve desde a virada do século uma outra consequência in-

teressante com a incorporação de técnicas de aprendizado de máquina [207]. As técnicas de aprendizado de máquina são particularmente aplicáveis no contexto de conjuntos de dados humanamente intratáveis, mas, dos quais se pode inferir padrões e, consequentemente, informação. Naturalmente, os últimos anos têm testemunhado uma convergência das técnicas de PLN baseadas em corpus com técnicas de aprendizagem de máquina, ou mais especificamente, técnicas de mineração de dados. Esse aumento significativo das ferramentas à disposição dos pesquisadores de PLN permitiu também um aumento significativo nas ambições da área. Retomomaram-se seriamente os trabalhos de tradução automática, apesar da consciência de ser inatingível uma tradução perfeita. Uma quantidade muito grande de tradutores automáticos está disponível na internet como é o caso dos *sites* especializados como *babelfish* [8], mas também de serviços de tradução embutidos como os disponíveis automaticamente pelo gigante de pesquisas Google [80].

2.1.2 Web Semântica

Igualmente, a evolução de outras áreas da computação, como é o caso da computação pervasiva (recursos computacionais presentes em atividades cotidianas) ou aplicações web, abriu espaço para novos sistemas de reconhecimento e geração automática da linguagem. Alguns sistemas recentes utilizam PLN com o objetivo de responder perguntas de forma clara e direta através de conhecimento semântico. Exemplos práticos desse tipo de aplicação, disponíveis na internet, são os sistemas Ask [6], Lexxe [115] e Hakia [86]. Outro exemplo desenvolvido com o mesmo princípio é o sistema True Knowledge [190], que, além de direcionar a vários *links* relacionados com perguntas simples feitas pelo usuário, também insere a resposta direta a pergunta do usuário. Esse sistema permite ainda que os usuários acrescentem informações, tornando-o cada vez mais completo e preciso.

Porém, talvez o objetivo mais ambicioso do processamento de linguagem natural resida atualmente na construção da web semântica que pretende estabelecer uma ponte entre o enorme volume de dados disponível na Internet e as demandas de informação e conhecimento de seus milhões de usuários [19]. A web semântica é uma iniciativa que busca identificar e representar o significado de páginas na web de forma que tanto pessoas como máquinas possam identificá-los. Nesse sentido, o grande desafio é a representação do conhecimento em um formato adequado, que nesse contexto é feito através de Ontologias.

2.2 Ontologias

Segundo Gruber [84], "Ontologia é uma especificação explícita de uma conceitualização". Dessa forma, ontologias podem ser consideradas representações formais (um conjunto concreto de especificações) de um modelo de domínio que exista de forma abstrata.

Geralmente, uma ontologia é entendida como um conjunto de conceitos organizados hierarquicamente, um conjunto de relações e um conjunto de atributos. Essa seção apresenta uma definição formal de ontologias, mas o leitor interessado pode encontrar um extenso material nos trabalhos de Ehrig [60], Biemann [21] e Buitelaar *et al.* [32]. Cabe salientar que existem diversas definições formais de ontologias na literatura [22, 182, 61, 175], e qualquer uma delas se prestaria aos propósitos dessa tese. Apesar disso, no contexto dessa tese, é apresentada a definição proposta por Cimiano [42], por ser uma referência formal de ampla difusão na área empregada por diversos autores da comunidade nacional.

2.2.1 Definição Formal de Ontologias

De um ponto de vista formal [42] uma ontologia é uma estrutura:

$$\mathcal{O} := (C, \leqslant_C, R, \sigma_R, \leqslant_R, A, \sigma_A, T)$$

Composta de:

- Quatro conjuntos disjuntos:
  - C - identificadores de conceitos;
  - R - identificadores de relação;
  - A - identificadores de atributos; e
  - T - tipos de dados (inteiros, *strings*, *etc.*);
- Um semireticulado superior $\leqslant_C$ definido sobre os elementos de C (conceitos) chamado de hierarquia de conceitos ou taxonomia, que possui:
  - um supremo $raiz_C$;
  - uma relação de subconceito e superconceito entre dois conceitos c_1 e c_2 pertencentes a C que diz que c_1 é um subconceito de c_2, caso $c_1 \leqslant_C c_2$, e, por simetria, c_2 é um superconceito de c_1;
  - adicionalmente caso não exista um conceito c_3 tal que $c_1 \leqslant_C c_3 \leqslant_C c_2$, diz-se que c_1 é um subconceito direto de c_2 e, analogamente, c_2 é um superconceito direto de c_1, essas relações denota-se como $c_1 \prec c_2$;
  - as relações simétricas de subconceito e superconceito são relações taxonômicas, ou seja, quando c_2 é um superconceito de c_1 pode se dizer que c_1 tem uma relação "é_um" (em inglês "*is_a*") com c_2;
- Uma função σ_R: $R \rightarrow C^+$ que estabelece relações entre conceitos, chamada assinatura de relação. Essas funções definem uma relação do conjunto R e dois conjuntos de conceitos de C, respectivamente:
  - domínio (*domain*) que diz quais conceitos podem originar a relação; e
  - intervalo (*range*) que diz que conceitos podem ser destino da relação;
- Uma ordem parcial $\leqslant_R$ sobre R que estabelece uma ordem de precedência de certas relações sobre outras, chamada hierarquia de relação, que de forma análoga à hierarquia de conceitos define:
  - os conceitos de subrelação e superelação que dizem que duas relações r_1 e r_2 pertencentes a R onde $r_1 \leqslant_R r_2$ são: r_1 uma subrelação de r_2 e, analogamente, r_2 uma superelação de r_1; e
  - os conceitos de subrelação e superelação diretas quando não existe uma relação r_3 tal que $r_1 \leqslant_R r_3 \leqslant_R r_2$, que denota-se $r_1 \prec r_2$;
- Uma função σ_A: $A \rightarrow C \times T$, similar à função σ_R, mas que relaciona atributos ao invés de conceitos, chamada assinatura de atributos.

2.2.2 Sistema de Axiomas, Base de Conhecimentos e Extensões

Usualmente, define-se ao mesmo tempo que uma ontologia $\mathcal{O}$ um conjunto de axiomas que permite estabelecer as propriedades necessárias entre os conceitos, as relações e os atributos dessa ontologia.

De um ponto de vista formal, um sistema de axiomas $\mathcal{S}$ de uma ontologia $\mathcal{O}$ é definido pela tripla:

$$\mathcal{S} := (AS, \alpha, \mathcal{L})$$

Composta de:

- uma linguagem lógica $\mathcal{L}$;
- um conjunto de axiomas AS que pode fazer referência a conceitos, relações e atributos;
- um mapeamento $\alpha : AS \rightarrow AS_{\mathcal{L}}$

Uma vez definida uma ontologia $\mathcal{O}$ e um sistema de axiomas $\mathcal{S}$, a definição geral de uma ontologia é completada através da definição de instâncias para os conceitos, as relações e os atributos.

De um ponto de vista formal, isto é feito através da definição de uma base de conhecimento:

$$\mathcal{KB} := (I, \iota_C, \iota_R, \iota_A)$$

Composta de:

- um conjunto I de identificadores de instâncias, ou simplesmente instâncias;
- uma função $\iota_C : C \rightarrow \mathfrak{P}(I)$, chamada instanciação de conceitos, que define para cada conceito $c \in C$ qualquer subconjunto[2] de I;
- uma função $\iota_R : R \rightarrow \mathfrak{P}(I^+)$, chamada instanciação de relações, que define para cada relação $r \in R$ qualquer tupla[3] contendo elementos de I;
- uma função $\iota_A : A \rightarrow (I \cup_{t \in T} [t])^+$, chamada instanciação de atributos, que define para cada atributo $a \in A$ um par com uma instância de I e um elemento do seu tipo de dados t.

Aplicando-se a uma ontologia $\mathcal{O}$, instanciada por uma base de conhecimentos $\mathcal{KB}$, e levando-se em consideração um sistema de axiomas $\mathcal{S}$, é possível popular esse conjunto $\{\mathcal{O}, \mathcal{KB}, \mathcal{S}\}$ com instanciações adicionais decorrentes do semireticulado $\leqslant_C$, da ordem parcial $\leqslant_R$ e da aplicação dos axiomas. Essas extensões são definidas como $[c]$, para conceitos $c \in C$, $[r]$, para relações $r \in R$ e $[a]$, para atributos $a \in A$.

2.2.3 Hierarquia de Conceitos

Frente à complexidade de uma ontologia completa, com sua estrutura básica ($\mathcal{O}$), seu sistema de axiomas ($\mathcal{S}$), sua base de conhecimentos ($\mathcal{KB}$) e suas extensões, muitos trabalhos [172, 28, 43, 217, 56, 158] estão baseados em construir apenas uma hierarquia de conceitos. Formalmente, uma hierarquia de conceitos é definida por um conjunto de conceitos e um semirreticulado superior, ou seja:

$$\mathcal{H} := (C, \leqslant_C)$$

[2] A notação $\mathfrak{P}(I)$ representa o conjunto com todos os subconjuntos possíveis do conjunto I.

[3] A notação I^+ representa todos os conjuntos possíveis de tuplas formadas por elementos de I.

Obter um conjunto qualificado de conceitos torna possível construir melhores hierarquias, que, por sua vez, é a estrutura base para definir uma ontologia ($\mathcal{O}$). O propósito dessa tese é a extração automática e qualificada de conceitos, ou seja, a definição qualificada do conjunto C para um *corpus* de domínio. Porém, como será visto posteriormente (Seção 6.4), desenvolve-se também uma hierarquia de conceitos, ou seja, infere-se as relações expressas pelo semirreticulado $\leqslant_C$. Cabe salientar que o processo de construção de hierarquia de conceitos apresentado como exemplo de aplicação na Seção 6.4, é apenas um exercício de estruturação sem maiores ambições científicas, pois o foco científico dessa tese é a extração de conceitos.

2.2.4 Construção Automática de Ontologias

Para construção de ontologias é necessário realizar um processo bastante complexo e trabalhoso, que pode ser feito manualmente por um engenheiro de ontologias com o auxílio de um ou vários especialistas de um determinado domínio. No entanto, essa construção manual demanda muito tempo e trabalho de todos os envolvidos. Uma alternativa é automatização da construção de ontologias, porém essa tarefa representa grandes desafios computacionais.

Diversas abordagens de aprendizagem, técnicas e ferramentas para a construção de ontologias podem ser encontradas atualmente, pois dada a complexidade do processo, é difícil imaginar a construção de uma ontologia sem o auxílio de ferramentas computacionais. Curiosamente, pela mesma razão, a alta complexidade, ainda não existem sistemas efetivos capazes de construir uma ontologia completa de forma totalmente automática.

A construção de ontologia feita através de métodos automáticos ou semi-automáticos de extração de conhecimento é denominada "Aprendizagem de Ontologia" (*Ontology Learning*). Esse termo foi introduzido originalmente por Madche e Staab em 2001 [126] que inicialmente incorporaram o uso de técnicas oriúndas da área de aprendizagem de máquina [140]. Apesar disto, o termo "Apredizagem de Ontologia" não se restringe apenas a técnicas baseadas em aprendizagem de máquina, podendo envolver diversas outras áreas do conhecimento como linguística computacional e recuperação de informações.

Os esforços semiautomáticos são baseados na utilização de ferramentas, *e.g.*, software de edição, que permitam organizar ontologias que serão projetadas por um usuário que conheça o domínio a ser descrito. Dentre essas ferramentas, provavelmente a mais popular é o Protégé [77, 157], que permite ao usuário construir e manipular ontologias. As funcionalidades básicas desta ferramenta incluem algumas verificações e visualizações automáticas. Porém, o Protégé oferece a possibilidade de adicionar *plugins* capazes de realizar operações sobre ontologias, *e.g.*, OntoLP [165], um extrator de termos de fontes textuais (textos). Protégé permite armazenar ontologias modeladas segundo dois protocolos: OKBC - *Open Knowledge Base Connectivity* [37] e OWL - *Ontology Web Language* [133, 185].

Outra ferramenta semi-automática de construção de ontologias é o OntoGen [70, 144] que combina técnicas de mineração de textos com uma interface de utilização, que facilita a escolha dos conceitos e relações. Dessa forma, o OntoGen, parte de um *corpus* e oferece ao usuário conjuntos de termos candidatos a conceitos, e cabe ao usuário estabelecer manualmente a hierarquia entre os conceitos, bem como as relações entre eles. Nesse sentido, OntoGen, assim como alguns *plugins* do Protégé, é uma ferramenta para edição de ontologias que possui um processo de extração de termos a partir de textos.

Na verdade, muitas ferramentas buscam em fontes textuais (textos) o conhecimento a ser armazenado em uma ontologia. Segundo Maedche e Staab [126], a busca de informações em textos se justifica, pois a grande maioria do conhecimento disponível encontra-se em fontes textuais. Nesse sentido, o trabalho desenvolvido nessa tese está baseado na extração de informações contidas em textos.

O primeiro problema para gerar ontologias a partir de textos é identificar quais tarefas são

necessárias para a construção efetiva de uma ontologia. Segundo Buitelaar *et al.* [32], esse processo divide-se em cinco[4] etapas básicas: extração de termos candidatos a conceitos de um domínio; determinação de sinônimos entre os termos candidatos e escolha dos conceitos; identificação da relação hierárquica entre os conceitos; identificação de relações entre os conceitos; e identificação de instâncias (população da ontologia). A aprendizagem de ontologias pode ser representada em camadas de acordo com a Figura 2.1.

Figura 2.1: Etapas de aprendizagem de ontologias.

Logicamente, os passos descritos na Figura 2.1 devem ser executados sequencialmente, sendo a extração de termos candidatos a conceitos a primeira e mais importante tarefa, pois da qualidade dos resultados dessa etapa depende a qualidade dos resultados de todas as demais etapas. Note-se que essa afirmação não significa que as outras etapas sejam mais simples, ou que não seja necessário preocupar-se com a eficiência de cada uma delas. A qualidade da ontologia é dependente de todas as etapas, porém, caso a extração de termos candidatos seja deficiente, o resultado de todas as demais etapas não poderá compensar essa deficiência. Essa opinião é compartilhada por diversos autores da área [32, 153, 134, 71].

2.3 Extração de Termos e Conceitos

A importância da extração de termos para a construção automática de ontologias é clara [170, 180, 203, 162, 186, 213]. No entanto, em mecanismos de busca e mineração de textos em geral a importância de uma correta extração de termos também vem sendo tema de pesquisas há mais de quatro décadas [183, 167, 95, 111, 212, 26, 169, 2].

[4]Segundo diversos autores [42, 43, 126], e certas vezes até em publicações de um mesmo autor, é possível encontrar diversas variações na definição das etapas de construção automática de ontologias. A versão considerada nessa tese é uma ligeira adaptação realizada a partir da publicação de Buitelaar *et al.* [32] que reflete a organização do processo proposto nessa tese.

2.3.1 Abordagens de Extração de Termos

Uma das primeiras observações relevantes, no que diz respeito à extração de termos, é o fato de que existem diferenças entre extração de termos simples, ou seja, termos com uma única palavra, e extração de termos compostos. Um termo composto é um conjunto de duas ou mais palavras que possui um significado comum, e que por sua natureza são mais difíceis de detectar do que termos simples (uma única palavra).

Historicamente, os trabalhos de extração iniciaram, e ainda têm uma importante vertente, com contabilizações do número de termos simples extraídos [183, 167, 170]. Em seguida, por volta da década de 80, um grande número de trabalhos centrou seu interesse na extração de termos compostos [180, 95, 10, 124, 186]. De qualquer maneira, devido à importância da qualidade na extração de termos, muitos trabalhos científicos dedicam-se a aperfeiçoar esse processo, e como é comum em PLN, as abordagens para a extração de termos se dividem em abordagens estatísticas e linguísticas.

As abordagens estatísticas de extração de termos têm no extrator NSP [11] sua ferramenta mais popular. Essa ferramenta alia simplicidade da busca de termos por combinação de palavras adjacentes com um método de descarte de termos através de *stop list*, ou seja, listas de termos comuns que não possuem grande valor terminológico. Na verdade, a eficiência da abordagem utilizada pela ferramenta NSP depende muito da escolha de termos a incluir na *stop list*.

Outras abordagens estatísticas, como a ferramenta BootCat [14], oferecem recursos mais sofisticados, principalmente, no que concerne a extração de termos compostos. Apesar disso, a abordagem utilizada pela ferramenta BootCat também depende da especificação de *stop lists*, como toda abordagem estatística.

Ainda se inclui dentre as abordagens estatísticas de extração de termos as iniciativas que tentam calcular índices, como os populares *tf-idf* [131, 111] e *loglikelihood* [146, 130], que sejam mais efetivos do que a simples frequência de ocorrência dos termos. Porém, segundo Wermter e Udo [203], não é possível, sem o uso de informações linguísticas, obter melhores resultados do que a simples frequência absoluta de termos. Essa conclusão, de certa forma, explica o sucesso de uma abordagem simplista como a implementada na ferramenta NSP.

Por outro lado, as abordagens baseadas em informações linguísticas tendem a oferecer bons resultados na extração de termos. Ainda que tenham como desvantagem o fato de precisarem de ferramentas de anotação linguística eficazes, e que sejam, quase sempre, específicas para textos em um único idioma.

Dentre as abordagens linguísticas, alguns métodos têm apresentado resultados bastante precisos, como é o caso das abordagens baseadas no método *C-value* e sua versão estendida *NC-value* [73]. Esse método baseia-se na observação de padrões sintáticos para detectar, com grande sucesso, termos compostos aninhados, que são particularmente frequentes em inglês[5]. Infelizmente, esse método não parece ter a mesma eficiência quando portado para outras línguas, tipicamente línguas latinas [23].

Um exemplo recente de abordagem linguística para extração de termos é o trabalho de Bui e Sloot [31], onde através de padrões sintáticos buscam-se termos específicos de eventos biológicos[6]. A abordagem desse artigo não procura termos gerais, mas sim padrões específicos que possuam uma semântica clara e um conjunto de termos conhecidos previamente (por exemplo, nomes de proteínas). Abordagens como essa são facilitadas pela especificidade, e chegam a taxas de acerto com valores médios de precisão em torno de 50%.

[5]Termos compostos aninhados não são uma exclusividade da língua inglesa. No entanto, seu uso em inglês apresenta uma dificuldade adicional devido à possibilidade de composição de diversos substantivos como na expressão "*movie actor studio*" (estúdio de atores de filme), onde três substantivos são utilizados para descrever, além do termo geral, dois termos aninhados: "*movie actor*" (ator de filme) e "*movie*" (filme).

[6]Eventos biológicos são termos específicos da área de biologia que descrevem um momento de interesse, por exemplo, a interação entre duas proteínas.

De maneira genérica, é possível afirmar que a extração de termos é uma tarefa que, apesar de ser objeto de estudo há um longo tempo, ainda apresenta desafios consideráveis. Uma das formas mais eficazes de extração de termos é realizar a anotação linguística de *corpora* e em seguida extrair termos segundo uma análise estatística. O processo de extração proposto no decorrer dessa tese se enquadra nesse tipo de abordagem híbrida. Alguns exemplos similares são os trabalhos de Drouin [55], Teixeira *et al.* [186], e Bonin *et al.* [23].

2.3.2 Identificação de Conceitos

Um aspecto importante para a recuperação de informações textuais é o passo posterior à extração de termos, que consiste em escolher dentre os termos extraídos aqueles que são portadores de valor conceitual, e não apenas terminológico [108]. Uma distinção importante, segundo Petasis *et al.* [153], é a definição de conceitos, que se presta a controvérsias. No entanto, um bom número de autores [32, 193, 42, 153] parece concordar que um conceito é uma generalização associada a uma ideia, podendo ter várias manifestações textuais.

No processo proposto nessa tese, alguns dos termos relevantes extraídos e identificados como conceitos, poderiam ser melhor classificados como instâncias A subclassificação de um termo relevante como conceito ou instância é um processo de grande complexidade. Para atacar esse problema, faz-se uso de técnicas de análise sintática, desambiguação, coreferência, *etc.* dentro de uma área denominada população de ontologias, que foge ao escopo dessa tese. O leitor interessado pode achar grande material sobre o assunto em publicações específicas [42, 128, 154, 104, 63].

Apesar de não estabelecer uma distinção teórica entre conceitos e instâncias, a grande vantagem da abordagem proposta, reside no fato de que o processo, baseado na estimativa da relevância dos termos, permite automatizar a identificação dos principais conceitos de um domínio sem maiores intervenções humanas. Dessa forma, o esforço de extração de conceitos de um *corpus* de domínio feito nessa tese se alinha com outros trabalhos científicos que partem de um processo básico de extração de termos, e, em seguida, se empenham em estimar a relevância dos termos extraídos a fim de identificar os conceitos. Alguns exemplos desse tipo de trabalho, são os esforços de Pantel e Lin [146], Chung [41], Milios *et al.* [138], Drouin [55], Park *et al.* [148], Kim *et al.* [102].

2.3.3 Medidas de Avaliação

Uma questão importante que se coloca nessa área de extração de informação é que todas as iniciativas de identificação de conceitos são, pela natureza do objetivo, obrigatoriamente empíricas [95]. Assim sendo, uma das questões fundamentais de pesquisa é definir uma forma de verificar a qualidade do processo proposto.

Nessa tese optou-se por utilizar, quando disponível, uma lista de termos relevantes do domínio previamente estabelecida como referência para o sucesso do processo (*gold standard*). Dessa forma, é possível comparar listas de termos resultantes da extração segundo diversas abordagens com as listas de referência.

Com o propósito de comparar listas de termos ao longo dessa tese, definem-se três índices oriúndos da área de teoria da informação e de uso frequente na área de recuperação de informação. Esses índices são as tradicionais medidas de precisão (em inglês: *precision* - P), abrangência (em inglês: *recall* - R) e medida F (em inglês: *f-measure* - F) [192].

Essas medidas são utilizadas para comparar dois conjuntos, por exemplo, duas listas de termos. Um desses conjuntos, denominado $\mathcal{LR}$ (lista de referência), contém os termos de referência considerados corretos para o propósito, ou seja, o alvo da identificação de conceitos. O outro conjunto, denominado $\mathcal{LE}$ (lista extraída), contém os termos a comparar com a referência,

ou seja, os termos extraídos que por alguma métrica foram escolhidos pela aplicação do ponto de corte.

A precisão (P) é dada pela equação abaixo que expressa a razão entre o número de termos da lista de referência que foram extraídos e considerados (tamanho da intersecção entre os conjuntos $\mathcal{LR}$ e $\mathcal{LE}$) e o tamanho da lista de termos extraídos e considerados ($\mid \mathcal{LE} \mid$). Dessa forma, a precisão (em inglês: *precision*) expressa o percentual de termos corretamente extraídos, ou seja, o percentual dos termos localizados como corretos, quantos são efetivamente corretos.

$$P = \frac{\mid \mathcal{LR} \cap \mathcal{LE} \mid}{\mid \mathcal{LE} \mid} \tag{2.1}$$

A abrangência (R) é semelhante à precisão, porém expressa a razão entre o número de termos da lista de extraídos e considerados ($\mathcal{LE}$) presentes na lista de referência ($\mathcal{LR}$) e o tamanho da lista de referência ($\mid \mathcal{LR} \mid$). Dessa forma, a abrangência (em inglês: *recall*) expressa o percentual de termos da lista de referência coberta pela extração de termos feita.

$$R = \frac{\mid \mathcal{LR} \cap \mathcal{LE} \mid}{\mid \mathcal{LR} \mid} \tag{2.2}$$

A medida F (F) expressa o equilíbrio entre os valores de precisão e abrangência. A sua expressão numérica é a média harmônica entre os valores de P e R. Os valores da medida F (em inglês: *f-measure*) são valores situados entre P e R, e quanto maior for a diferença entre esses valores, mais próxima a medida F será do menor valor entre eles.

$$F = \frac{2 \times P \times R}{P + R} \tag{2.3}$$

O uso desses índices de qualidade é bastante difundido em diversas áreas, *e.g.* [141, 25, 188, 65]. Na área de PLN, e em especial nas tarefas de extração de termos, diversos trabalhos justificam a sua validade baseados em seus resultados numéricos, *e.g.*, [91, 12, 123].

3. EXTRAÇÃO DE TERMOS

A primeira etapa do trabalho desenvolvido no contexto dessa tese consiste em extrair um conjunto de termos sobre um *corpus* de domínio específico. O ponto de entrada nessa tarefa é um *corpus* linguisticamente anotado, e como saída gera-se uma lista com todos os termos empregados no *corpus*, bem como uma série de informações sobre o contexto no qual cada termo foi empregado.

Dessa forma, nesse capítulo faz-se uma breve descrição de um conjunto de *corpora* que serão utilizados como exemplos ao longo dessa tese (Seção 3.1). Após, descreve-se informações sobre a anotação realizada, bem como, noções básicas de gramática necessárias à compreensão da tarefa de extração (Seção 3.2). Em seguida, são propostas heurísticas de ajuste, descarte e inclusão aplicadas aos termos linguisticamente anotados, ou seja, a contribuição central desse capítulo (Seção 3.3). Após, são avaliadas as heurísticas propostas através de uma série de experimentos práticos que comparam as listas extraídas às listas de referência (Seção 3.4). Finalmente, sumariza-se na Seção 3.5 o processo de extração exemplificando todas as informações extraídas.

Os experimentos práticos relativos às heurísticas apresentadas nesse capítulo fazem parte de uma publicação recentemente aceita na conferência *PROPOR 2012* que será realizada em Abril de 2012 em Coimbra, Portugal [122].

3.1 *Corpora* de Domínio Utilizados nessa Tese

O objetivo central dessa tese é a extração automática de conceitos a partir de um um *corpus* de domínio específico. Logicamente, para que se possa alcançar esse objetivo é necessário ter disponível um certo número de *corpora* para que o procedimento possa ser experimentado.

Formalmente, *corpora* (o plural de *corpus*) são conjuntos de dados linguísticos pertencentes ao uso oral ou escrito de uma linguagem devidamente sistematizado de acordo com critérios suficientemente abrangentes para ser considerados representativos do uso linguístico [171].

Segundo Perini [152], "O uso de *corpora* no processo científico se torna relevante por causa de sua imparcialidade e indicação confiável de frequências das formas, posto que eles representam a realidade da linguagem sem preconceitos teóricos". Apesar de longo e laborioso, o processo de construção de *corpus* é válido, pois, uma vez criado, ele pode ser utilizado para diferentes aplicações como extração automática de termos, análises de estilo de escrita, construção de glossários, *etc.*

Muitos trabalhos na área de PLN são baseados no uso de *corpus* de domínio. Um *corpus* de domínio é um conjunto de textos que pode ser considerado suficientemente representativo de uma área específica (o domínio). Exemplos de trabalhos científicos baseados em manipulação de *corpus* são muito abundantes [172, 99, 27, 143, 100, 124, 158]. Isto se explica por que o formato textual (bases não estruturadas) é, segundo Maedche e Staab [126], o formato no qual se encontra a maior parte do conhecimento disponível.

Frequentemente, os *corpora* são constituídos sobre um domínio específico com o intuito de servir como descrição/definição/caracterização desse domínio. é possível afirmar que o uso desse tipo de *corpora* permite economizar os esforços de especialistas do domínio para realizar tarefas de extração de termos e outras formas de descoberta de conhecimento em geral.

Diversos *corpora* estão disponíveis, sendo a maior parte deles em língua inglesa. Alguns *corpora* de ampla divulgação são: *Brown corpus* [106], *Lancaster-Olso-Bergen corpus* [75], *Penn Treebank* [132], *Lonely Planet corpus* [97] e *Genia corpus* [101].

Além do inglês, outros idiomas possuem uma relativa abundância de *corpora*, como é o caso do *corpus* utilizado por Kietz *et al.* com textos em alemão coletados na *intranet* de uma companhia de seguros [99]. Outro exemplo é o *corpus* utilizado por Bourigault e Lame [28] composto por códigos legais franceses. Eventualmente, encontram-se *corpora* bilíngues, *e.g.*, o *corpus* desenvolvido por Kilgarriff *et al.* [100] que reúne textos em irlandês e inglês.

Infelizmente, para o português o número de *corpora* disponíveis é consideravelmente menor, principalmente tratando-se de *corpora* de domínios científicos. Uma das exceções é o *corpus* de Pediatria (PED) desenvolvido por Coulthard [49] a partir de 183 textos do Jornal de Pediatria, um periódico bilíngue da Sociedade Brasileira de Pediatria. Devido a essa escassez de *corpora* sobre domínios científicos em português, e para suprir as necessidades dessa tese, foi construído um conjunto de *corpora* sobre domínios específicos [120].

Especificamente, foram criados quatro *corpora* sobre os seguintes domínios específicos[1]: Modelagem estocástica (ME); Mineração de dados (MD); Processamento paralelo (PP); e Geologia (GEO). De um ponto de vista prático, nessa tese utilizam-se, então, cinco *corpora*, cujas características estão descritas na Tabela 3.1 que apresenta o número de textos, frases e palavras de cada um dos *corpora*.

Tabela 3.1: Características dos *Corpora*.

corpora		Número de textos	Número de frases	Número de palavras
Pediatria	PED	281	27.724	835.412
Modelagem estocástica	ME	88	44.222	1.173.401
Mineração de dados	MD	53	42.932	1.127.816
Processamento paralelo	PP	62	40.928	1.086.771
Geologia	GEO	234	69.461	2.010.527

3.2 Anotação Linguística, Processo Básico de Extração

A anotação linguística de um *corpus* é um processo complexo e empregado em abordagens de processamento de linguagem natural que não sejam puramente estatísticas. Diversas possibilidades de anotação linguística estão disponíveis em várias línguas [51, 179, 4], mas, em português, poucas opções estão operacionais enquanto *parsers* completos [116].

Dentre as opções disponíveis para português existe a ferramenta *LX parser* [177], recentemente disponibilizada *online* pela equipe dirigida por António Branco da Universidade de Lisboa. Uma outra opção de *parser* para o português é a ferramenta de software PALAVRAS [20] desenvolvida por Eckhard Bick na Universidade de Arhus (Dinamarca) desde 2000. Ao contrário do *LX parser*, o PALAVRAS vem sendo utilizado por diversos pesquisadores da área de processamento de linguagem natural há vários anos [34, 124, 194, 155, 159, 17, 204], e portanto seu uso se configura em uma verdadeira referência no tratamento de língua portuguesa.

Dessa forma, o *parser* PALAVRAS foi utilizado como ferramenta de anotação linguística para os trabalhos desenvolvidos nessa tese. No entanto, cabe salientar que o uso de outros *parsers* não inviabiliza nenhuma das contribuições científicas aqui apresentadas. Na verdade, conforme será visto nas conclusões dessa tese, um trabalho futuro natural será experimentar

[1] Os domínios escolhidos para compor os *corpora* se justificam pela disponibilidade de especialistas disponíveis no Programa de Pós-Graduação em Ciência da Computação durante essa etapa desse trabalho de tese.

todas as técnicas expostas nesse capítulo, bem como os métodos descritos nos demais, para outros *parsers*. O *LX parser* é o primeiro candidato natural a ser testado como alternativa ao PALAVRAS para a anotação linguística.

3.2.1 Anotação Linguística

O processo de anotação linguística do PALAVRAS é aplicado individualmente a cada frase dos documentos. A base linguística para esse processo foge ao escopo dessa tese, e por isso, todas as descrições dessa seção irão limitar-se à apresentação do processo de extração empregado, sem se aprofundar em questões linguísticas ou terminológicas. O leitor interessado em maiores detalhes sobre o PALAVRAS deve consultar a bibliografia original em [20] e também visitar o site Floresta Sintáctica [196] que apresenta alguns detalhes específicos além da anotação *on line* de frases.

Cada frase reconhecida é armazenada pelo *parser* como uma estrutura em árvore composta por nós terminais (as folhas da árvore) que representam as palavras e nós não-terminais que representam estruturas gramaticais. Um exemplo disso é apresentado na Figura 3.1, em que está representada a anotação linguística realizada pelo *parser* para a frase "*Essas duas cidades são os maiores e mais importantes centros de pesquisa no Brasil.*".

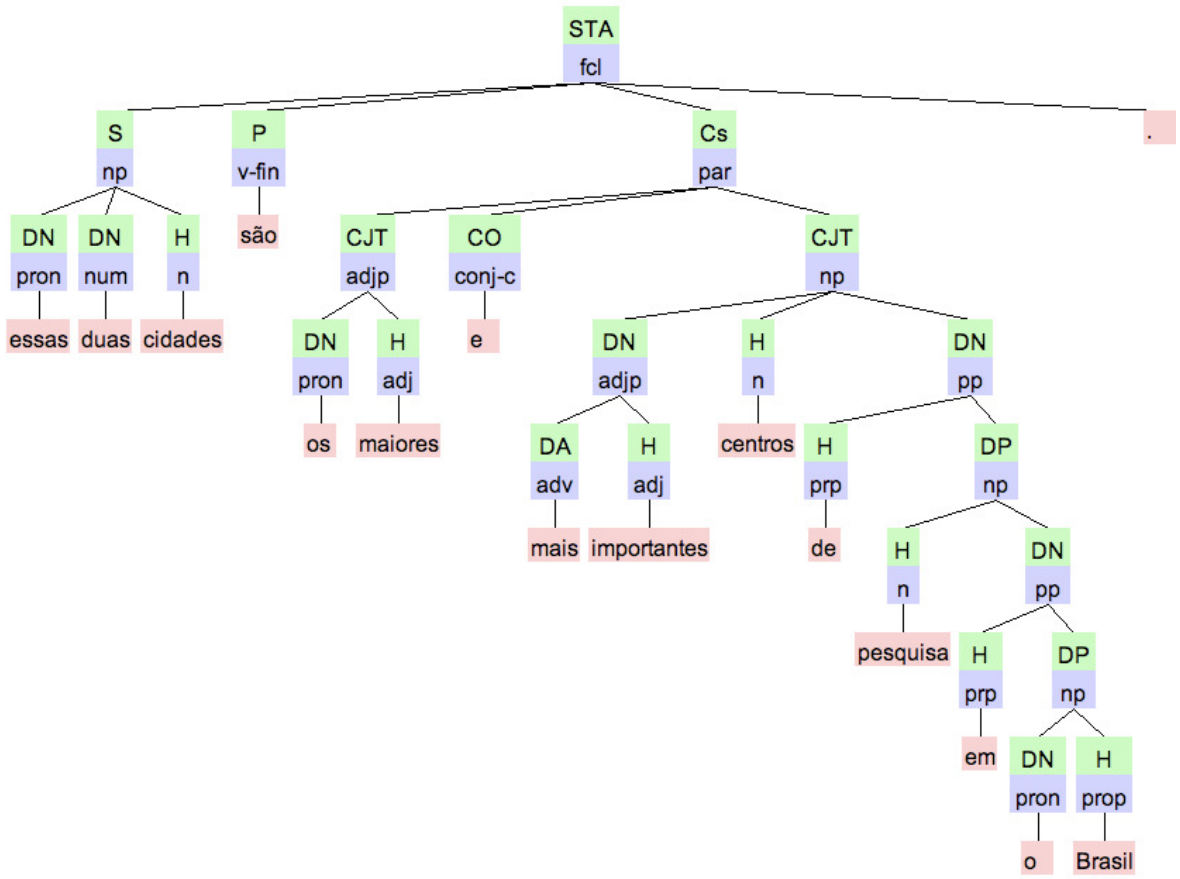


Figura 3.1: Anotação feita pelo *parser* para a frase: "Essas duas cidades são os maiores e mais importantes centros de pesquisa no Brasil.".

A primeira observação quanto ao exemplo da Figura 3.1 é que utilizam-se nós não-terminais para representar estruturas gramaticais que podem ser tão complexas como orações, mas também estruturas mais simples como uma única palavra. Importa saber que cada estrutura, seja uma oração ou uma palavra única, receberá do PALAVRAS pelo menos duas etiquetas:

uma que define sua função gramatical na frase (*e.g.*, sujeito - "S", predicado - "P", *etc.*); outra que define sua função sintática (*e.g.*, sintagma nominal - "np", adjetivo - "adj", *etc.*).

Já os nós terminais serão utilizados para representar palavras (ou *tokens*) que compõem as frases. Para cada *token*, o *parser* associa um conjunto maior de informações como a forma canônica de cada palavra, sua morfologia, sua função sintática e sua provável semântica.

3.2.2 Processo Básico de Extração de Termos

A primeira informação importante relativa ao processo de extração de termos, no contexto dessa tese, é considerar os sintagmas nominais (SNs) como os portadores de informação conceitual [107]. Em função disso, todo SN é, em princípio, um termo candidato a conceito do domínio. Nesse sentido, somente critérios arbitrários (que serão vistos em detalhe nos próximos capítulos) irão definir quais SNs serão efetivamente considerados conceitos. Porém antes disso, é necessário definir como os SNs serão detectados a partir da saída do *parser* PALAVRAS.

O processo de identificação de SNs passa inicialmente pela detecção dos não-terminais identificados pela etiqueta "np", que para o PALAVRAS são todos SNs compostos por mais de um *token*.

Um exemplo claro dessa detecção pode ser visto na Figura 3.2, em que encontram-se os SNs indicados pelas etiquetas "np". Na frase: "A gastroesquise é um defeito da parede abdominal anterior.", esses SNs são:

- "A gastroesquise", indicado como SN (etiqueta "np"), que cumpre a função de sujeito (etiqueta "S");
- "um defeito de a parede abdominal anterior", indicado como SN (etiqueta "np"), que cumpre a função de complemento do sujeito (etiqueta "Cs");
- "a parede abdominal anterior", indicado como SN (etiqueta "np"), que cumpre a função de argumento da preposição "de" (etiqueta "DP").

Porém, SNs que são compostos por um único *token* não são identificados pelo PALAVRAS com a etiqueta "np". Por exemplo, reescrevendo a frase da Figura 3.2, retirando o artigo que inicia a frase, temos a nova frase anotada na Figura 3.3.

Nesse novo exemplo (Figura 3.3), o primeiro SN é composto por um único *token* ("Gastroesquise"), que é anotado pelo *parser* como sujeito da oração (etiqueta "S") e substantivo próprio (etiqueta "prop"). Mesmo não estando indicado pela anotação do PALAVRAS com a etiqueta "np", sem dúvida esse SN deve ser considerado para a extração.

A diferença entre as duas frases (Figura 3.2 e 3.3) se resume a uma diferença de estilo de escrita, logo é natural que, para esse exemplo, sejam extraídos praticamente os mesmos SNs, ou seja:

- "Gastroesquise", termo com único *token* indicado como sujeito da oração (etiqueta "S") e indicado como um substantivo próprio (etiqueta "prop");
- "um defeito de a parede abdominal anterior", indicado como sintagma nominal (etiqueta "np"), que cumpre a função de complemento do sujeito (etiqueta "Cs");
- "a parede abdominal anterior", indicado como sintagma nominal (etiqueta "np"), que cumpre a função de argumento da preposição "de" (etiqueta "DP").

Dessa forma, o método básico de extração proposto analisa o resultado da anotação linguística feita pelo *parser* para extrair todos os termos *multi-token* marcados com a etiqueta "np" e todos os termos com um *token* único que estejam marcados como sujeito (etiqueta "S"), objeto (etiquetas "Od", "Oi" e "Op") ou seus complementos (etiquetas "Cs" e "Co").

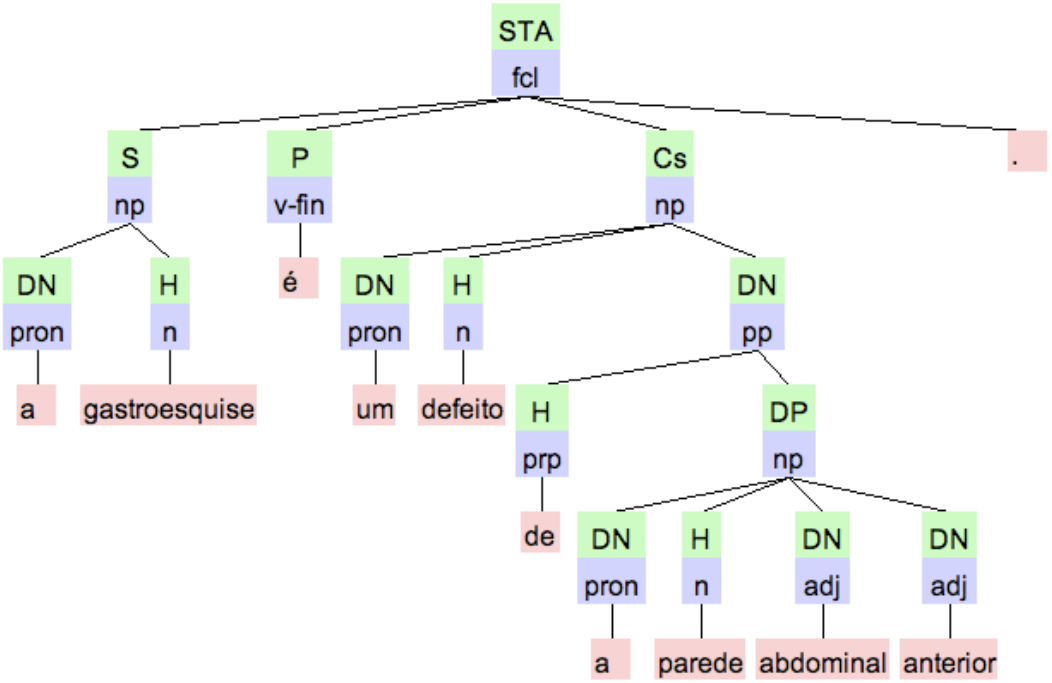


Figura 3.2: Anotação feita para a frase: "A gastroesquise é um defeito da parede abdominal anterior.".

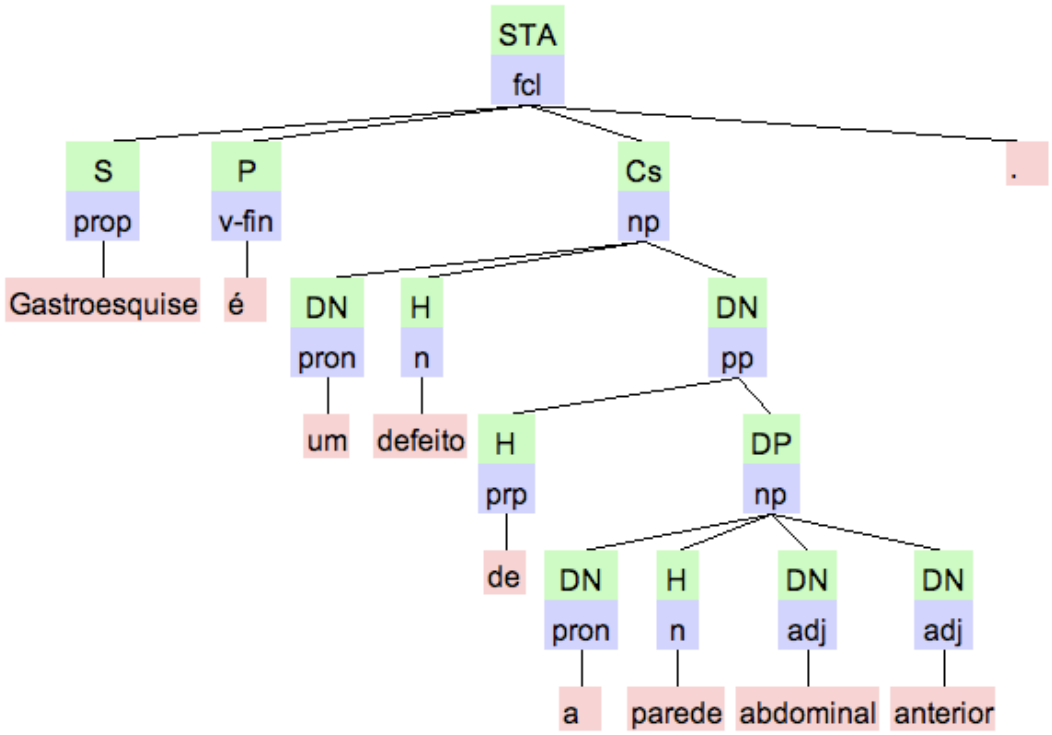


Figura 3.3: Anotação feita para a frase: "Gastroesquise é um defeito da parede abdominal anterior.".

Aplicando o processo básico de extração de termos a todos os *corpora* citados na Seção 3.1 (Pediatria - PED, Modelagem estocástica - ME, Mineração de dados - MD, Processamento paralelo - PP, e Geologia - GEO) são apresentados, na Tabela 3.2, os números de SNs em cada *corpus*, devidamente divididos segundo o número de palavras que cada termo contém (unigramas, bigramas, *etc.*). Nessa tabela a penúltima linha (N-grama) indica o número de SNs extraídos com 10 ou mais palavras e a última linha indica o total de termos extraídos.

Tabela 3.2: Número de termos extraídos originalmente de cada *corpora*.

número de palavras	PED	ME	MD	PP	GEO
unigramas	5.583	9.141	8.362	8.279	10.909
bigramas	58.504	81.723	74.939	75.822	120.477
trigramas	25.485	35.454	34.373	32.484	61.909
4-gramas	17.150	26.510	27.304	25.950	46.484
5-gramas	16.994	24.291	22.940	22.265	39.310
6-gramas	12.334	16.395	16.323	15.429	30.254
7-gramas	8.850	13.011	12.786	12.144	23.994
8-gramas	7.128	10.204	9.844	9.427	19.275
9-gramas	5.879	7.687	7.681	7.207	15.442
N-gramas	31.239	36.964	42.048	38.699	89.421
total	189.146	264.380	256.600	247.706	457.475

Os SNs extraídos pelo método básico, no entanto, carecem de um tratamento para que possam ser considerados candidatos a conceitos de um domínio. Nesse sentido, as duas próximas seções propõem e avaliam um conjunto de regras heurísticas que visa refinar o conjunto de SNs extraídos com o processo básico.

3.3 Heurísticas Propostas

Todas as heurísticas propostas são baseadas em análise linguísticas, logo, a efetividade das heurísiticas é dependente da qualidade da anotação. Para aplicar as heurísticas assume-se que sejam recebidos todos sintagmas nominais (SNs) do *corpus*, a anotação sintática de cada palavra que o compõe e a função gramatical que o SN desempenha na frase (sujeito, objeto ou complemento). As heurísticas propostas estão divididas em três grupos: Heurísticas de ajuste: heurísticas que adaptam SNs extraídos; Heurísticas de descarte: heurísticas que recusam SNs inadequados; Heurísticas de inclusão: heurísticas que detectam SNs implícitos que não podem ser encontrados pela simples leitura sequencial dos textos do *corpus*.

3.3.1 Heurísticas de Ajuste

As heurísticas de ajuste têm por objetivo remover palavras que não carregam significado para o termo representado pelo SN. As regras propostas são a remoção de:

- artigos no começo de um SN;
- artigos em qualquer posição de um SN;
- pronomes no começo de um SN; além de
- pronomes em qualquer posição de um SN.

Ferramentas de extração que seguem abordagens estatísticas também oferecem técnicas semelhantes através do uso de listas de "stop word". No entanto, é importante salientar que as regras propostas de remoção de palavras são baseadas em uma anotação linguística prévia, logo, elas tendem a ser mais precisas do que abordagens estatísticas.

Outro ponto importante das heurísticas de ajuste é que sua utilização reduz o número de palavras de um SN. Por exemplo, sua aplicação em um trigrama pode transformá-lo em um bigrama.

3.3.1.1 A1 - Regra de Ajuste 1 – Remoção de Artigos no Início de SNs

A primeira heurística é a simples remoção de artigos que aparecem no início do SN. Ainda que artigos tenham um papel importante como determinantes, a remoção do primeiro artigo de um SN é coerente com o objetivo de extrair termos candidatos a conceitos. O SN "o leite materno" é, sem dúvida, diferente do SN "um leite materno", porém ambos SNs fazem referência ao candidato a conceito de domínio "leite materno".

Posto que os artigos são um conjunto finito de palavras, a remoção de artigos no início de um SN é um processo simples que pode ser feito sem o auxílio de uma anotação linguística. Apesar disso, a anotação linguística permite uma remoção mais precisa, pois nem sempre palavras usadas como artigo possuem uma única classe gramatical. Por exemplo, o artigo definido feminino "a", escreve-se igual à preposição "a", ou ainda ao pronome oblíquo feminino "a". Portanto, ao colocar a palavra "a" em uma "stop list", uma extração puramente estatística iria remover esta palavra sendo ela empregada como artigo, preposição ou pronome.

A aplicação dessa heurística sobre os 189.146 SNs encontrados no *corpus* de Pediatria resulta no ajuste de 81.031 SNs (cerca de 43%).

3.3.1.2 A2 - Regra de Ajuste 2 – Remoção de Todos Artigos de SNs

A segunda heurística de ajuste é a remoção de todos artigos encontrados em um SN, e não apenas artigos que aparecem no início de um SN. Dessa forma, a regra A2 é uma generalização da anterior (A1), e todas considerações feitas para a regra A1 sobre como a remoção de artigos altera o significado de um termo, continuam verdadeiras para a regra A2. No entanto, essa segunda heurística dificilmente poderia ser aplicada em um método puramente estatístico, pois com ela é possível considerar termos compostos por palavras não contíguas.

Um exemplo de aplicação da regra A2 sobre o SN "o leite da mãe" resulta no termo "leite de mãe". Note-se que a preposição "de" e o artigo definido feminino "a" estão contraídos na palavra "da". A aplicação dessa heurística é a mais impactante, pois dos 189.146 SNs inicialmente anotados no *corpus* de Pediatria, pouco menos da metade (92.754) possuíam artigos e foram, portanto, ajustados.

3.3.1.3 A3 - Regra de Ajuste 3 – Remoção de Pronomes no Início de SNs

Semelhante à primeira heurística, a remoção de pronomes no início de SNs, a regra A3 tem por objetivo manter o SN genérico o suficiente para ser considerado candidato a conceito. Portanto, essa heurística só é aplicada quando o pronome a ser removido não é o núcleo do SN. Cabe lembrar que, embora o usual seja que o núcleo de um SN seja um substantivo (comum ou próprio), é possível ter como núcleo um adjetivo ou um verbo no particípio passado fazendo o papel de um substantivo, ou até um pronome referenciando um substantivo citado em outro lugar (uma anáfora).

A aplicação da regra A3 sobre os 189.146 SNs extraídos do *corpus* de Pediatria ajusta 12.793 SNs.

3.3.1.4 A4 - Regra de Ajuste 4 – Remoção de Todos Pronomes de SNs

De forma análoga aos artigos, a regra A4 propõe a remoção de pronomes que se encontram em qualquer posição de um SN. As mesmas considerações feitas na regra A3 são válidas, principalmente, a restrição que, só é possível remover pronomes que não sejam núcleo do SN.

Dessa forma, a aplicação da regra A4 ao SN "o objetivo de seu movimento" transforma-o no termo "o objetivo de movimento". Note-se que o artigo "o" não é removido, pois exemplifica-se a aplicação da regra A4 sozinha. A aplicação dessa heurística sobre as frases que compõem o *corpus* de Pediatria resulta no ajuste de 18.230 termos do total de 189.146 termos extraídos.

3.3.2 Heurísticas de Descarte

As heurísticas de descarte são regras que recusam SN anotados que provavelmente não são termos representativos de um domínio. Essas heurísticas são regras que descartam SNs que:

- contém numerais;
- contém outros símbolos além de letras, dígitos ou hífen;
- o núcleo é um pronome; ou
- começam com um advérbio.

Ao contrário das heurísticas de ajuste, as heurísticas de descarte não alteram o número de palavras dos SNs, mas reduzem significativamente o número total de SNs extraídos. Considerando a aplicação de todas heurísticas de descarte sobre os 189.146 SNs originalmente extraídos do *corpus* de Pediatria recusou 55.896 SNs, ou seja, um pouco menos de 30% dos termos originalmente extraídos são descartados.

3.3.2.1 D1 - Regra de Descarte 1 – Recusa de SNs com Numerais

A primeira heurística de descarte recusa SNs que contêm numerais, seja na forma escrita ou utilizando caracteres numéricos (dígitos). Apesar de ser uma heurística bastante restritiva que ignora termos como "as sete maravilhas" ou "os três mosqueteiros", essa heurística é frequentemente válida para descartar SNs que expressam quantidades que são comuns em textos científicos.

Exemplos de sucesso da aplicação dessa regra no *corpus* de Pediatria é o descarte dos SNs "três meses" e "ano 2000". Na verdade, essa heurística é bastante eficiente por excluir SNs que fazem referências a datas. A aplicação da regra D1 sobre os 186.146 SNs extraídos do *corpus* de Pediatria resultou na recusa de 30.969 termos.

3.3.2.2 D2 - Regra de Descarte 2 – Recusa de SNs com Símbolos

Analogamente à recusa de SNs com numerais, a regra D2 descarta SNs que contém símbolos, ou seja, só aceita SNs compostos por letras e dígitos. Porém, aceita-se também o caracter hífen ("-") que é usual em palavras compostas, como por exemplo: "recém-nascido" e "bem-estar".

Muitos dos SNs recusados pela presença de símbolos também possuem numerais, como por exemplo, valores percentuais ("46%"). Encontra-se também símbolos em endereços eletrônicos ("info@saude.gov.br") ou representações abreviadas de números ordinais ("2º").

A aplicação da regra D2 nos 189.146 SNs extraídos do *corpus* de Pediatria resultaram na recusa de 40.989 SNs, tornando essa regra a mais restritiva dentre as heurísticas de descarte, ou seja, mais de 21% dos termos extraídos são descartados devido a essa heurística.

3.3.2.3 D3 - Regra de Descarte 3 – Recusa de SNs com um Pronome como Núcleo

Usualmente o núcleo de um SN é um substantivo comum ou próprio. No entanto, o núcleo de um SN também pode ser um adjetivo, um verbo no particípio passado ou um pronome. A terceira heurística de descarte visa aceitar somente SNs cujo o núcleo possui um significado autocontido, ou seja, o núcleo é um substantivo, adjetivo ou verbo no particípio passado. Consequentemente, recusa-se SNs quando o núcleo é um pronome, ou seja, quando o SN indica um termo explicitamente mencionado em outro ponto do texto (anáfora).

Algumas situações de SNs com núcleos de diferentes classes gramaticais são exemplificados nas frases indicadas na Tabela 3.3.

Tabela 3.3: Frases com núcleos de SN de diferentes classes gramaticais.

	Frase Exemplo (SN de interesse em **negrito**)	núcleo	classe gramatical
1	**Os alunos espertos** podem prever dificuldades.	alunos	substantivo comum
2	**Os espertos** podem prever dificuldades.	espertos	adjetivo
3	O aleitamento materno é fundamental para **os recém-nascidos**.	nascidos	particípio passado
4	O aleitamento materno é fundamental para **as crianças recém-nascidas**.	crianças	substantivo comum
5	**A Madalena arrependida** teve dificuldade em explicar-se.	Madalena	substantivo próprio
6	**A arrependida** teve dificuldade em explicar-se.	arrependida	particípio passado
7	**Eles** não foram encontrados apesar dos esforços empregados.	eles	pronome pessoal
8	O esforço empregado gerou grandes expectativas, mas frustou **as nossas**.	nossas	pronome possessivo
9	**Aqueles que sabiam**, perguntaram.	aqueles	pronome demonstrativo

Como pode ser observado nas frases 1 e 2 da Tabela 3.3, o SN com o núcleo "espertos" (um adjetivo) pode não ser tão adequado como conceito quanto o SN "os alunos espertos" que possui como núcleo um substantivo. Por outro lado, observando a frase 3 da Tabela 3.3, o SN "os recém-nascidos", que também não possui substantivo, é bastante significativo, sendo talvez mais significativo que o SN encontrado na frase 4, "crianças recém-nascidas". Por essa razão, opta-se por aceitar SNs que possuem como núcleo um adjetivo ou verbo no particípio passado, e não só substantivos.

Para a frase 5 observa-se a utilização de um nome próprio com uma função que se assemelha mais à de um substantivo, logo aceitar nomes próprios pode ser adequado. Isso fica claro se comparado ao exemplo da frase 6, em que o adjetivo "arrependida" traz menos informação que o SN utilizado na frase 5.

Finalmente, para os exemplos nas frases 7, 8 e 9, fica claro que SNs que possuem pronomes como núcleo não fornecem bons candidatos a conceitos. Isso verifica-se tanto em utilizações comuns, como na frase 7, quanto em estruturas mais complexas, como nas frases 8 e 9.

Note-se que, de acordo com o propósito da extração de termos, pode ser interessante descartar SNs segundo a classe gramatical do núcleo. Para os trabalhos desenvolvidos nessa tese, são aceitos SNs que possuem como núcleo substantivos comuns ou próprios, adjetivos ou verbos no particípio passado, ou seja, recusa-se SNs cujo núcleo é um pronome. A aplicação da regra D3 sobre os 189.146 SNs extraídos do *corpus* de Pediatria causou a recusa de 6.109 termos.

3.3.2.4 D4 - Regra de Descarte 4 – Recusa de SNs que Iniciam com Advérbio

A última heurística de descarte baseia-se no fato de que alguns SNs não se referem explicitamente a um termo, mas apenas fazem referência a termos previamente mencionados. Nesses casos, usualmente o SN começa com um advérbio e possui como núcleo um adjetivo. Esses SNs não são adequados a serem considerados candidatos a conceitos, pois eles não carregam uma informação completa.

Por exemplo, no *corpus* de Pediatria o SN "mais frequente" foi encontrado 11 vezes, mas nessas ocorrências ele foi empregado 5 vezes para referenciar o uso frequente de um medicamento, e 6 vezes para referenciar a adoção frequente de um hábito por um paciente. No entanto, é inútil considerar o SN "mais frequente" como um candidato a conceito, pois somente observando os contextos onde o termo é empregado torna-se possível saber se ele está se referindo a um medicamento ou um hábito de pacientes.

A aplicação da heurística D4 sobre os 189.146 SNs do *corpus* de Pediatria fez com que apenas 650 termos fossem descartados. Esse número é relativamente baixo, porém é importante perceber que sua remoção representa uma clara melhora no processo de extração, pois descarta-se SNs que não carregam informação conceitual.

3.3.3 Heurísticas de Inclusão

As heurísticas de inclusão têm por objetivo detectar SNs implícitos. De um ponto de vista linguístico, essas regras são as mais sofisticadas dentre as propostas, pois através delas considera-se SNs que não aparecem no texto, mas podem ser inferidos pela anotação linguística. As heurísticas de inclusão são:

- remoção sucessiva de adjetivos;
- uso de predicado múltiplo; e
- conjunção de adjetivos.

O efeito prático das heurísticas de inclusão é o aumento no número total de SNs. Por exemplo, os 133.250 SNs extraídos do *corpus* de Pediatria, que não foram descartados após a aplicação das heurísticas de descarte, dão origem a 46.617 SNs implícitos. Dessa forma, as heurísticas de inclusão são responsáveis por adicionar um número menor do que os 55.896 SNs que foram removidos pelas heurísticas de descarte.

3.3.3.1 I1 - Regra de Inclusão 1 – Detecção de SNs Implícitos por Remoção Sucessiva de Adjetivos

A primeira heurística de inclusão está baseada na detecção de SNs contidos em SNs maiores pela remoção sucessiva de adjetivos. Por exemplo, a frase "Estudos realizados mostram o perigo de doenças virais hemorrágicas." (Figura 3.4), mostra um caso em que o processo básico de extração detectaria apenas os seguintes SNs: "Estudos realizados"; "perigo de doenças virais hemorrágicas"; e "doenças virais hemorrágicas".

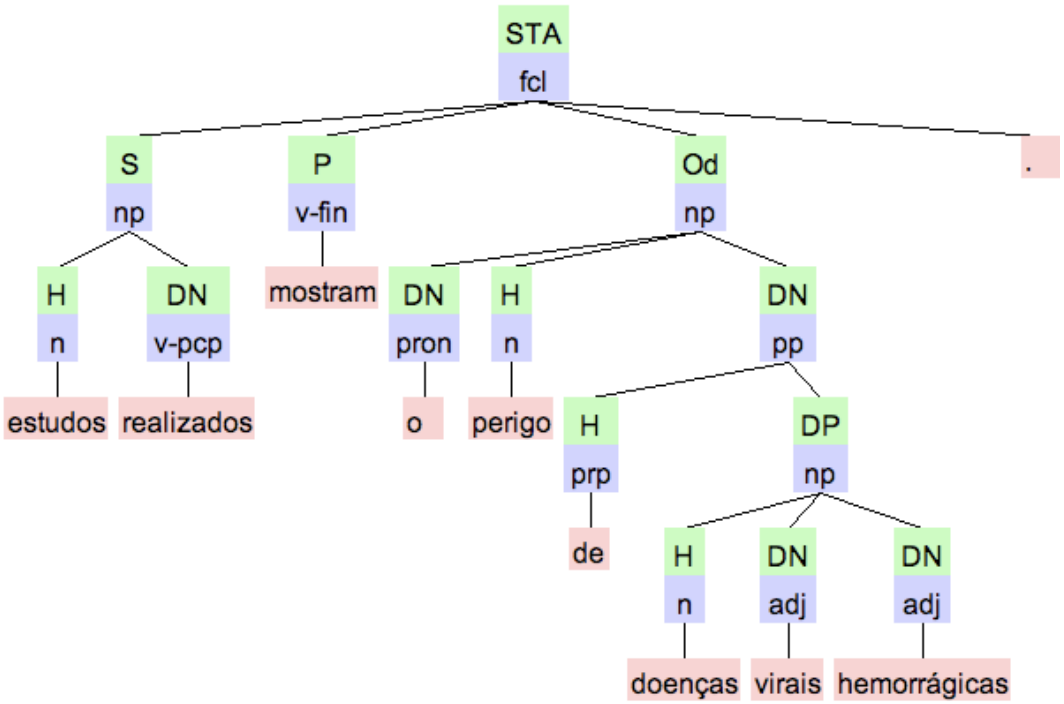


Figura 3.4: Anotação feita para a frase: "Estudos realizados mostram o perigo de doenças virais hemorrágicas.".

A proposta dessa heurística consiste em gerar termos adicionais pela remoção dos adjetivos (ou verbos no particípio passado) ao fim de cada termo. Dessa maneira, a Tabela 3.4 apresenta os termos que seriam extraídos da frase exemplificada na Figura 3.4.

Tabela 3.4: Termos extraídos por remoção sucessiva de adjetivos ou verbos no particípio passado.

	Termo Extraído	Termo Completo (removido em **negrito**)	classe gramatical do removido
1	Estudos realizados	–	–
2	Estudos	Estudos **realizados**	particípio passado
3	perigo de doenças virais hemorrágicas	–	–
4	perigo de doenças virais	o perigo de doenças virais **hemorrágicas**	adjetivo
5	perigo de doenças	o perigo de doenças **virais**	adjetivo
6	perigo	o perigo **de doenças**	sintagma preposicional
7	doenças virais hemorrágicas	–	–
8	doenças virais	doenças virais **hemorrágicas**	adjetivo
9	doenças	doenças **virais**	adjetivo

No *corpus* de Pediatria 40.156 SNs terminam com pelo menos um adjetivo. A aplicação da heurística I1 a esses SNs resultou na inclusão de 44.020 novos SNs, ou seja, essa heurística é responsável por quase todos os 46.617 SNs incluídos pelas heurísticas desse terceiro grupo.

3.3.3.2 I2 - Regra de Inclusão 2 – Detecção de SNs Replicados pelo Uso de Predicado Múltiplo

Na língua portuguesa é comum encontrar o uso de predicados com mais de um verbo. Nesses casos, a sentença representa múltiplas frases com o mesmo sujeito e objeto, cada uma delas utilizando um dos verbos do predicado. A segunda heurística de inclusão atua nesse tipo de situação, considerando como se as frases com predicado múltiplo, fossem desmembradas em diversas frases com um único verbo.

Dessa forma, a regra I2 não cria SNs diferentes dos originalmente encontrados, pois ela somente replica ocorrências de SNs, que são sujeito ou objeto de uma sentença que possui predicado com mais de um verbo. Por exemplo, a frase "Pacientes idosos compram e tomam remédios mais caros.", ilustrada na Figura 3.5, mostra esse tipo de situação.

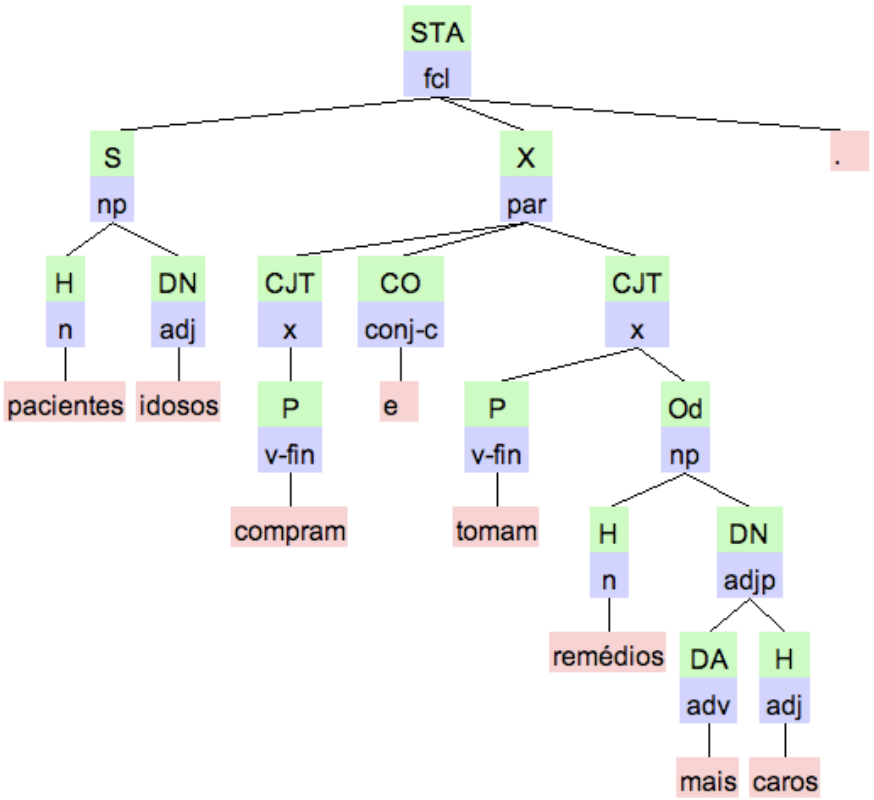


Figura 3.5: Anotação feita para a frase: "Pacientes idosos compram e tomam remédios mais caros.".

Percebe-se, pela atribuição de etiquetas feita pelo *parser*, assim como pelo próprio sentido da frase descrita na Figura 3.5, que ela poderia ser reescrita por duas frases iguais em tudo exceto pelo predicado:

"Pacientes idosos compram remédios mais caros."
"Pacientes idosos tomam remédios mais caros."

Caso essa frase seja desdobrada em duas, os SNs relacionados ao predicado duplo serão computados com duas ocorrências cada. Nesse sentido, a regra I2 propõe que SNs que estejam relacionados com predicados múltiplos sejam computados tantas vezes quantos forem os verbos do predicado. A aplicação da regra I2 sobre a frase da Figura 3.5 faz com que os SNs "Pacientes idosos" (sujeito) e "remédios mais caros" (objeto) sejam considerados duas vezes cada um, ou seja, como se a extração fosse feita sobre as frases desmembradas.

No *corpus* de Pediatria foram encontradas 3.413 frases com predicado múltiplo que deram origem a 3.472 novas ocorrências de SNs. Cabe salientar que os predicados múltiplos podem ocorrer entre dois verbos, ou ainda em uma lista de três ou mais verbos separados por vírgulas.

3.3.3.3 I3 - Regra de Inclusão 3 – Detecção de SNs implícitos por Conjunção de Adjetivos

A última heurística de inclusão também é baseada na detecção de estruturas gramaticais múltiplas com o uso de conjunções, mas ao contrário de regra I2, a regra I3 detecta SNs implícitos quando um mesmo substantivo é qualificado por dois ou mais adjetivos. Por exemplo, a frase "Os pacientes idosos ou obesos possuem maior risco de diabetes.", ilustrada na Figura 3.6, mostra um caso em que o *parser* identifica o primeiro SN corretamente como "Os pacientes idosos", porém fica implícito, nessa frase, também o SN "Os pacientes obesos".

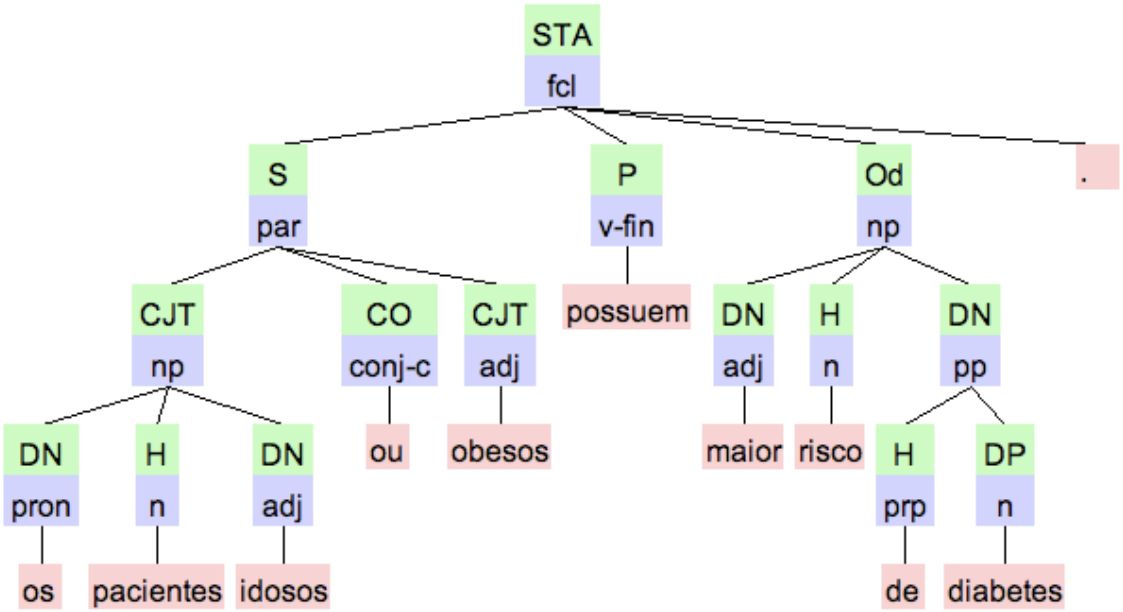


Figura 3.6: Anotação feita para a frase: "Os pacientes idosos ou obesos possuem maior risco de diabetes.".

A Tabela 3.5 apresenta alguns exemplos de termos implícitos criados. Observando a primeira e a segunda frase dessa tabela, percebe-se que a heurística I3 pode ser empregada sem riscos quando a conjunção é alternativa ("ou"), porém, quando a conjunção aditiva "e" é empregada, a semântica da frase se presta a diferentes interpretações. Enquanto que, da primeira frase, se compreende que basta a uma pessoa ser "esperta" ou ser "sábia" para prever dificuldades, a segunda frase sugere que somente a pessoa que for ao mesmo tempo "esperta e sábia" poderá

prever dificuldades. No entanto, essa duplicidade de interpretação não invalida a existência, enquanto termos portadores de informação, dos SNs "pessoas espertas" e "pessoas sábias".

Tabela 3.5: Frases com termos implícitos e sua detecção.

	Frase Exemplo (SN explícito em **negrito**)	SN implícito
1	**As pessoas espertas** ou sábias podem prever dificuldades.	pessoas sábias
2	**As pessoas espertas** e sábias podem prever dificuldades.	pessoas sábias
3	O aleitamento materno é vital para **recém-nascidos normais** e prematuros.	recém-nascidos prematuros
4	O defeito pode aparecer n**a parede abdominal anterior** ou torácica posterior.	a parede torácica posterior

As frases 3 e 4 mostram que é necessário substituir o mesmo número de adjetivos do SN explícito, quantos forem os adjetivos encontrados após a conjunção. Por exemplo, na frase 4 não seria correto gerar o termo implícito "A parede abdominal torácica posterior", pois os adjetivos que seguem a conjunção "ou" correspondem a duas palavras ("torácica posterior") e portanto devem substituir em igual medida os adjetivos do termo explícito ("abdominal anterior").

A aplicação da regra I3 sobre os termos originalmente extraídos do *corpus* de Pediatria resultou na inclusão de 861 novos SNs.

3.4 Avaliação Numérica das Heurísticas Propostas

Nessa seção são relatados os experimentos realizados sobre o *corpus* de Pediatria com o intuito de avaliar o uso das heurísticas propostas. Foi escolhido esse *corpus* pelo fato de possuir associado a ele listas de termos de referência construídas por um grupo externo (`www.ufrgs.br/textecc`). Essas listas são compostas por 1.534 bigramas e 2.660 trigramas e estão disponíveis no anexo A. Logo, torna-se possível comparar termos extraídos com a referência, utilizando medidas usuais da área de recuperação de informação [192].

Especificamente, exemplifica-se o benefício trazido pelas 11 heurísticas propostas para a extração de bigramas e trigramas do *corpus* de Pediatria. A quantificação desses benefícios é feita pelo cálculo da precisão, abrangência e medida F (Seção 2.3.3).

Para avaliar a aplicação de cada heurística, compara-se as listas de bigramas e trigramas extraídos mais frequentes às listas de referência. A extração básica de termos do *corpus* de Pediatria detecta 58.504 bigramas e 25.485 trigramas (veja Tabela 3.2), porém observando o número de bigramas e trigramas distintos[2], contabiliza-se apenas 17.407 e 15.577, respectivamente. Logo, para as experiências dessa seção escolheu-se considerar listas com 10% desses termos, ou seja, listas com os termos mais frequentes. Essa escolha de considerar os 10% mais frequentes organizados segundo a frequência absoluta é consistente com resultados preliminares publicados por Lopes *et al.* [123]. Dessa forma, compara-se os 1.741 bigramas e os 1.558 trigramas com maior frequência absoluta (denominada $\mathcal{LE}$) com os 1.534 bigramas e 2.660 trigramas das listas de referência (denominada $\mathcal{LR}$), respectivamente.

3.4.1 Resultados Numéricos para as Heurísticas de Ajuste

A Tabela 3.6 ilustra os benefícios trazidos pela aplicação das heurísticas de ajuste sobre os SNs extraídos do *corpus* de Pediatria. Além dos valores de precisão (P), abrangência (R) e medida-F (F), a última coluna indica quantos termos da lista de referência foram encontrados na lista dos termos extraídos mais frequentes (10%), ou seja, a intersecção entre $\mathcal{LE}$ e $\mathcal{LR}$.

[2]A lista de termos extraídos contém diversas ocorrências de termos repetidos. Porém ao contabilizar o número de ocorrências de cada termo, reduz-se o tamanho da lista, pois considera-se apenas o número de termos distintos.

A primeira linha (*nenhuma*) mostra os resultados obtidos sem a aplicação de nenhuma das heurísticas. A próximas 4 linhas indicam os resultados obtidos aplicando cada uma das heurísticas de ajuste individualmente. Finalmente, a última linha (*todas*) indica os resultados obtidos aplicando todas heurísticas de ajuste simultaneamente.

Tabela 3.6: Benefícios obtidos com as heurísticas de ajuste.

Bigramas				
heurísticas de ajuste	P	R	F	$\mid \mathcal{LR} \cap \mathcal{LE} \mid$
nenhuma	12%	13%	13%	206
A1	38%	43%	40%	653
A2	38%	43%	40%	654
A3	14%	16%	15%	252
A4	15%	17%	16%	257
todas A	48%	55%	51%	839
Trigramas				
heurísticas de ajustes	P	R	F	$\mid \mathcal{LR} \cap \mathcal{LE} \mid$
nenhuma	13%	8%	10%	202
A1	55%	32%	40%	852
A2	59%	34%	43%	914
A3	15%	9%	11%	229
A4	16%	9%	12%	242
todas A	60%	35%	44%	934

A primeira observação dos dados da Tabela 3.6 é que a extração de SNs sem nenhuma heurística resulta em valores baixos de precisão e abrangência. Esses valores são similares àqueles encontrados em métodos básicos de extração baseados no uso de anotação linguística feita pelo PALAVRAS [165]. No entanto, após a remoção de artigos (heurísticas A1 e A2) percebe-se um grande aumento (de 25% a 43%) nos valores de precisão e abrangência.

As heurísticas de remoção de pronomes (A3 e A4) foram menos efetivas, mas ainda assim essas permitem um aumento razoável de 2% a 3% na precisão. Note-se que a aplicação combinada de todas heurísticas de ajuste (linha *todas*) traz benefícios enormes como pode ser visto pelo aumento de 38% e 35% nos valores de medida-F para bigramas e trigramas, respectivamente.

3.4.2 Resultados Numéricos para as Heurísticas de Descarte

A análise das heurísticas de descarte inicia considerando os resultados já obtidos com a aplicação de todas as heurísticas de ajuste. Dessa forma, os resultados apresentados na primeira linha (*todas A*) da Tabela 3.7 consideram a aplicação de todas heurísticas de ajuste e nenhuma das heurísticas de descarte. As 4 linhas seguintes representam os resultados obtidos aplicando todas heurísticas de ajuste e cada uma das heurísticas de descarte individualmente. Finalmente, a última linha da Tabela 3.7 (*todas A D*) apresenta os resultados obtidos com todas as heurísticas de ajuste, bem como todas as heurísticas de descarte.

Observando as informações na Tabela 3.7 é possível perceber que a maior parte dos benefícios (até 10% de medida-F) das heurísticas de descarte ocorre devido à regra de recusa de SNs com símbolos (D2). A recusa de SN com numerais (D1) também causou um aumento interessante da medida-F (até 3%). Além disso, para essas duas heurísticas (D1 e D2) percebeu-se um aumento mais significativo para bigramas, enquanto que para trigramas os benefícios foram menos impactantes.

Tabela 3.7: Benefícios obtidos com as heurísticas de descarte.

Bigramas				
heurísticas de descarte	P	R	F	$\mid \mathcal{LR} \cap \mathcal{LE} \mid$
todas A	48%	55%	51%	839
D1	52%	60%	56%	914
D2	57%	65%	61%	993
D3	48%	55%	51%	842
D4	48%	55%	51%	840
todas A D	57%	65%	61%	1.001
Trigramas				
heurísticas de descarte	P	R	F	$\mid \mathcal{LR} \cap \mathcal{LE} \mid$
todas A	60%	35%	44%	934
D1	61%	36%	45%	947
D2	64%	37%	47%	995
D3	61%	36%	45%	953
D4	60%	35%	44%	936
todas A D	65%	38%	48%	1.006

As outras duas heurísticas (D3 e D4), ainda que afetando um número razoável de SNs (6.759 termos), conforme informado nas Seções 3.3.2.3 e 3.3.2.4, tiveram efeitos menores tanto na precisão como abrangência. Apesar disso, tanto para bigramas como para trigramas, as heurísticas ainda contribuíram com a recusa de termos inadequados, aumentando, portanto, o número de termos encontrados nas listas de referência.

Adicionalmente, o uso combinado de todas as heurísticas de descarte trouxe um inegável benefício na precisão das listas de 9% para bigramas e 5% para trigramas. Este aumento de precisão é ainda mais notável devido a ser acompanhado por um aumento de 10% e 3% de abrangência, para bigramas e trigramas respectivamente.

3.4.3 Resultados Numéricos para as Heurísticas de Inclusão

Analogamente à análise feita para as heurísticas de descarte, a avaliação quantitativa das heurísticas de inclusão é feita considerando a aplicação de todas heurísticas dos dois grupos anteriormente citados. A primeira linha (*todas A D*) da Tabela 3.8 apresenta os resultados obtidos com a aplicação de todas heurísticas de ajuste e descarte, e nenhuma das heurísticas de inclusão. As 3 linhas seguintes indicam os resultados com a aplicação de todas heurísticas de ajuste e descarte com cada uma das heurísticas de inclusão aplicada individualmente. Finalmente, a última linha (*todas*) indica os resultados obtidos com a aplicação de todas as 11 heurísticas propostas.

Observando os resultados da Tabela 3.8 é possível perceber que todas as heurísticas de inclusão apresentam incrementos na precisão e abrangência. Observando cada heurística de inclusão individualmente percebe-se um aumento de 1% a 2% em precisão e abrangência.

Numericamente, mesmo a aplicação das 3 heurísticas traz um incremento entre 2% e 3% para todos os índices. No entanto, cabe salientar que após a aplicação das heurísticas de ajuste e descarte os valores de precisão e abrangência já estavam altos em comparação com outras abordagens com o mesmo propósito de extração de termos [165, 127]. Dessa forma, mesmo o incremento de 1% de precisão obtido já é significativo quando se passa de uma precisão de 57% a 58%.

Tabela 3.8: Benefícios obtidos com as heurísticas de inclusão.

Bigramas				
heurísticas de inclusão	P	R	F	$\mid \mathcal{LR} \cap \mathcal{LE} \mid$
todas A D	57%	65%	61%	1.001
I1	59%	67%	63%	1.027
I2	58%	65%	61%	1.004
I3	58%	66%	62%	1.010
todas	60%	68%	64%	1.041
Trigramas				
heurísticas de inclusão	P	R	F	$\mid \mathcal{LR} \cap \mathcal{LE} \mid$
todas A D	65%	38%	48%	1.006
I1	67%	39%	50%	1.044
I2	65%	38%	48%	1.011
I3	65%	38%	48%	1.009
todas	68%	40%	50%	1.058

3.4.4 Resultado Final das Heurísticas Propostas

O benefício trazido pela aplicação das heurísticas é claro. Os resultados combinados mostram um aumento consistente que trouxe os valores de 7% a 13%, somente com o processo básico de extração, a valores entre 40% e 68%, com a aplicação de todas as heurísticas. Cabe salienar que a ordem de aplicação das heurísticas não afeta o resultado final das listas extraídas.

Outro fator importante a observar é que os resultados foram testados a partir de uma anotação linguística feita pelo *parser* PALAVRAS, considerando especificamente os sintagmas nominais. Em um trabalho anterior [124], uma outra ferramenta chamada OntoLP [165] seguindo os mesmos passos, ou seja, anotação pelo PALAVRAS e detecção de SNs, chegou a valores de precisão semelhantes aos valores iniciais sem o uso de heurísticas (cerca de 10%). Ainda que seja difícil comparar trabalhos distintos devido aos *corpora* utilizados, listas de referência e número de termos extraídos, percebe-se que a precisão obtida anteriormente ao uso das heurísticas propostas era sensivelmente inferior aos valores por volta de 60% de precisão conseguidos com o uso de todas as heurísticas.

Por essas razões, acredita-se que as heurísticas propostas são uma contribuição clara para qualificar o processo de extração automática de termos. Ainda que os testes de precisão, abrangência e medida-F tenham sido realizados somente sobre o *corpus* de Pediatria, os resultados obtidos para bigramas e trigramas foram consistentes entre si. Cabe lembrar que a razão pela qual não foram feitos mais testes, foi a inexistência de listas de referência a serem usadas como paradigma de qualidade do processo automático de extração de termos.

O processo de extração com todas heurísticas propostas aplicado aos *corpora* citados anteriormente (Seção 3.1) resultou no número de termos descritos na Tabela 3.9. Nessa tabela temos o número de termos gerados para cada *corpora* (Pediatria - PED, Modelagem estocástica - ME, Mineração de dados - MD, Processamento paralelo - PP, e Geologia - GEO) e divididos segundo o número de palavras dos termos (unigramas, bigramas, *etc.*). Essa tabela atualiza o número de termos originalmente extraídos expresso na Tabela 3.2.

Uma observação comparativa do número de termos antes e após a aplicação das heurísticas, respectivamente, Tabelas 3.2 e 3.9, mostra que o número total de termos varia pouco. No entanto, há um incremento de qualidade, pois descartou-se termos inadequados e incluiu-se termos adequados. A Figura 3.7 mostra graficamente essa variação para cada *corpus*.

Tabela 3.9: Número de termos extraídos de cada *corpora* após aplicação de heurísticas.

número de palavras	PED	ME	MD	PP	GEO
unigramas	71.327	100.425	91.370	93.433	151.755
bigramas	33.340	37.608	35.727	35.233	78.490
trigramas	27.587	43.905	45.450	43.303	71.377
4-gramas	15.555	19.905	19.212	19.354	39.625
5-gramas	10.067	16.388	17.199	15.897	28.785
6-gramas	6.973	9.893	9.683	9.612	19.877
7-gramas	4.659	7.159	7.440	6.901	13.597
8-gramas	3.186	4.700	5.013	4.756	9.493
9-gramas	2.218	3.402	3.628	3.424	6.547
N-gramas	5.208	8.783	9.717	9.232	16.855
total	180.120	252.168	244.439	241.145	436.401

Na Figura 3.7 indica-se o número de termos extraídos de cada *corpus* com cores distintas para os termos segundo o número de palavras (unigramas, bigramas, *etc.*). Mostra-se ainda nessa figura o número de termos considerando apenas o processo básico de extração sem nenhuma heurística nas colunas onde aparece apenas o nome do *corpus* (PED, ME, MD, PP e GEO), e o número de termos considerando a aplicação de todas as heurísticas propostas nas colunas onde aparece o nome do *corpus* marcado com um asterisco (PED*, ME*, MD*, PP* e GEO*).

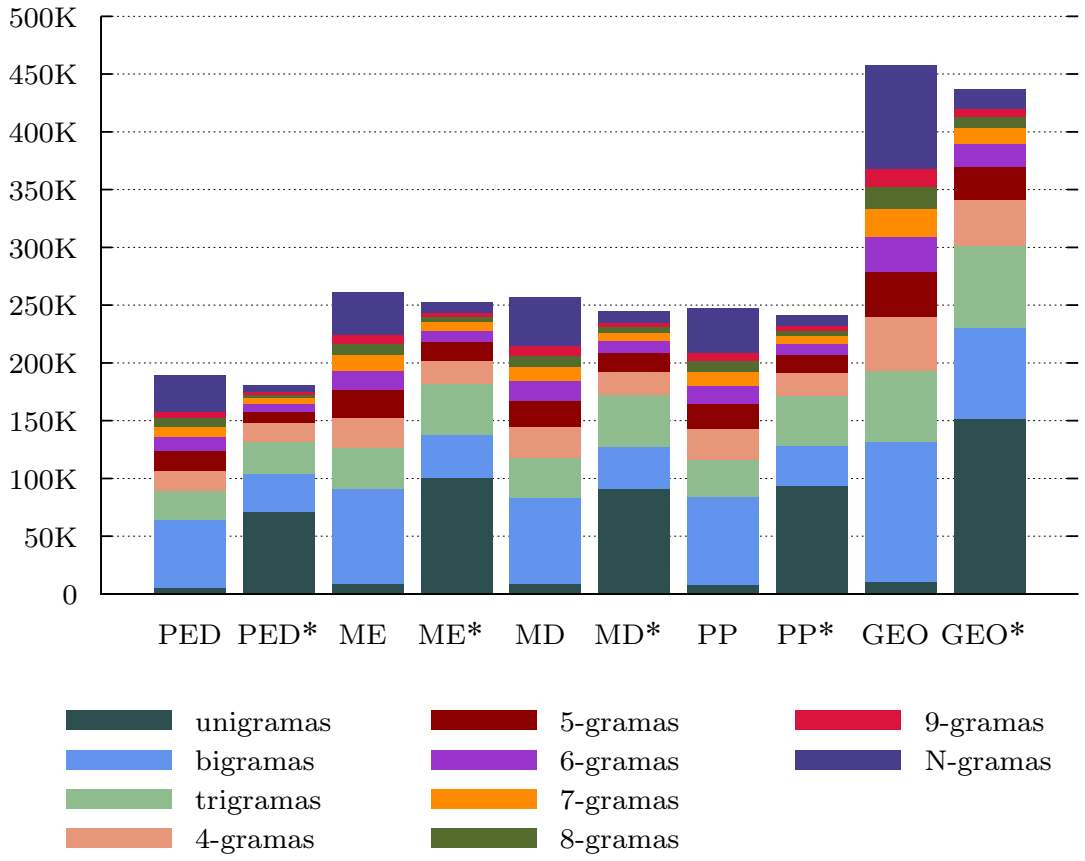


Figura 3.7: Comparativo do número de termos extraídos com a aplicação das heurísticas.

Observando a Figura 3.7, percebe-se um grande aumento no número de termos com menos palavras, especialmente unigramas, enquanto que o número de termos com muitas palavras diminui bastante. O mais interessante é que essa alteração na distribuição do número de termos acontece com um incremento de qualidade, pois descartou-se termos inadequados e incluiu-se termos adequados, como indicam os testes de precisão vistos anteriormente.

3.5 Produto Final da Extração

Como produto final da extração realizada gera-se um recurso linguístico composto por um conjunto de termos (SNs) extraídos ao qual associa-se informações contextuais que podem ser muito úteis em várias aplicações dos conceitos. Essas informações oferecem dados relevantes de cada SN extraído por si só, como sua forma original e sua forma lematizada (forma canônica), mas também informações que remetem ao contexto no qual cada termo foi encontrado, como por exemplo, a função gramatical que o SN desempenha na frase, ou o verbo ao qual o termo está relacionado.

Especificamente, para cada SN extraído associam-se as seguintes informações:

1. o termo na sua forma original;
2. o termo na sua forma canônica;
3. o número de palavras que compõem o termo (1 para unigramas, 2 para bigramas, 3 para trigramas, *etc.*);
4. a palavra indicada como núcleo do termo na sua forma canônica;
5. a etiqueta sintática do núcleo (substantivo, adjetivo, *etc.*);
6. a(s) etiqueta(s) semântica(s) do núcleo (uma estimativa feita pelo *parser*);
7. a etiqueta morfológica do núcleo (gênero, número, *etc.*);
8. a função gramatical do termo na oração (sujeito, objeto, *etc.*);
9. a posição ocupada pelo termo na frase (onde situam-se as palavras que compõem termo);
10. o número total de palavras da frase;
11. o predicado ao qual o termo exerce sua função gramatical na forma original;
12. o predicado ao qual o termo exerce sua função gramatical na forma canônica;
13. a etiqueta sintática do predicado ao qual o termo exerce a sua função gramatical;
14. a etiqueta morfológica do predicado ao qual o termo exerce a sua função gramatical;
15. a posição ocupada pelo predicado ao qual o termo exerce sua função gramatical na frase (onde situam-se as palavras que compõem o predicado);
16. um identificador da frase de onde o termo foi extraído;
17. um identificador do documento de onde o termo foi extraído.

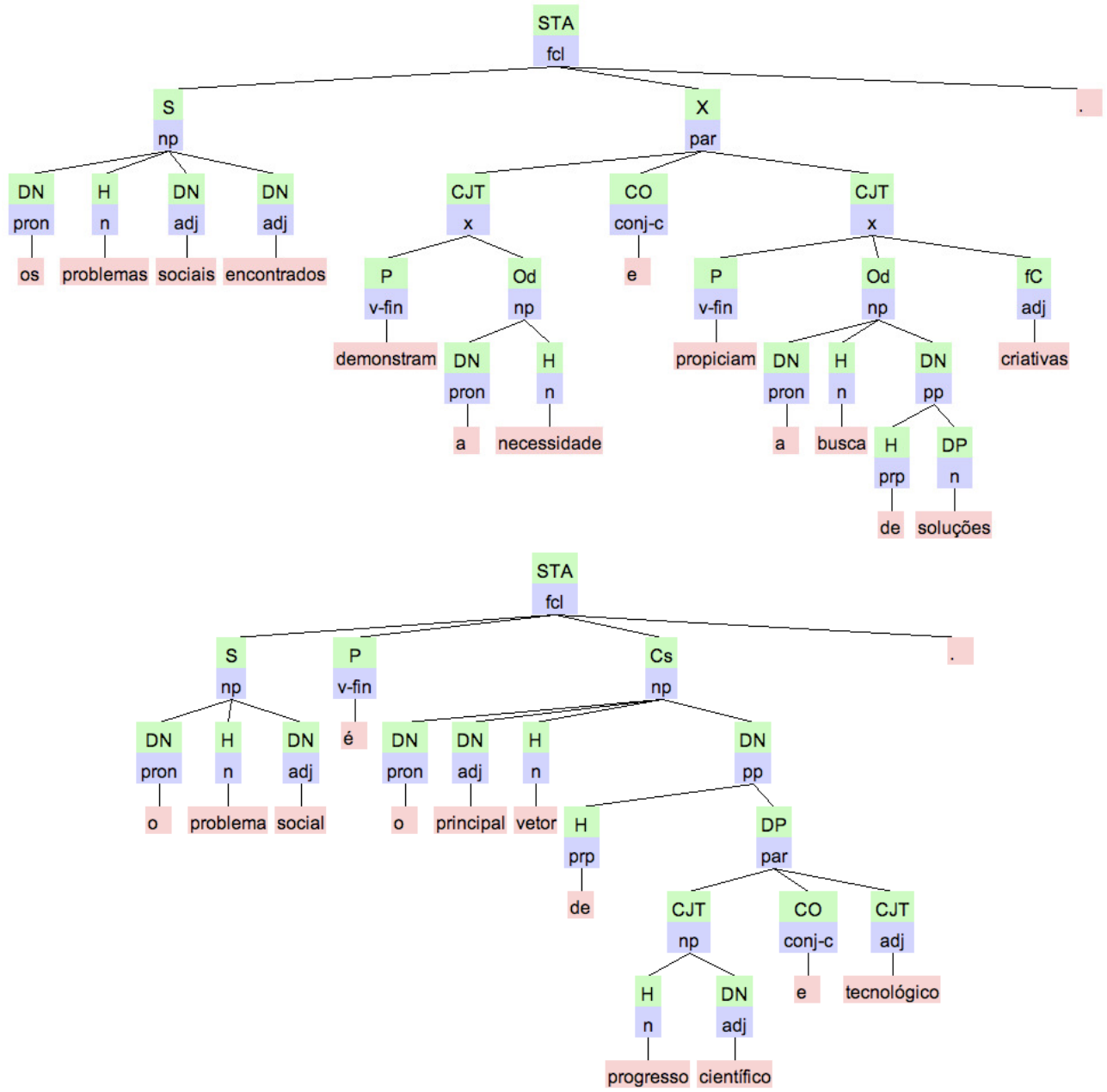


Figura 3.8: Anotação para as frases do documento exemplo *d*.

Para exemplificar o recurso linguístico disponível após a extração, considere-se um documento *d* composto pelas frases "Os problemas sociais encontrados demonstram a necessidade e propiciam a busca de soluções criativas." e "O problema social é o principal vetor de progresso científico e tecnológico.", cuja anotação linguística está descrita na Figura 3.8.

O processo de extração proposto, aplicado a esse documento, resulta nos 17 termos apresentados na Tabela 3.10 com suas respectivas informações associadas (o número em negrito identifica o campo, segundo a enumeração definida nessa seção). Uma vez extraídas essas informações, é possível transformá-las de diversas maneiras que servirão de base para as próximas etapas desenvolvidas nessa tese.

Tabela 3.10: Termos extraídos do documento exemplo com duas frases.

#	informações extraídas				
1	**1:** problemas sociais encontrados		**2:** problema social encontrar		**3:** 3
	4: problema	**5:** n	**6:** ac	**7:** M P	**8:** S
	9: 1–4	**10:** 14	**11:** demonstram	**12:** demonstrar	**13:** v-fin
	14: IND PR 3P	**15:** 5–5	**16:** 1	**17:** *d*	
2	**1:** problemas sociais encontrados		**2:** problema social encontrar		**3:** 3
	4: problema	**5:** n	**6:** ac	**7:** M P	**8:** S
	9: 1–4	**10:** 14	**11:** propiciam	**12:** propiciar	**13:** v-fin
	14: IND PR 3P	**15:** 9–9	**16:** 1	**17:** *d*	
3	**1:** problemas sociais		**2:** problema social		**3:** 2
	4: problema	**5:** n	**6:** ac	**7:** M P	**8:** S
	9: 1–3	**10:** 14	**11:** demonstram	**12:** demonstrar	**13:** v-fin
	14: IND PR 3P	**15:** 5–5	**16:** 1	**17:** *d*	
4	**1:** problemas sociais		**2:** problema social		**3:** 2
	4: problema	**5:** n	**6:** ac	**7:** M P	**8:** S
	9: 1–3	**10:** 14	**11:** propiciam	**12:** propiciar	**13:** v-fin
	14: IND PR 3P	**15:** 9–9	**16:** 1	**17:** *d*	
5	**1:** problemas		**2:** problema		**3:** 1
	4: problema	**5:** n	**6:** ac	**7:** M P	**8:** S
	9: 1–2	**10:** 14	**11:** demonstram	**12:** demonstrar	**13:** v-fin
	14: IND PR 3P	**15:** 5–5	**16:** 1	**17:** *d*	
6	**1:** problemas		**2:** problema		**3:** 1
	4: problema	**5:** n	**6:** ac	**7:** M P	**8:** S
	9: 1–2	**10:** 14	**11:** propiciam	**12:** propiciar	**13:** v-fin
	14: IND PR 3P	**15:** 9–9	**16:** 1	**17:** *d*	
7	**1:** necessidade		**2:** necessidade		**3:** 1
	4: necessidade	**5:** n	**6:** am	**7:** F S	**8:** Od
	9: 6–7	**10:** 14	**11:** demonstram	**12:** demonstrar	**13:** v-fin
	14: IND PR 3P	**15:** 5–5	**16:** 1	**17:** *d*	
8	**1:** busca de soluções		**2:** busca de solução		**3:** 3
	4: busca	**5:** n	**6:** activity	**7:** F S	**8:** Od
	9: 10–13	**10:** 14	**11:** propiciam	**12:** propiciar	**13:** v-fin
	14: IND PR 3P	**15:** 9–9	**16:** 1	**17:** *d*	
9	**1:** busca		**2:** busca		**3:** 1
	4: busca	**5:** n	**6:** activity	**7:** F S	**8:** Od
	9: 10–10	**10:** 14	**11:** propiciam	**12:** propiciar	**13:** v-fin
	14: IND PR 3P	**15:** 9–9	**16:** 1	**17:** *d*	
10	**1:** problema sociai		**2:** problema social		**3:** 2
	4: problema	**5:** n	**6:** ac	**7:** M P	**8:** S
	9: 1–3	**10:** 12	**11:** é	**12:** ser	**13:** v-fin
	14: IND PR 3P	**15:** 4–4	**16:** 2	**17:** *d*	
11	**1:** problema		**2:** problema		**3:** 2
	4: problema	**5:** n	**6:** ac	**7:** M P	**8:** S
	9: 1–2	**10:** 12	**11:** é	**12:** ser	**13:** v-fin
	14: IND PR 3P	**15:** 4–4	**16:** 2	**17:** *d*	
12	**1:** principal vetor de progresso científico		**2:** principal vetor de progresso científico		**3:** 6
	4: vetor	**5:** n	**6:** ac-sign	**7:** M P	**8:** Cs
	9: 5–10	**10:** 12	**11:** é	**12:** ser	**13:** v-fin
	14: IND PR 3P	**15:** 4–4	**16:** 2	**17:** *d*	
13	**1:** principal vetor de progresso tecnológico		**2:** principal vetor de progresso tecnológico		**3:** 6
	4: vetor	**5:** n	**6:** ac-sign	**7:** M P	**8:** Cs
	9: 5–9;12	**10:** 12	**11:** é	**12:** ser	**13:** v-fin
	14: IND PR 3P	**15:** 4–4	**16:** 2	**17:** *d*	
14	**1:** principal vetor de progresso		**2:** principal vetor de progresso		**3:** 5
	4: vetor	**5:** n	**6:** ac-sign	**7:** M P	**8:** Cs
	9: 5–8	**10:** 12	**11:** é	**12:** ser	**13:** v-fin
	14: IND PR 3P	**15:** 4–4	**16:** 2	**17:** *d*	
15	**1:** principal vetor		**2:** principal vetor		**3:** 2
	4: vetor	**5:** n	**6:** ac-sign	**7:** M P	**8:** Cs
	9: 5–7	**10:** 12	**11:** é	**12:** ser	**13:** v-fin
	14: IND PR 3P	**15:** 4–4	**16:** 2	**17:** *d*	
16	**1:** progresso científico		**2:** progresso científico		**3:** 2
	4: progresso	**5:** n	**6:** am	**7:** M S	**8:** --
	9: 9–10	**10:** 12	**11:** –	**12:** –	**13:** –
	14: –	**15:** –	**16:** 2	**17:** *d*	
17	**1:** progresso tecnológico		**2:** progresso tecnológico		**3:** 2
	4: progresso	**5:** n	**6:** am	**7:** M S	**8:** --
	9: 9;12	**10:** 12	**11:** –	**12:** –	**13:** –
	14: –	**15:** –	**16:** 2	**17:** *d*	

4. ORDENAÇÃO DE TERMOS

Após a extração dos termos exposta no capítulo anterior, o próximo passo é ordenar os termos extraídos segundo sua relevância para o domínio de interesse. Conforme discutido na introdução, ainda que técnicas linguísticas sejam utilizadas para detectar os termos, técnicas estatísticas são a base para identificar a importância de cada termo no domínio representado pelo *corpus*. Admite-se que os termos mais frequentes tendem a ser mais importantes do que os menos frequentes [183].

Cabe lembrar que, no contexto dessa tese, o que consideramos termos candidatos a conceito são fruto de uma extração linguística refinada, que fornece somente termos portadores de informação, posto que são sintagmas nominais cuidadosamente tratados pelas heurísticas descritas no capítulo anterior. Essa qualidade de termos extraídos é um fato importante no contexto da abordagem de ordenação adotada.

Dessa forma, podemos assumir que a frequência é adequada como critério de importância, ao contrário do que acontece com termos obtidos através de extração puramente estatística, em que palavras muito frequentes podem ser completamente desprovidas de informação conceitual.

As abordagens tradicionais da área de recuperação de informações que buscam índices para estimar a relevância dos termos extraídos [183, 170, 168, 146] baseiam-se na análise de um único *corpus*. Abordagens mais recentes [41, 166, 203, 148, 212, 102], porém, se valem do uso de *corpora* contrastantes, ou seja, *corpora* de outros domínios, para melhor estimar a relevância de termos no domínio de interesse.

Nesse sentido, a abordagem proposta nesse capítulo segue a linha dessas abordagens recentes propondo um novo índice que leva em consideração a frequência do termo, mas também a sua disjunção no *corpus* de domínio em relação aos *corpora* contrastantes. Por essa razão, o novo índice proposto é denominado *tf-dcf*, do inglês *term frequency, disjoint corpora frequency*, ou seja, frequência de termo, frequência de disjunção de *corpora*.

Esse capítulo, inicialmente (Seção 4.1), descreve alguns índices já existentes para a ordenação de termos segundo a relevância. Em seguida, na Seção 4.2, apresenta-se o novo índice *tf-dcf*, proposto para ordenar, segundo a relevância, os termos extraídos. Na Seção 4.3 avalia-se esse índice através de sua aplicação prática sobre o *corpus* de Pediatria descrito anteriormente (Seção 3.1) e a comparação com a aplicação de índices já existentes. A Seção 4.4 apresenta uma análise dos limites do índice proposto examinando o impacto na sua precisão em função da escolha de diferentes *corpora* contrastantes.

4.1 Índices Previamente Propostos na Literatura

Essa seção apresenta índices tradicionais para estimar a relevância de termos extraídos de um *corpus*. Especificamente, apresentam-se os seguintes índices: frequência absoluta de termo (*tf* - Seção 4.1.1); frequência de termo e frequência inversa de documento (*tf-idf* - Seção 4.1.2) segundo Manning e Schültz [131]; índice de especificidade de domínio (*tds* - Seção 4.1.3) segundo Park [148]; índice *termhood* (*thd* - Seção 4.1.4) segundo Kit e Liu [103]; e frequência de termo, frequência inversa de domínio (*TF-IDF* - Seção 4.1.5) segundo Kim *et al.* [102].

4.1.1 Frequência Absoluta de Termo - *tf*

A maneira mais direta de estimar a relevância de um termo extraído é contar sua frequência absoluta, ou seja, o número de vezes que esse termo aparece nos textos [170]. Esse índice, chamado frequência absoluta de termo, tem o apelo de ser intuitivo e fácil de calcular. Apesar disso, é necessário decidir algumas questões práticas frequentemente encontradas no processamento de textos em linguagem natural. Especificamente, nessa tese discute-se brevemente o tratamento de termos na forma canônica, de sinônimos e de anáforas. Após essas discussões, apresenta-se formalmente a definição da frequência absoluta de termos adotada nessa tese.

4.1.1.1 Tratamento de Termos na Forma Canônica

A primeira das questões relevantes para o cálculo da frequência absoluta diz respeito às variações morfológicas dos termos extraídos. Variações de número são particularmente frequentes. Por exemplo, o termo "recém-nascido" no *corpus* de Pediatria, aparece 122 vezes na sua forma singular "recém-nascido" e 177 vezes na sua forma plural "recém-nascidos".

Variações de gênero também são encontradas, como é o caso do termo "paciente hospitalizado" que aparece 6 vezes na sua forma plural masculina "pacientes hospitalizados" e 2 vezes na sua forma plural feminina "pacientes hospitalizadas", nesse mesmo *corpus* de Pediatria. Parece razoável considerar essas diferentes variações linguísticas de um termo como ocorrências do mesmo termo.

Dessa forma, os termos são comparados e computados segundo sua forma canônica, ou seja, sempre considerados no singular, masculino e infinitivo (para verbos). Por exemplo, as ocorrências dos termos "recém-nascido" (122 vezes) e "recém-nascidos" (177 vezes) são agrupadas com uma frequência absoluta de 299 ocorrências, devido a todas essas terem a mesma forma canônica: "recém-nascer".

Note-se que essa aglutinação de termos com diversos formatos, mas a mesma forma canônica, só é possível quando os recursos computacionais empregados na extração automática de alguma forma disponibilizam essa informação. No contexto do processo utilizado nessa tese, é necessário que o *parser* associe a forma canônica a cada termo extraído e a extração preserve essas informações. Na eventualidade de utilizar um *parser* que não disponibiliza forma canônica, é necessário prover a associação de termos segundo as variações, por exemplo, através de técnicas de redução a formas radicais (*stemming*) [214, 69, 145].

4.1.1.2 Tratamento de Sinônimos

A segunda questão relevante para o cálculo da frequência absoluta é a ocorrência de sinônimos. O uso de sinônimos se presta a discussão, pois, recomendações de estilo de escrita sugerem que não se escreva de forma repetitiva, ocasionando um maior uso de sinônimos. Por exemplo, os termos "rocha magmática" e "rocha ígnea", presentes no *corpus* de Geologia, representam, na imensa maioria dos contextos de utilização, o mesmo conceito. Segundo literatura especializada em geologia [184, 181] os termos "rocha magmática" e "rocha ígnea" referem-se ao mesmo tipo de rocha que é a rocha gerada por cristalização de magma.

Infelizmente, esse tipo de sinônimo é bastante difícil de ser detectado somente a partir dos textos que compõem o *corpus*. Assim, esse processo poderia ser feito com o auxílio de outros recursos linguísticos, como, por exemplo, um dicionário de sinônimos.

Um caso mais fácil de detectar durante a extração são termos como "areia marinha" e "areia de mar", pois, é possível reconhecer radicais em um adjetivo ("marinha") e inferir que o uso de um sintagma preposicional com a preposição "de" e o substantivo correspondente ("de mar") [152]. No entanto, algumas vezes, como é o caso com esses termos, o uso de jargão especializado pode invalidar essa tentativa. Por exemplo, "areias marinhas" e "areia de mar"

não são sinônimos no contexto de Geologia. Um sinônimo mais adequado para "areia marinha", nesse contexto, seria o termo "areia de praia".

Essa situação, onde sinônimos de fato são de difícil identificação, e termos onde a semelhança é mais facilmente detectável, mas não necessariamente confiável, motiva a decisão de desconsiderar a busca por sinônimos nos trabalhos dessa tese. Note-se que abandona-se a busca de sinônimos para o propósito de identificação de conceitos, mas esse fato não implica que essa busca não possa ser bem mais relevante para outros propósitos.

4.1.1.3 Tratamento de Anáforas

Uma situação semelhante aos sinônimos é o problema da identificação de anáforas[1] nas frases do *corpus*. O processo de identificação de anáforas também poderia ser extremamente útil na identificação do número total de vezes que um determinado termo está sendo referenciado em um texto, pois além das referências explícitas, as anáforas representam ocorrências implícitas do termo ao qual elas se referem. Se contabilizarmos somente as ocorrências explícitas de um termo e ignorar diversas referências que podem ter sido feitas a esses termos através de outras expressões que, muitas vezes, são empregadas por questões de estilo de escrita.

Porém, o processo de resolução (identificação) de anáforas é bastante complexo. Em contraposição a essa dificuldade, é natural assumir que o número de anáforas referenciando a cada termo seja proporcional ao seu número de ocorrências explícitas [24]. Obviamente, essa suposição não será matematicamente precisa para todos os termos, mas é razoável esperar que, em linhas gerais, ela ocorra de forma relativamente homogênea para os termos mais frequentes. Devido à dificuldade do tratamento de anáforas e à proporcionalidade no aumento de ocorrências implícitas, decidiu-se ignorar anáforas no escopo dessa tese.

4.1.1.4 Definição Formal da Frequência Absoluta de Termo

Partindo do número de ocorrências de cada termo em cada um dos documentos de um corpus c, a definição formal da frequência absoluta de um termo t é expressa por:

$$tf_t^{(c)} = \sum_{\forall\, d \in \mathcal{D}^{(c)}} tf_{t,d} \tag{4.1}$$

onde $tf_{t,d}$ é o número de ocorrências do termo t no documento d que pertence ao conjunto $\mathcal{D}^{(c)}$ de documentos que compõem o corpus c.

4.1.2 Frequência de Termo e Inversa de Documento - *tf-idf*

O uso da frequência absoluta como medida de relevância para listas obtidas com métodos puramente estatísticos é uma abordagem muito simples, que pode produzir resultados precários. Termos muito frequentes, como expressões usuais em uma língua, podem ter frequências absolutas muito altas, apesar de não possuir grande relevância para o *corpus* de domínio. Essa é a motivação do uso de "*stop lists*" que define termos (ou palavras) que devem ser desconsideradas durante o processo de extração. Na verdade, sem "*stop lists*" qualquer método puramente estatístico indica como os mais frequentes, termos sem relevância conceitual, como preposições e expressões usuais.

O uso de frequência de termos como índice de relevância é menos prejudicial para métodos de extração baseados em abordagens linguísticas. Por exemplo, a anotação sintática de um *corpus*,

[1] Uma anáfora é uma expressão que se refere a, ou substitui, outra expressão no texto [57, 110, 195, 1]. Por exemplo, sejam as frases "Os sedimentos preenchem os espaços criados pela subida relativa do nível do mar. Eles são depositados episodicamente e possuem distribuição local.". A expressão "Eles" que inicia a segunda frase é uma anáfora que refere-se à expressão "Os sedimentos" que inicia a primeira frase.

ao identificar sintagmas nominais, permite à extração evitar termos que não são adequados ao papel de conceitos, como verbos e pronomes. No entanto, até métodos sofisticados de extração, como o desenvolvido no capítulo anterior dessa tese, não evitam que termos comuns a diversos textos científicos sejam muito frequentes e, por consequência, considerados candidatos a conceitos. Por exemplo, o termo "trabalhos futuros" é muito frequente em textos científicos, mas dificilmente pode ser considerado relevante para um domínio específico.

Uma alternativa para a frequência absoluta de termo, bem conhecida na área de recuperação de informação, é considerar de maneira distinta a frequência dos termos entre os vários documentos do *corpus*. O trabalho seminal de Spärck-Jones [183] mostra a importância de considerar termos frequentes e infrequentes para a recuperação de documentos. Essas ideias levaram ao modelo probabilístico de relevância de termos para documentos de Robertson e Spärck-Jones [167].

Croft e Harper [50], e mais tarde Robertson e Walker [168], propuseram formulações para um índice que leva positivamente em consideração a frequência do termo (*tf*), *i.e.*, o número de ocorrências de um termo t em um documento d, e negativamente o número de documentos onde esse termo aparece pelo menos uma vez (*idf*). Esse índice, denominado *tf-idf* possui muitas formulações, *e.g.*, [111, 126, 131], mas nessa tese será considerado a formulação proposta por Bell *et al.* [208], por ser uma definição mais robusta que as demais citadas. O índice *tf-idf* é formalmente definido para o termo t, para cada documento d que pertence ao *corpus* c e possui pelo menos uma ocorrência de t ($\forall d \in \mathcal{D}^{(c)}$ e $tf_{t,d} > 0$), da seguinte forma:

$$tf\text{-}idf_{t,d} = \underbrace{(1 + \log(tf_{t,d}))}_{\text{parte } tf} \times \underbrace{\log\left(1 + \frac{\mid \mathcal{D}^{(c)} \mid}{\mid \mathcal{D}_t^{(c)} \mid}\right)}_{\text{parte } idf} \tag{4.2}$$

onde $tf_{t,d}$ é o número de ocorrências do termo t no documento d; $\mathcal{D}^{(c)}$ é o conjunto de todos documentos de um *corpus* c; e $\mathcal{D}_t^{(c)}$ é o subconjunto desses documentos onde t ocorre pelo menos uma vez.

Observando a equação (4.2), é possível observar as partes *tf* e *idf*. A parte *tf* considera a variação logarítmica da frequência do termo, pois a variação das ocorrências dos termos se aproxima da distribuição exponencial. Dessa forma, um termo que possui 10 ocorrências não é 10 vezes mais importante que um termo que ocorre uma única vez, mas sua relevância é uma ordem de magnitude maior. A parte *idf* corresponde a um valor numérico que varia de $\log(2)$ para um termo que aparece em todos documentos do *corpus*, até $\log(1+ \mid \mathcal{D}^{(c)} \mid)$ para um termo que aparece apenas em um documento.

A ideia por trás da fórmula do *tf-idf* é que um termo t é mais relevante para um documento d, se ele é muito frequente nesse documento, e aparece em poucos documentos, ou idealmente em um único documento. A popularidade desse índice é justificada em parte porque ele evita que termos frequentes presentes em vários documentos sejam considerados mais relevantes do que mereçam. Na verdade, *tf-idf* é um índice eficaz para identificar palavras chave, pois ele atribui relevância a termos adequados para indexação ou categorização de documentos.

O uso de *tf-idf* para estabelecer a relevância de termos para *corpus* de domínio, ao invés de pares termo-documento, foi proposta por Manning and Schütze [131, 130]. Com esse propósito, é necessário obter um índice único por termo, logo, a proposta desses autores é somar os valores de um mesmo termo para todos os documentos, de forma a obter um valor único para cada termo. De acordo com esses autores, a expressão formal desse índice para estimar a relevância de um termo t em um *corpus* c é dada por:

$$tf\text{-}idf_t^{(c)} = \sum_{\forall d \in \mathcal{D}_t^{(c)}} tf\text{-}idf_{t,d} \tag{4.3}$$

4.1.3 Índice de Especificidade de Domínio - *tds*

Além das iniciativas baseadas em analisar um único *corpus*, a comunidade científica vem propondo novas abordagens baseadas na observação de um conjunto de *corpora* que permita uma visão em perspectiva de quais termos são relevantes para um *corpus* de domínio. As primeiras iniciativas para considerar a relevância de termos em um *corpus* de domínio com o auxílio de *corpora* contrastantes inclui os trabalhos de Chung em 2003 [41] e Drouin em 2004 [55].

Entretanto, é somente com o trabalho de Park *et al.* [148], em 2008, que aparecem as primeiras definições formais de um índice para expressar a relevância de termos baseadas em *corpora* contrastantes. Nesse trabalho, um índice chamado Especificidade de Domínio de Termo (em inglês, *term domain specificity*) foi expresso como a razão entre a probabilidade de um termo t em um *corpus* de domínio c e a probabilidade desse mesmo termo em um *corpus* genérico contrastante g. Formalmente, o índice proposto por Park *et al.* é expresso por:

$$tds_t^{(c)} = \frac{p_t^{(c)}}{p_t^{(g)}} = \left.\begin{array}{l} \frac{tf_t^{(c)}}{N^{(c)}} \\ \frac{tf_t^{(g)}}{N^{(g)}} \end{array}\right\} \frac{\text{prob. no domínio } c}{\text{prob. no } \textit{corpus } g} \tag{4.4}$$

onde $p_t^{(c)}$ expressa a probabilidade de ocorrência do termo t no *corpus* c; e $N^{(c)}$ é o número total de termos nesse *corpus* c, *i.e.*, $N^{(c)} = \sum_{\forall t'} tf_{t'}^{(c)}$.

Adaptando a definição original de Park *et al.* para considerar, não um único *corpus*, mas um conjunto de *corpora* contrastantes $\mathcal{G}$, e adotando uma notação mais simples, a definição do índice de especificidade de domínio de termo t em um *corpus* de domínio c redefine-se por:

$$tds_t^{(c)} = \frac{\frac{tf_t^{(c)}}{|V^{(c)}|}}{\frac{tf_t^{(\mathcal{G})}}{|V^{(\mathcal{G})}|}} \tag{4.5}$$

onde $V^{(c)}$ corresponde ao vocabulário do *corpus* c, ou seja, todos os termos que fazem parte do *corpus* c; $V^{(\mathcal{G})}$ corresponde a união dos vocabulários de todos os *corpora* contrastantes, ou seja, todos *corpora* que pertencem a $\mathcal{G}$.

4.1.4 Índice *Termhood* - *thd*

Seguindo a mesma abordagem de *corpora* contrastantes, o trabalho de Kit e Liu, em 2008, propõe um índice denominado *termhood* [103]. Esse índice, assim como o *tds*, segue a ideia que um termo relevante para um domínio é mais frequente no *corpus* desse domínio do que em outros *corpora*. A principal diferença da proposta de Kit e Liu é que, ao invés de considerar a probabilidade de ocorrência do termo, considera-se a ordenação (*rank*) do termo no vocabulário (conjunto de todos os termos) do *corpus*. A definição formal de Kit e Liu para o índice *termhood* do termo t no *corpus* c, considerando a existência de um *corpus* contrastante g, é expressa por:

$$thd_t^{(c)} = \underbrace{\frac{r_t^{(c)}}{|\,V^{(c)}\,|}}_{\substack{\text{valor de rank} \\ \text{norm. para } c}} - \underbrace{\frac{r_t^{(g)}}{|\,V^{(g)}\,|}}_{\substack{\text{valor de rank} \\ \text{norm. para } g}} \tag{4.6}$$

onde $|\,V^{(c)}\,|$ é o tamanho do vocabulário de c; e $r_t^{(c)}$ é o valor de ordenação (*rank*) do termo t no *corpus* c. O valor de $r_t^{(c)}$ é definido como $|\,V^{(c)}\,|$ para o termo mais frequente do *corpus* c. Para o segundo termo mais frequente, o valor de $r_t^{(c)}$ é igual a $|\,V^{(c)}\,| - 1$, e assim por diante até $r_t^{(c)}$ igual a 1 para o termo menos frequente.

Observando o índice *termhood* (Eq. 4.6), percebe-se que ele é a diferença entre o valor de ordenação normalizado para o *corpus* de domínio c e o valor de ordenação normalizado para o *corpus* contrastante g. A normalização é feita com o objetivo de manter o valor do índice *termhood* dentro do intervalo $[-1, 1]$, pois assim o tamanho do vocabulário dos *corpora* c e g não desequilibra o valor do índice *thd*.

Intuitivamente expandindo a definição de Kit e Liu para uma situação onde exista um conjunto $\mathcal{G}$ de *corpora* contrastantes, a expressão formal do índice *termhood* é generalizada por:

$$thd_t^{(c)} = \frac{r_t^{(c)}}{\mid V^{(c)} \mid} - \frac{r_t^{(\mathcal{G})}}{\mid V^{(\mathcal{G})} \mid} \tag{4.7}$$

onde $r_t^{(\mathcal{G})}$ é o valor de ordenação do termo t para o *corpus* composto pela união de todos os *corpora* contrastantes em $\mathcal{G}$, ou seja, o termo mais frequente da união de todos *corpora* contrastantes será igual a cardinalidade da união dos vocabulários de todos os *corpora* contrastantes ($\mid V^{(\mathcal{G})} \mid$). Cabe salientar que, mesmo nessa situação com vários *corpora* contrastantes, o valor do índice *thd* ficará dentro do intervalo $[-1, 1]$.

4.1.5 Frequência de Termo e Inversa de Domínio - *TF-IDF*

Kim *et al.* [102] propuseram, em um artigo publicado em 2009, uma ideia intuitiva de índice de relevância considerando o princípio básico do índice *tf-idf*, cujo propósito original é identificar quando um termo é adequado para representar um documento específico. Dessa forma, a proposta de Kim *et al.* não propõe um índice verdadeiramente novo, mas sim faz uma releitura do índice *tf-idf* que originalmente considera a frequência de termo e frequência inversa de documento. A proposta de Kim *et al.* aplica a mesma ideia, porém considera, ao invés das ocorrências de termos em documentos individualmente, as ocorrências de termos em cada *corpus* individualmente.

A nomenclatura utilizada por Kim *et al.* utiliza as mesma letras (*TF-IDF*), porém utiliza-as para abreviar a expressão frequência de termo e frequência inversa de domínio, em inglês, *term frequency, inverse domain frequency*. Para evitar confusão com a definição original do índice *tf-idf*, o índice proposto por Kim *et al.* será escrito com letras maiúsculas.

Conforme proposto por Kim *et al.*, o índice *TF-IDF* é formalmente definido por:

$$TF\text{-}IDF_t^{(c)} = \underbrace{\frac{tf_t^{(c)}}{\mid V^{(c)} \mid}}_{\text{parte } TF} \times \underbrace{\log\left(\frac{\mid \mathcal{G}^* \mid}{\mid \mathcal{G}_t^* \mid}\right)}_{\text{parte } IDF} \tag{4.8}$$

onde $tf_t^{(c)}$ é a frequência absoluta do termo t no *corpus* c; $\mathcal{G}^*$ é o conjunto de todos os *corpora* contrastantes e o *corpus* c; e $\mathcal{G}_t^*$ é o subconjunto de $\mathcal{G}^*$ onde o termo t aparece pelo menos uma vez.

Cabe salientar que a definição formal do índice *tf-idf* usada como inspiração da proposta feita por Kim *et al.* não é tão robusta como a proposta por Bell *et al.* (Eq. 4.3). Por exemplo, se um termo t aparece em todos *corpora*, a parte *IDF* da equação 4.8 será igual a 0, portanto, o valor do índice *TF-IDF* para o termo t será igual a 0, ou seja, o termo t será considerado menos relevante do que qualquer um dos demais, independente do número de vezes que ele possa ocorrer.

Outra diferença significativa entre as equações 4.3 e 4.8, ocorre na parte *tf*. A formulação de Bell *et al.*(Eq. 4.3) utiliza o logaritmo da frequência absoluta, enquanto Kim *et al.*(Eq. 4.8) considera diretamente a frequência relativa de termo.

4.2 Proposta de um Novo Índice de Relevância

O objetivo de todos os índices apresentados na seção anterior é obter, para cada termo, um valor numérico diretamente proporcional a sua relevância no domínio. Dessa forma, ordenando os termos segundo os índices é possível descobrir quais deles são os mais relevantes dentre os extraídos do *corpus*. Aplicações da área de engenharia de conhecimento [186, 169], como a extração de termos candidatos a conceitos de uma ontologia, podem, então, se valer desses índices de relevância.

A frequência absoluta de termo (Eq. 4.1), obviamente indica relevância, pois um termo que é muito frequente será provavelmente relevante para o domínio. Da mesma forma, o índice *tf-idf* (equação 4.3) pode ser visto como uma alternativa para indicar a relevância, pois ele permite detectar termos que são típicos de documentos do *corpus* de domínio, ou seja, palavras chaves para indexar os documentos.

No entanto, os índices *tds* (Eq. 4.5), *thd* (Eq. 4.7) e *TF-IDF* (Eq. 4.8) possuem um diferencial para indicar relevância de termos, pois eles permitem uma observação dos termos do domínio de interesse em perspectiva com a observação de *corpora* de outros domínios. Apesar disso, esses índices que utilizam *corpora* contrastantes possuem particularidades bastante distintas, que revelam iniciativas empíricas de contornar o problema de ordenação de termos segundo a relevância. A Tabela 4.1 sumariza essas diferenças que são, na sequência, discutidas em detalhe.

Tabela 4.1: Comparação teórica entre os índices que utilizam *corpora* contrastantes.

índice (equação)	fórmula	ocorrências no domínio (indicação primária de relevância)	ocorrências nos *corpora* contrastantes (mecanismo de recompensa/penalização)	
tds (4.5)	$\frac{\frac{tf_t^{(c)}}{\lvert V^{(c)}\rvert}}{\frac{tf_t^{(\mathcal{G})}}{\lvert V^{(\mathcal{G})}\rvert}}$	probabilidade de ocorrência no *corpus* (frequência relativa de termo)	penaliza	divide pela probabilidade de ocorrência na união dos *corpora* contrastantes
thd (4.7)	$\frac{r_t^{(c)}}{\lvert V^{(c)}\rvert} - \frac{r_t^{(\mathcal{G})}}{\lvert V^{(\mathcal{G})}\rvert}$	valor de ordenação (*rank*) normalizado no *corpus*	penaliza	subtrai pelo valor de ordenação (*rank*) normalizado na união dos *corpora* contrastantes
TF-IDF (4.8)	$\frac{tf_t^{(c)}}{\lvert V^{(c)}\rvert} \times \log\left(\frac{\lvert\mathcal{G}^*\rvert}{\lvert\mathcal{G}_t^*\rvert}\right)$	frequência relativa de termo	recompensa	multiplica pelo log do número total de *corpora* divido pelo número de *corpora* onde o termo aparece

A primeira diferença entre eles é a forma como esses índices consideram as ocorrências de termos no *corpus* de domínio. Os índices *tds* (Eq. 4.5) e *TF-IDF* (Eq. 4.8) calculam uma frequência relativa de termo, pois a probabilidade de termo ($p_t^{(c)}$) do índice *tds* e a parte *tf* do índice *TF-IDF* são calculadas com a frequência absoluta dividida pelo número total de termos no *corpus* de domínio.

O índice *thd* (Eq. 4.7), porém, calcula um valor de ordenação (*rank*) normalizado que, ainda que seja função da frequência absoluta, fornece uma relação linear entre os termos. Cabe salientar que a distribuição dos valores de frequência absoluta tende a seguir uma lei de Zipf [218], *i.e.*, o termo mais frequente tende a ter o dobro de ocorrência do segundo mais frequente, o triplo de ocorrências que o terceiro mais frequente, e assim por diante. Eventualmente, de acordo com a língua escolhida a distribuição dos termos pode não seguir uma distribuição de acordo com a lei de Zipf, mas para os propósitos da análise feita nesse capítulo a distribuição permanece equivalente [76].

A segunda diferença consiste na forma como ocorrências nos *corpora* contrastantes afetam o valor numérico do índice. O índice *tds* (Eq. 4.5) penaliza termos que ocorrem nos *corpora* contrastantes dividindo a probabilidade de ocorrência no *corpus* de domínio pela probabilidade

no conjunto de *corpora* contrastantes. O índice *thd* (Eq. 4.7) também penaliza termos que são encontrados nos *corpora* contrastantes, mas nesse caso, é subtraído o valor de ordenação normalizado no *corpus* de domínio pelo valor equivalente nos *corpora* contrastantes.

Por outro lado, a abordagem utilizada no índice *TF-IDF* (Eq. 4.8) segue uma outra ideia ao recompensar termos que aparecem apenas no *corpus* de domínio, através da multiplicação da parte *tf* pelo logaritmo do número total de *corpora*. Essa recompensa atribuída pelo índice *TF-IDF* (Eq. 4.8) vai decaindo conforme o termo aparece em um número maior de *corpora* contrastantes, até cair para 0 quando o termo aparece em todos *corpora*. Cabe salientar que apesar do valor da recompensa atribuída decair proporcionalmente ao número de *corpora* onde o termo aparece, a recompensa não depende do número total de ocorrências do termo nos *corpora* contrastantes.

Tendo essas questões em mente, propõe-se um novo índice para estimar a relevância de termos para um domínio, seguindo a ideia geral de observar *corpora* contrastantes. No entanto, esse novo índice se distingue dos demais pela forma como são consideradas as ocorrências de um termo no *corpus* de domínio, e, principalmente, na forma como as ocorrências do termo em *corpora* contrastantes afetam numericamente o índice. Especificamente, propõe-se modelar o efeito de ocorrências de um termo em *corpora* contrastantes com um mecanismo chamado Frequência de Disjunção de *Corpora* (em inglês: *disjoint corpora frequency - dcf*), que é uma forma matemática de penalizar um termo proporcionalmente ao número de *corpora* contrastantes em que ele aparece, e também ao número de ocorrências desse termo em cada um desses *corpora*.

4.2.1 Frequência de Termo e Disjunção de *Corpora* - *tf-dcf*

A proposta feita nessa tese, assim como outras iniciativas com *corpora* contrastantes, baseia-se em uma indicação primária de relevância de termo (devido a ocorrências no *corpus* de domínio) e de um mecanismo de recompensa/penalização (devido a ocorrência em *corpora* contratantes). A base do índice *tf-dcf* é considerar a frequência absoluta de termo como indicação primária da relevância de um termo. Em seguida, escolhe-se penalizar termos que aparecem nos *corpora* contrastantes dividindo a frequência absoluta do termo no *corpus* de domínio pela composição geométrica da sua frequência absoluta em cada um dos *corpora* contrastantes. A definição formal do índice *tf-dcf*, para o termo t no *corpus* c, considerando um conjunto de *corpora* contrastantes $\mathcal{G}$, é:

$$\textit{tf-dcf}_t^{(c)} = \frac{tf_t^{(c)}}{\prod\limits_{\forall\, g \,\in\, \mathcal{G}} 1 + \log\left(1 + tf_t^{(g)}\right)} \tag{4.9}$$

A escolha da frequência absoluta como indicação primária da relevância do termo t no *corpus* c, ao invés da frequência relativa (como *tds* e *TF-IDF*), ou *rank* (como *thd*), visa manter a simplicidade do índice por duas razões principais:

- Acredita-se que não existe a necessidade de linearização, como o uso de *rank* no índice *thd*, nem existe a necessidade de normalizar o valor pelo tamanho do *corpus*, como nos índices *tds* e *TF-IDF*, na verdade, se desejado, qualquer normalização permanece possível após o cálculo do índice *tf-dcf*;

- Acredita-se que, manter uma relação numérica direta do índice *tf-dcf* com a frequência absoluta (*tf*), preserva uma interpretação intuitiva dos valores do índice, ou seja, o valor numérico do *tf-dcf* será igual ao valor de *tf*, caso o termo não ocorra nos *corpora* contrastantes, ou inferior ao valor de *tf*, caso o termo ocorra nos *corpora* contrastantes.

A composição geométrica das frequências absolutas de um termo nos *corpora* contrastantes foi escolhida para expressar a penalização aos termos que não são exclusivos ao *corpus* de domínio. Essa penalização se materializa pela divisão expressa na equação 4.9, que tenta abranger as seguintes premissas:

- O número de ocorrências de um termo em cada *corpora* contrastante se distribui segundo a lei de Zipf [218] ou outra lei com comportamento semelhante [76][2], logo para estimar corretamente a influência das ocorrências nos outros *corpora* é necessário linearizar esse número de ocorrências;

- Um termo que aparece somente no *corpus* de domínio não deve ser penalizado, ou seja, termos que não ocorrem nos *corpora* contrastantes devem ter o divisor da equação 4.9 igual a 1, ou seja, o valor de *tf-dcf* será igual ao valor de *tf*; e

- Um termo que ocorre em vários *corpora* contrastantes tende a ser menos relevante do que se ele ocorresse em poucos *corpora*.

Devido à primeira premissa, decide-se utilizar uma função logarítmica da frequência absoluta do termo em cada *corpora* contrastante ($tf_t^{(g)}$). Essa decisão segue o mesmo princípio adotado na proposição original do índice *tf-idf* feita por Robertson and Spärck-Jones [167].

A segunda premissa motivou uma adaptação na função logarítmica com a adição do valor 1 dentro e fora da função logarítmica, para retornar um valor 1 quando o número de ocorrências de um termo nos *corpora* contrastantes é igual a 0. Essa adaptação segue o mesmo princípio adotado por Bell *et al.* [208] na sua definição formal do índice *tf-idf* (Eq. 4.2).

Finalmente, a terceira premissa leva ao uso do produto do logaritmo das ocorrências em cada *corpora* contrastante. O produto representa que a importância das ocorrências deverá crescer geometricamente, conforme o termo ocorra em diversos *corpora* contrastantes. A definição formal, proposta na equação 4.9, faz com que um termo seja menos relevante para o *corpus* de domínio, caso ele ocorra poucas vezes em muitos *corpora* contrastantes, do que se ele ocorresse muitas vezes em poucos *corpora* contrastantes. Adicionalmente, o uso do produto do logarítimo das ocorrências é compatível com a intenção de que o divisor da equação 4.9 seja igual a 1 quando o termo não ocorra nos *corpora* contrastantes.

4.3 Análise Comparativa da Precisão do Índice Proposto

A exemplo da avaliação das heurísticas propostas feita no capítulo anterior (Seção 3.4), essa seção apresenta a avaliação do índice *tf-dcf* como indicador da relevância de termos extraídos. Mostra-se que os resultados obtidos com o índice proposto são superiores aos resultados obtidos com todos os demais índices da literatura apresentados nesse capítulo.

Mais uma vez retoma-se o *corpus* de Pediatria e suas listas de referência para bigramas e trigramas. Cabe salientar, porém, que ao contrário dos experimentos realizados para as heurísticas propostas, o processo aqui desenvolvido para avaliar o índice *tf-dcf* proposto é independente da língua, ou mesmo do processo de extração que disponibilizou os termos extraídos.

Os testes realizados a seguir utilizaram o *corpus* de Pediatria (PED) como *corpus* de domínio, e os demais *corpora*, apresentados anteriormente na Seção 3.1, como *corpora* contrastantes (Modelagem estocástica - ME, Mineração de dados - MD, Processamento paralelo - PP, e Geologia - GEO). Todos esses *corpora* foram submetidos ao processo de extração descrito no capítulo anterior, considerando a aplicação de todas as 11 heurísticas propostas.

[2]Nesse contexto, considera-se como lei com comportamento semelhante aquelas que seguem uma progressão geométrica.

4.3.1 Processo Geral de Experimentação

Dentre os termos extraídos de todos os *corpora*, apenas os bigramas e trigramas foram mantidos, pois as listas de referência disponíveis não possuíam unigramas, nem termos com mais do que 4 palavras. Cabe lembrar que, as listas de referência foram desenvolvidas por um grupo externo, são compostas de 1.534 bigramas e 2.660 trigramas, e que estão disponíveis no anexo A dessa tese.

Conforme dito no capítulo anterior (Tabela 3.9), o processo de extração aplicado ao *corpus* de Pediatria resultou em um total de 33.340 bigramas e 27.587 trigramas, considerando termos repetidos. Porém, feita a contabilização do número de termos distintos conforme descrito na Seção 4.1.1.1, o total de bigramas e trigramas é de 15.485 e 18.172, respectivamente. Para essas listas de bigramas e trigramas extraídos calculam-se os seguintes índices:

tf a frequência absoluta (Eq. 4.1), que é a forma mais simples de estimar a relevância de termos;

tf-idf a frequência de termo e inversa de documento (Eq. 4.3) segundo a formalização feita por Bell *et al.* [208] e com agregação através de uma soma sobre os documentos do *corpus*, proposta por Manning e Schütze [131];

tds o índice de especificidade de domínio (Eq. 4.5) proposta por Park *et al.* [148];

thd o índice *termhood* (Eq. 4.7) proposto por Kit and Liu [103];

TF-IDF a frequência de termo e inversa de domínio (Eq. 4.8) proposta por Kim *et al.* [102]; e

tf-dcf a frequência de termo e disjunção de *corpora* (Eq. 4.9), proposta nesse capítulo.

Os resultados numéricos calculados foram os valores de precisão[3] (Seção 2.3.3) das listas compostas pelos primeiros termos extraídos, ordenados segundo cada um dos seis índices citados (Eq. 4.1, 4.3, 4.5, 4.7, 4.8 and 4.9). Para avaliar incrementalmente o benefício de cada índice, foram consideradas listas com os 50, 100, 150, 200, 250, 300, 350, 400, 450 e 500 primeiros termos, ordenados de acordo com cada um dos índices.

4.3.2 Análise Numérica dos Índices

Observando em detalhe alguns termos extraídos, é possível entender melhor o efeito de cada índice, e, por consequência, perceber os benefícios do índice *tf-dcf* como indicador de relevância de termos. Os dez termos mais frequentes do *corpus* de Pediatria são apresentados na Tabela 4.2. Nessa tabela mostra-se o número de ocorrências de cada termo em cada *corpus* (Pediatria - PED, Modelagem estocástica - ME, Mineração de dados MD, Processamento paralelo - PP e Geologia - GEO). Adicionalmente, a última coluna (lista de ref.) indica se o termo pertence ("IN"), ou não ("OUT"), à lista de referência.

Os mesmos dez termos mais frequentes são mostrados na Tabela 4.3, mas nessa tabela são indicados os valores de cada um dos seis índices apresentados, bem como a posição que o termo ocupa na lista ordenada segundo cada índice. Por exemplo, na terceira linha da Tabela 4.3, o termo "faixa etária", que pertence à lista de referência, tem como frequência absoluta (*tf*) o valor 234, que o faz ocupar a terceira posição na lista organizada segundo o índice *tf*. O índice *tf-idf* desse mesmo termo é igual a 169.18, que também o coloca na terceira posição da lista

[3] Ao contrário de experimentos tradicionais da área de recuperação de informação e mesmo outros experimentos realizados nessa tese, limitou-se a observação da precisão, pois o cálculo de abrangência não acrescentaria informação devido a todas as listas terem tamanhos fixos.

Tabela 4.2: Número de ocorrência de termos frequentes do *corpus* de Pediatria.

termos	PED	ME	MD	PP	GEO	lista de ref.
aleitamento materno	306	0	0	0	0	IN
recém nascido	299	0	0	0	0	IN
faixa etária	234	0	6	0	0	IN
presente estudo	188	4	1	0	67	OUT
leite materno	163	0	0	0	0	IN
idade gestacional	144	0	0	0	0	IN
ventilação mecânica	138	0	0	0	0	IN
via aérea	120	0	0	0	0	IN
pressão arterial	112	0	0	0	0	IN
sexo masculino	109	7	8	0	0	OUT

organizada por esse índice. O índice *tds* do termo "faixa etária" possui valor igual a 0.98, e este valor o coloca na 13.281^a posição na lista organizada pelo índice *tds*.

Tabela 4.3: Análise de termos frequentes do *corpus* de Pediatria.

termos (lista de ref.)	*tf* Eq. 4.1	*tf-idf* Eq. 4.3	*tds* Eq. 4.5	*thd* Eq. 4.7	*TF-IDF* Eq. 4.8	*tf-dcf* Eq. 4.9
aleitamento materno (IN)	306 1^a	199,18 1^a	1,00 1^a	1,00 1^a	0,0027 1^a	306,00 1^a
recém nascido (IN)	299 2^a	184,98 2^a	1,00 1^a	0,99 2^a	0,0027 2^a	299,00 2^a
faixa etária (IN)	234 3^a	169,18 3^a	0,98 13.281^a	0,93 4^a	0,0012 6^a	61,46 15^a
presente estudo (OUT)	188 4^a	167,78 4^a	0,73 13.429^a	0,50 42^a	0,0002 57^a	3,99 $1{,}276^a$
leite materno (IN)	163 5^a	143,23 5^a	1,00 1^a	0,94 3^a	0,0015 3^a	163,00 3^a
idade gestacional (IN)	144 6^a	135,60 7^a	1,00 1^a	0,93 5^a	0,0013 4^a	144,00 4^a
ventilação mecânica (IN)	138 7^a	140,85 6^a	1,00 1^a	0,91 6^a	0,0012 5^a	138,00 5^a
via aérea (IN)	120 8^a	132,72 8^a	1,00 1^a	0,90 7^a	0,0011 7^a	120,00 6^a
pressão arterial (IN)	112 9^a	93.27 19^a	1,00 1^a	0,88 8^a	0,0010 8^a	112,00 7^a
sexo masculino (OUT)	109 10^a	125,70 9^a	0,88 $13{,}318^a$	0,77 14^a	0,0003 35^a	6,53 543^a

Observando as diferenças de posição dos termos nas listas ordenadas pelos índices *tf* (Eq. 4.1) e *tf-idf* (Eq. 4.3), percebe-se uma semelhança muito grande. A única diferença significativa ocorre para o termo "pressão arterial" que cai da 9^a posição, segundo *tf*, para a 19^a posição, segundo *tf-idf*. No entanto, este rebaixamento não justifica-se, pois, intuitivamente, o termo "pressão arterial" não parece menos relevante que o termo "via aérea", por exemplo. Contrário a essa situação, o termo genérico "presente estudo" não é afetado no que diz respeito a sua posição, devido ao uso do índice *tf-idf*.

A observação do efeito do uso do índice *tds* (Eq. 4.5) mostra uma falta de discernimento ao atribuir valores numéricos aos termos. A forma de cálculo do índice *tds* atribui valores iguais a 1,00 para todos os termos que ocorrem somente no *corpus* de Pediatria. Dessa forma os termos que ocorram pelo menos uma vez em algum dos *corpora* contrastantes serão banidos de qualquer lista de termos relevantes, pois existem mais de 13.000 bigramas que ocorrem somente no *corpus* de Pediatria. Apesar desse problema, o índice *tds* consegue fazer uma certa diferenciação entre os termos que ocorrem nos *corpora* contrastantes. Os termos "faixa etária" ($tds = 0.98$), "sexo

masculino" ($tds = 0.88$) e "presente estudo" ($tds = 0.73$) parecem ter uma ordem de relevância compatível com a ordem do valor do índice *tds*.

A lista ordenada segundo o índice *thd* (Eq. 4.7) mostra um efeito de rebaixamento nos três termos que ocorrem nos *corpora* contrastantes, mas esse rebaixamento não é muito grande. Por exemplo, mesmo o termo "presente estudo", que é bastante frequente nos *corpora* contrastantes, cai da quarta posição (segundo *tf*) e para a 42^a posição usando o índice *thd*.

A lista organizada de acordo com o índice *TF-IDF* (Eq. 4.8) mostra um efeito mais forte do que o obtido com o índice *thd* (Eq. 4.7), pois ele é dependente do número de *corpora* contrastantes onde o termo ocorre. Como consequência, o termo "faixa etária" cai da terceira posição, segundo *tf*, para a sexta posição, segundo o índice *TF-IDF*. Já o termo "presente estudo" cai da quarta para a 57^a posição, pois esse termo ocorre em todos *corpora* contrastantes, exceto o de Geologia.

É importante lembrar que o índice *tf-dcf* proposto é o único que considera tanto o número de ocorrências fora do *corpus* de domínio (assim como os índices *tds* e *thd*), mas também o número de *corpora* contrastantes onde o termo ocorre (assim como o índice *TF-IDF*). Por essa razão, o efeito de rebaixamento causado pelo índice *tf-dcf* é o mais forte dentre os experimentados. O termo "presente estudo" sofre o maior rebaixamento, caindo da quarta posição, segundo *tf*, para a 1.276^a posição, segundo *tf-dcf*. Um pouco menos impactante é o rebaixamento do termo "sexo masculino" que cai da vigésima para a 543^a posição. Por outro lado, a queda do termo faixa etária é bem pequena, pois ele cai da terceira para a décima quinta posição.

4.3.3 Análise da Precisão dos Índices

Seguindo o processo de experimentação definido (Seção 4.3.1), a precisão obtida para listas ordenadas segundo os seis índices apresentados é exposta nas Figuras 4.1 e 4.2, para bigramas e trigramas respectivamente. Nesses resultados, utiliza-se a frequência absoluta (*tf* - Eq. 4.1), representada por uma curva com quadrados cheios (em azul), como resultado padrão, pois esse índice é a escolha elementar para estimar a relevância de termos.

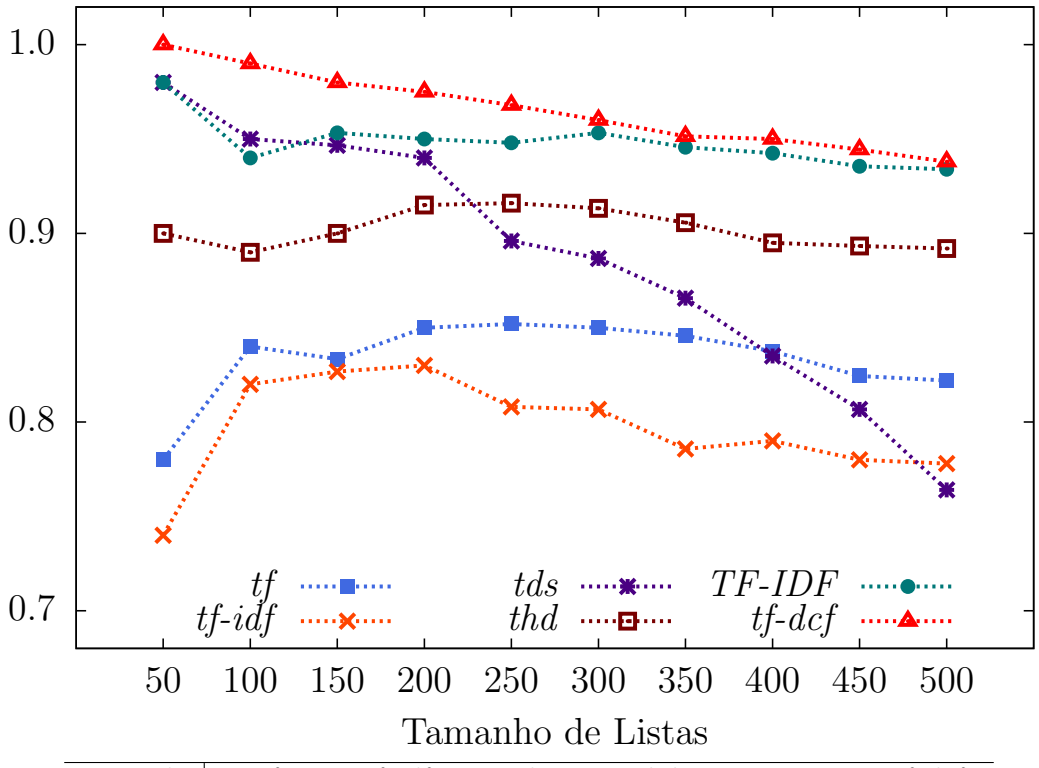
Os resultados obtidos com o uso do índice *tf-idf* (Eq. 4.3) exemplificam o esforço de estimar a relevância de termos sem utilizar *corpora* contrastantes. As demais 3 curvas (*tds* - Eq. 4.5, *thd* - Eq. 4.7, e *TF-IDF* - Eq. 4.8) representam os trabalhos da literatura que utilizam *corpora* contrastantes. Finalmente, a curva *tf-dcf*, representada com triângulos vazados (em vermelho), mostra os valores de precisão obtidos para as listas organizadas segundo o novo índice proposto nesse capítulo (Eq. 4.9).

A primeira observação para a precisão obtida com bigramas (Figura 4.1) são os baixos valores para listas ordenadas com o índice *tf-idf* (Eq. 4.3). Para listas ordenadas com a frequência absoluta (*tf* - Eq. 4.1), a precisão varia entre 78% e 85%, enquanto que para listas ordenadas com *tf-idf*, a precisão se situa entre 76% e 83%. De acordo com os tamanhos das listas, a queda de precisão resultante do uso do índice *tf-idf* frente ao índice *tf* vai de 1% (listas com 150 termos) a 6% (listas com 350 termos), mas a perda média fica em torno de 4%.

Também é fácil observar os melhores resultados alcançados com índices que usam *corpora* contrastantes, ou seja, *tds* (Eq. 4.5), *thd* (Eq. 4.7), *TF-IDF* (Eq. 4.8) e *tf-dcf* (Eq. 4.9). O ganho médio de precisão em listas organizadas por esses índices frente ao índice *tf* é de 9%, e, exceto pelo índice *tds* (Eq. 4.5) aplicado a listas com 400 ou mais termos, os valores de precisão foram sempre superiores àqueles obtidos com o índice *tf*.

Esses resultados ilustram a superioridade de índices que usam *corpora* contrastantes. No entanto, observando de perto os resultados de cada um desses índices, percebe-se comportamentos distintos.

Os resultados obtidos com o índice *tds* (Eq. 4.5) iniciam com o valor impressionante de 98% de precisão para listas de 50 termos. Porém, a precisão cai rapidamente conforme cresce



tam. de listas	*tf* Eq. 4.1	*tf-idf* Eq. 4.3	*tds* Eq. 4.5	*thd* Eq. 4.7	*TF-IDF* Eq. 4.8	*tf-dcf* Eq. 4.9
50	0.7800	0.7400	0.9800	0.9000	0.9800	1.0000
100	0.8400	0.8200	0.9500	0.8900	0.9400	0.9900
150	0.8333	0.8267	0.9467	0.9000	0.9533	0.9800
200	0.8500	0.8300	0.9400	0.9150	0.9500	0.9750
250	0.8520	0.8080	0.8960	0.9160	0.9480	0.9680
300	0.8500	0.8067	0.8867	0.9133	0.9533	0.9600
350	0.8457	0.7857	0.8657	0.9057	0.9457	0.9514
400	0.8375	0.7900	0.8350	0.8950	0.9425	0.9500
450	0.8244	0.7800	0.8067	0.8933	0.9356	0.9444
500	0.8220	0.7780	0.7640	0.8920	0.9340	0.9380

Figura 4.1: Precisão para bigramas do *corpus* de Pediatria ordenados segundo vários índices.

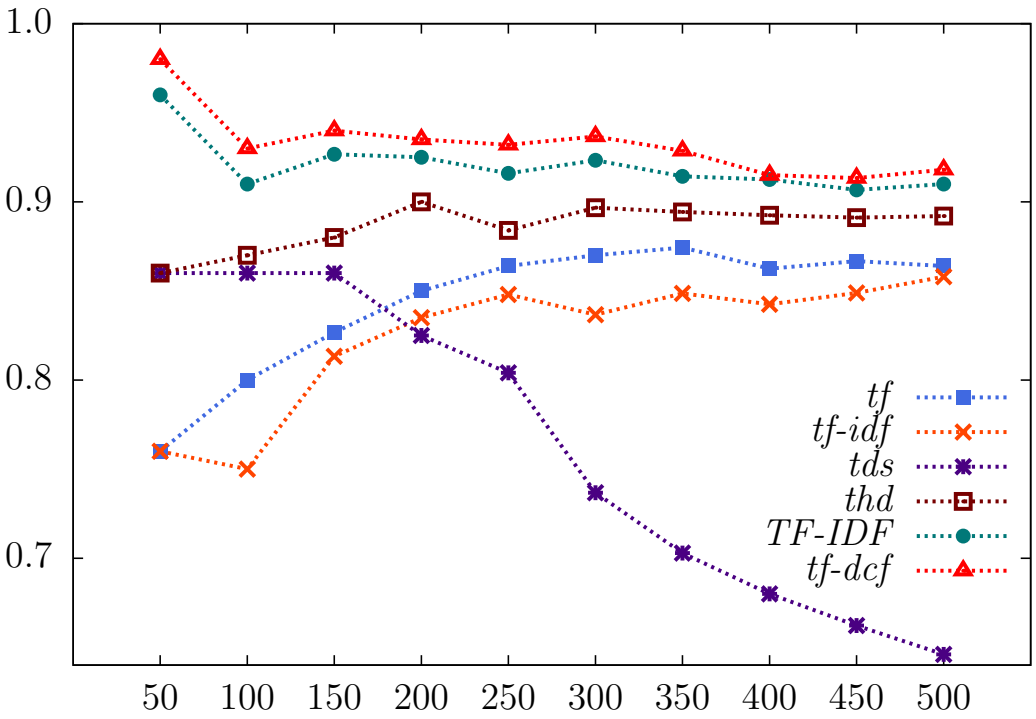
o tamanho da lista de termos. Na verdade, os valores de precisão ficam abaixo do valor padrão obtido com o índice *tf* para listas com 400, 450 e 500 termos. Esses resultados indicam que o índice de especificidade de domínio (*tds* - Eq. 4.5) não é uma opção escalável para aumentar a precisão obtida com o uso da frequência absoluta (*tf* - Eq. 4.1).

Os resultados obtidos com o índice *thd* (Eq. 4.7) apresentaram uma precisão em torno de 90% para listas de todos os tamanhos testados. O uso desse índice mostra ganhos frente ao uso do índice *tf* (Eq. 4.1) variando de 12% (listas com 50 termos) a 5% (listas com 100 termos), com uma média de 7% de ganho para todos os tamanhos de listas. Esses resultados indicam que o índice *termhood* oferece ganhos consistentes frente à frequência absoluta (*tf* - Eq. 4.1).

O uso do índice *TF-IDF* (Eq. 4.8) mostrou uma melhora significativa dos valores de precisão, variando de 98% (listas com 50 termos) a 93% (listas de 500 termos). Esses resultados representam um ganho médio de 11% frente àqueles obtidos com o índice padrão (*tf* - Eq. 4.1).

No entanto, a precisão obtida com listas organizadas pelo índice *tf-dcf* (Eq. 4.9), proposto nessa tese, são ainda mais impressionantes. Os valores de precisão resultantes do uso do índice

tf-dcf são os maiores dentre todos os experimentos realizados. Em especial, nos resultados para listas de bigramas de até 250 termos, percebe-se uma precisão nitidamente superior (2% ou mais) frente aos bons resultados obtidos com o índice *TF-IDF* (Eq. 4.8). Cabe salientar que para listas de 50 bigramas, consegue-se a precisão máxima (100%). Esse fato é reforçado pelo ganho médio de 13% obtido pelo uso do índice *tf-dcf* (Eq. 4.9) frente ao uso do índice padrão (*tf* - Eq. 4.1).



tam. de listas	*tf* Eq. 4.1	*tf-idf* Eq. 4.3	*tds* Eq. 4.5	*thd* Eq. 4.7	*TF-IDF* Eq. 4.8	*tf-dcf* Eq. 4.9
50	0.7600	0.7600	0.8600	0.8600	0.9600	0.9800
100	0.8000	0.7500	0.8600	0.8700	0.9100	0.9300
150	0.8267	0.8133	0.8600	0.8800	0.9267	0.9400
200	0.8500	0.8350	0.8250	0.9000	0.9250	0.9350
250	0.8640	0.8480	0.8040	0.8840	0.9160	0.9320
300	0.8700	0.8367	0.7367	0.8967	0.9233	0.9367
350	0.8743	0.8486	0.7029	0.8943	0.9143	0.9286
400	0.8625	0.8425	0.6800	0.8925	0.9125	0.9150
450	0.8667	0.8489	0.6622	0.8911	0.9067	0.9133
500	0.8640	0.8580	0.6460	0.8920	0.9100	0.9180

Figura 4.2: Precisão para trigramas do *corpus* de Pediatria ordenados segundo vários índices.

Os resultados obtidos para listas de trigramas (Figura 4.2) mostram um comportamento similar ao encontrado para listas de bigramas (Figura 4.1). Os resultados para listas ordenadas com o índice *tf-idf* (Eq. 4.3) são mais uma vez claramente abaixo dos valores obtidos com o índice padrão (*tf* - Eq. 4.1).

Os resultados para listas organizadas segundo índices que usam *corpora* contrastantes continuam sendo, em geral, superiores aos resultados utilizando o índice padrão. No entanto, percebe-se que o índice *tds* (Eq. 4.5) aplicado a trigramas mostra uma curva de precisão que cai um pouco mais rápido do que sua similar para bigramas. A precisão de listas com 200 ou

mais trigramas apresenta valores inferiores aos obtidos com a frequência absoluta (*tf* - Eq. 4.1).

A observação mais importante, porém, é que, também para os resultados da Figura 4.2, a precisão obtida para listas ordenadas com o índice *tf-dcf* (Eq. 4.9), proposto nesse capítulo, é superior aos valores de precisão de todos os demais índices. Esses resultados obtidos para listas de trigramas confirmam a impressão causada pelo sucesso obtido com as listas de bigramas (Figura 4.1).

4.4 Impacto da Escolha dos *Corpora* Contrastantes

Todos resultados apresentados até agora consideram a ordenação feita com o uso do *corpus* de Pediatria (PED) e de todos os demais *corpora* como contrastantes (Modelagem estocástica - ME, Mineração de dados - MD, Processamento paralelo - PP, e Geologia - GEO). Naturalmente, a escolha dos *corpora* contrastantes pode afetar a eficiência do índice *tf-dcf*. Logo, nessa seção são feitos três experimentos adicionais, variando o conjunto de *corpora* contrastantes conforme apresentado na Tabela 4.4.

Tabela 4.4: Experimentos com diferentes conjuntos de *corpora* contrastantes.

	corpus de domínio	*corpora* contrastantes
Experimento Original	PED	ME MD PP GEO
Experimento 1	PED	ME MD PP
Experimento 2	PED	GEO
Experimento 3	PED	PP

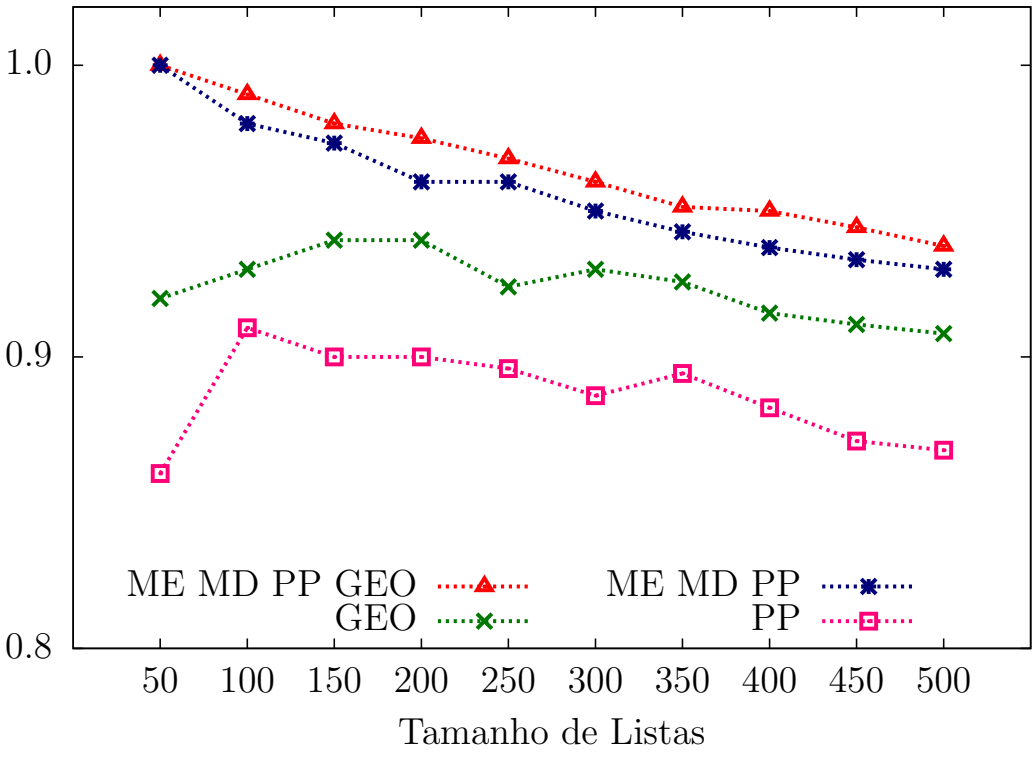
O primeiro experimento adicional (Exp. 1) corresponde à remoção do *corpus* de Geologia, mantendo apenas os *corpora* relacionados à Ciência da Computação (ME, MD e PP) como *corpora* contrastantes. O segundo experimento adicional (Exp. 2) corresponde ao complemento do Experimento 1, pois nele remove-se todos *corpora* relacionados à Ciência da Computação e mantém-se apenas o *corpus* de Geologia como *corpus* contrastante. Finalmente, o último experimento adicional (Exp. 3) também usa um único *corpus* contrastante, porém nesse caso utiliza-se apenas o pequeno *corpus* de Processamento paralelo.

A Figura 4.3 mostra os resultados de precisão para listas de 50 a 500 bigramas, de forma análoga aos experimentos feitos na Seção 4.3.3. Na verdade, os valores correspondentes à primeira curva (ME MD PP GEO) correspondem ao experimento original, ou seja, a curva referente ao índice *tf-dcf* apresentada na Figura 4.1.

Uma observação importante dos resultados apresentados na Figura 4.3 é que a remoção do *corpus* de Geologia (Exp. 1) causa uma pequena redução nos valores de precisão. Esse resultado é esperado, pois os três *corpora* relacionados à Ciência da Computação ainda fornecem uma boa comparação para o *corpus* de Pediatria.

O uso do *corpus* de Geologia como único contrastante (Exp. 2) reduz bem mais os valores de precisão. Essa redução poderia ser explicada por uma distância conceitual existente entre os *corpora* de Pediatria e Geologia, mas provavelmente, a redução possa ser consequência do tamanho do *corpus* de Geologia. Essa afirmação é consistente com o último experimento realizado (Exp. 3), onde o uso de um *corpus* ainda menor, produziu os mais baixos valores de precisão.

Reproduzindo os experimentos adicionais para os trigramas do *corpus* de Pediatria, a Figura 4.4 apresenta o efeito da variação dos *corpora* contrastantes (Tabela 4.4). Os resultados dessa figura são ligeiramente menos claros do que aqueles dos bigramas (Figura 4.3). Ainda assim, percebe-se o mesmo comportamento de redução de precisão, ou seja, a retirada do *corpus* de Geologia reduz um pouco da precisão, a retirada dos *corpora* de Ciência de Computação

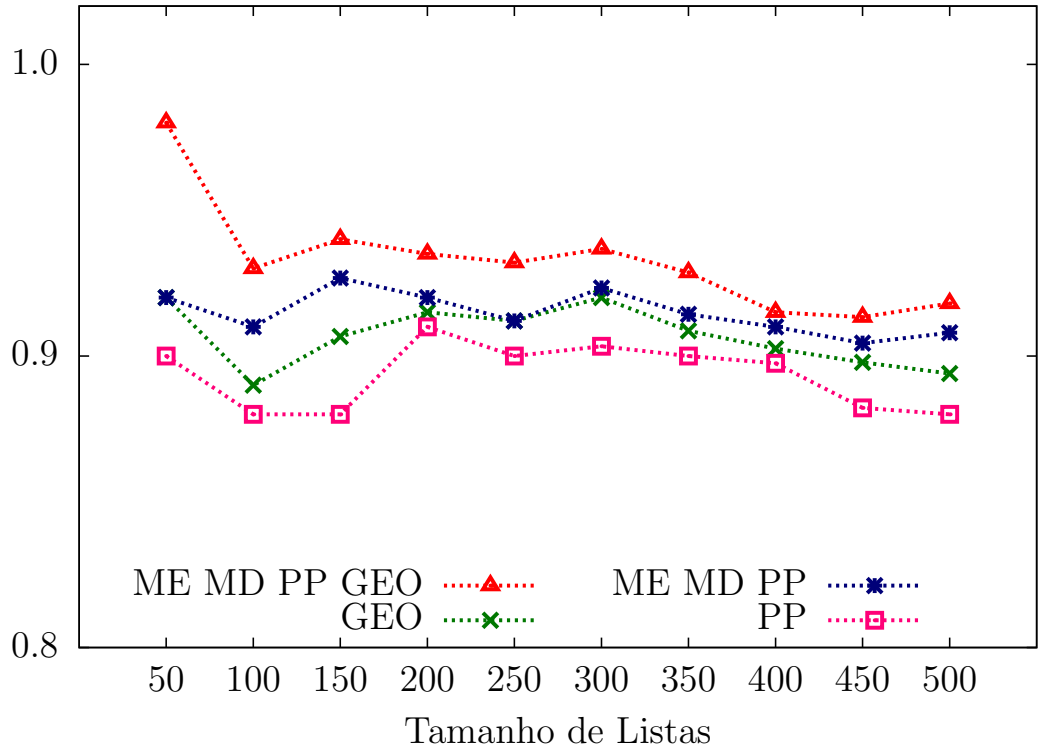


tam. de listas	ME MD PP GEO Experimento Original	ME MD PP Exp. 1	GEO Exp. 2	PP Exp. 3
50	1.0000	1.0000	0.9200	0.8600
100	0.9900	0.9800	0.9300	0.9100
150	0.9800	0.9733	0.9400	0.9000
200	0.9750	0.9600	0.9400	0.9000
250	0.9680	0.9600	0.9240	0.8960
300	0.9600	0.9500	0.9300	0.8867
350	0.9514	0.9429	0.9257	0.8943
400	0.9500	0.9375	0.9150	0.8825
450	0.9444	0.9333	0.9111	0.8711
500	0.9380	0.9300	0.9080	0.8680

Figura 4.3: Precisão para bigramas do *corpus* de Pediatria ordenados pelo índice *tf-dcf* usando diferentes conjuntos de *corpora* contrastantes.

reduz um pouco mais, e, finalmente, o pequeno *corpus* de Processamento paralelo apresenta os menores valores de precisão.

Os resultados das Figuras 4.3 e 4.4 mostram uma tendência que segue a intuição que a variabilidade dos *corpora* contrastantes é relevante. O uso de vários grandes *corpora* cobrindo diferentes domínios parece trazer uma vantagem considerável para o índice *tf-dcf*. Na verdade, o sucesso da abordagem *tf-dcf* é bastante dependente do uso de *corpora* contrastantes grandes e com domínios tão ortogonais quanto possível, como é o caso dos *corpora* utilizados nessa tese.



tam. de listas	ME MD PP GEO Experimento Original	ME MD PP Experimento 1	GEO Experimento 2	PP Experimento 3
50	0.9800	0.9200	0.9200	0.9000
100	0.9300	0.9100	0.8900	0.8800
150	0.9400	0.9267	0.9067	0.8800
200	0.9350	0.9200	0.9150	0.9100
250	0.9320	0.9120	0.9120	0.9000
300	0.9367	0.9233	0.9200	0.9033
350	0.9286	0.9143	0.9086	0.9000
400	0.9150	0.9100	0.9025	0.8975
450	0.9133	0.9044	0.8978	0.8822
500	0.9180	0.9080	0.8940	0.8800

Figura 4.4: Precisão para trigramas do *corpus* de Pediatria ordenados pelo índice *tf-dcf* usando diferentes conjuntos de *corpora* contrastantes.

5. IDENTIFICAÇÃO DE CONCEITOS

Uma vez estabelecido o índice para ordenar termos, o índice *tf-dcf*, assume-se essa ordenação como uma expressão da relevância dos termos no *corpus* de domínio. Cabe salientar, que as listas geradas tendem a ser bastante extensas. Apesar disso, devido à ordenação feita, existe a clara expectativa de que os termos mais relevantes estejam mais concentrados nas primeiras posições.

Dessa forma, deve buscar-se um ponto da lista que maximize a densidade de termos relevantes acima, e minimize o número de termos relevantes abaixo, ou seja, definir um ponto de corte que equilibre a precisão e a abrangência. O trabalho a fazer, então, é definir pontos de corte adequados para escolher, automaticamente, quais termos considerar ou não conceitos do domínio. Nesse sentido, conforme discutido no referencial teórico (Seção 2.3.2), assume-se no contexto dessa tese que os conceitos são os termos mais relevantes do domínio, e a relevância é definida segundo o índice *tf-dcf* que apresentou a melhor precisão nos experimentos do capítulo anterior.

O índice *tf-dcf* serve para ordenar os termos extraídos, ou seja, a aplicação do ponto de corte vai apenas indicar quais termos serão desprezados. Portanto, o objetivo desse capítulo é definir uma forma de escolher e aplicar pontos de corte às listas de termos extraídos, identificando aqueles que serão considerados conceitos do domínio.

Nesse sentido, esse capítulo analisa o comportamento de diversas políticas de escolha de pontos de corte aplicados sobre listas devidamente ordenadas segundo o índice *tf-dcf* proposto no capítulo anterior. Especificamente, são vistos pontos de corte tradicionalmente encontrados na literatura (Seção 5.1): os pontos de corte absolutos, os pontos de corte por limiares e pontos de corte relativos.

Em seguida, na Seção 5.2 é proposta uma forma de escolher automaticamente pontos de corte para listas de termos extraídos e ordenados. Finalmente, sumariza-se o resultado da aplicação de pontos de corte a todas as listas de termos extraídos de todos os *corpora* utilizados nessa tese. Uma parte das contribuições relativas às políticas de pontos de corte apresentadas nesse capítulo foi originalmente publicada no *Journal of the Brazilian Computer Society – JBCS/Springer* em Novembro de 2010 [123].

5.1 Pontos de Corte Tradicionais

A maneira mais simples de se aplicar pontos de corte é escolher um número arbitrário de termos que serão considerados. Vários trabalhos da literatura definem arbitrariamente pontos de corte de forma empírica [147, 28, 138, 202, 119, 7]. Essa escolha pode ser feita de várias formas. Por exemplo, as diversas curvas apresentadas no capítulo anterior (Figuras 4.1, 4.2, 4.3 e 4.4) mostram resultados obtidos para listas com tamanhos arbitrários, ou seja, com a aplicação de pontos de corte absolutos.

Nesses resultados do capítulo anterior, e em muitos trabalhos da literatura [216, 48, 124, 62, 215, 16, 54], as listas geradas separam os termos segundo o número de palavras que os compõem, ou seja, trata-se separadamente listas de unigramas, bigramas, trigramas, *etc.* Essa análise em separado faz sentido, uma vez que os termos tendem a apresentar variações distintas para os índices, segundo o número de palavras que os compõem.

Por exemplo, ordenando os unigramas do *corpus* de Geologia (Seção 3.1) segundo a frequência absoluta de termo (tf), o termo "topo" ocupa a 60^a posição com 433 ocorrências. Porém, dentre os demais termos (bigramas, trigramas, *etc.*), o termo mais frequente ("matéria orgânica") ocorre 430 vezes. Portanto, uma lista dos 60 termos mais frequentes, que não distingue os termos pelo número de palavras, será composta apenas por unigramas. Dessa forma, o estudo de pontos de corte será feito escolhendo um ponto de corte para unigramas, outro para bigramas, e assim por diante.

5.1.1 Pontos de Corte Absolutos

O *corpus* de Pediatria descrito anteriormente (Seção 3.1) possui uma lista de termos de referência desenvolvida por um grupo externo, composta por 1.534 bigramas e 2.660 trigramas considerados conceitos desse domínio (vide anexo A). Nesse sentido, a primeira experiência feita consiste em extrair os termos do *corpus* de Pediatria, organizá-los segundo o índice de relevância *tf-dcf* proposto no capítulo anterior (Eq. 4.9), e aplicar pontos de corte absolutos às listas de bigramas e trigramas. Para cada uma das listas organizadas e reduzidas pela aplicação dos pontos de corte, calcula-se a precisão, abrangência e medida F para listas obtidas com diversos pontos de corte absolutos.

Aplicam-se pontos de corte considerando os 100 termos mais frequentes, ou seja, os primeiros 100 termos dessas listas. Sucessivamente, analisam-se pontos de corte considerando os 200, 300, e assim por diante até 3.500 termos mais frequentes. A Figura 5.1 apresenta os tamanhos de listas conforme os pontos de corte, além de valores de precisão (P), abrangência (R) e medida F (F) obtidos para cada uma das listas de bigramas e trigramas. A tabela contida nessa figura indica também o número de termos encontrados na intersecção entre a lista de termos extraídos ($\mathcal{LE}$) e lista de referência ($\mathcal{LR}$).

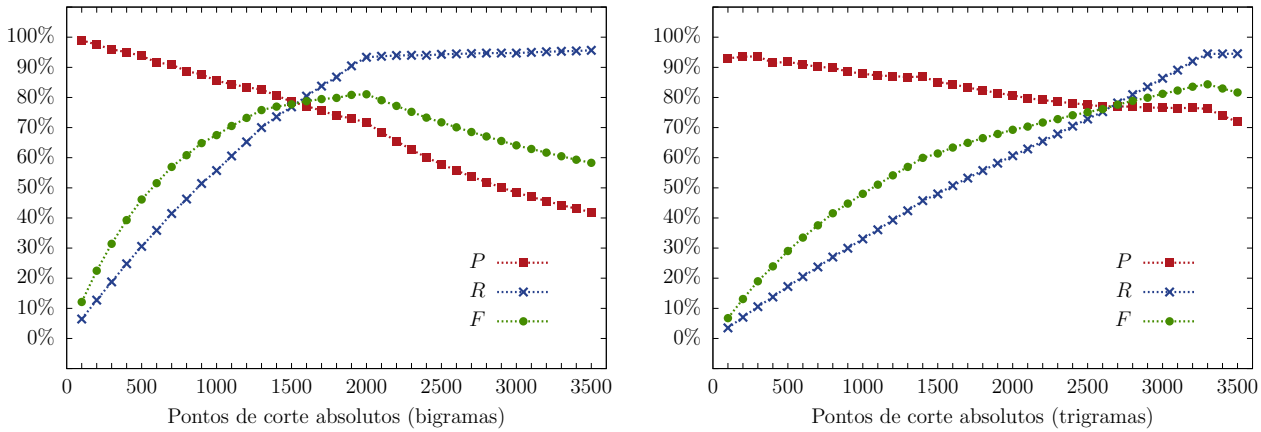
A observação dos resultados apresentados na Figura 5.1 mostra inicialmente que existe um ponto de cruzamento entre as curvas de precisão e abrangência. Observando as listas obtidas com aplicação de pontos de corte crescentes, esse ponto de cruzamento indica quando uma lista deixa de ser excessivamente restritiva. Esse ponto acontece para listas de 1.600 bigramas e 2.700 trigramas. Não por acaso, esses são valores próximos ao tamanho das listas de referência (1.534 bigramas e 2.660 trigramas), pois listas com menos termos do que a referência tem obrigatoriamente abrangência inferior a 100% e listas com mais termos que a referência sempre tem precisão inferior a 100%.

No entanto, os valores máximos da medida F, que representa o melhor equilíbrio entre precisão e abrangência, ocorrem um pouco depois desse cruzamento de curvas, respectivamente, nas listas de 2.000 bigramas e 3.300 trigramas. Isto se deve ao fato das curvas de precisão e abrangência terem um comportamento diferente, pois a queda dos valores de precisão é mais lenta, em comparação com o aumento rápido dos valores de abrangência.

Para os bigramas, percebe-se que o valor de abrangência se estabiliza por volta da aplicação do ponto de corte com 2.000 termos, onde atinge-se cerca de 93% de abrangência. Esse fato é curioso, posto que a lista de referência de bigramas possui 1.534 termos, ou seja, é necessário estender o ponto de corte absoluto para extrair cerca de 500 termos a mais do que 1.534 (tamanho da lista de referência) para atingir uma alta abrangência. A precisão, ao contrário, se mantém em valores altos (acima de 70%) até esse mesmo ponto de corte de 2.000 bigramas, caindo de maneira mais acentuada para pontos de corte menos restritivos.

Para os trigramas, que possuem 2.660 termos na lista de referência, percebe-se um comportamento análogo das curvas de precisão e abrangência somente para listas com 3.300 termos. Nesse caso, mais de 600 termos adicionais tiveram de ser extraídos para se chegar a uma alta abrangência (94%).

Se quiséssemos escolher um ponto de corte absoluto para os termos extraídos do *corpus*



pontos de corte absolutos	bigramas				trigramas			
	P	R	F	$\mid \mathcal{LE} \cap \mathcal{LR} \mid$	P	R	F	$\mid \mathcal{LE} \cap \mathcal{LR} \mid$
100	99%	6%	12%	99	93%	3%	7%	93
200	98%	13%	22%	195	94%	7%	13%	187
300	96%	19%	31%	288	94%	11%	19%	281
400	95%	25%	39%	380	92%	14%	24%	366
500	94%	31%	46%	469	92%	17%	29%	459
600	92%	36%	52%	550	91%	21%	33%	546
700	91%	41%	57%	636	90%	24%	38%	631
800	89%	46%	61%	710	90%	27%	42%	719
900	88%	51%	65%	789	89%	30%	45%	797
1.000	86%	56%	67%	855	88%	33%	48%	879
1.100	84%	61%	71%	929	87%	36%	51%	960
1.200	83%	65%	73%	1.001	87%	39%	54%	1.045
1.300	83%	70%	76%	1.074	87%	42%	57%	1.128
1.400	81%	74%	77%	1.129	87%	46%	60%	1.217
1.500	79%	77%	78%	1.179	85%	48%	61%	1.277
1.600	77%	80%	79%	1.234	84%	51%	63%	1.350
1.700	76%	84%	79%	1.285	83%	53%	65%	1.416
1.800	74%	87%	80%	1.331	82%	56%	66%	1.483
1.900	73%	90%	81%	1.388	81%	58%	68%	1.548
2.000	72%	93%	**81%**	1.432	81%	61%	69%	1.614
2.100	68%	94%	79%	1.437	80%	63%	70%	1.674
2.200	66%	94%	77%	1.441	79%	65%	72%	1.742
2.300	63%	94%	75%	1.442	79%	68%	73%	1.806
2.400	60%	94%	73%	1.442	78%	70%	74%	1.875
2.500	58%	94%	72%	1.447	78%	73%	75%	1.938
2.600	56%	94%	70%	1.449	77%	75%	76%	2.003
2.700	54%	95%	69%	1.451	77%	78%	78%	2.080
2.800	52%	95%	67%	1.453	77%	81%	79%	2.154
2.900	50%	95%	66%	1.453	77%	84%	80%	2.222
3.000	48%	95%	64%	1.453	77%	86%	81%	2.298
3.100	47%	95%	63%	1.457	76%	89%	82%	2.370
3.200	46%	95%	62%	1.460	77%	92%	84%	2.449
3.300	44%	95%	60%	1.462	76%	94%	**84%**	2.514
3.400	43%	95%	59%	1.464	74%	94%	83%	2.514
3.500	42%	96%	58%	1.467	72%	95%	82%	2.515

Figura 5.1: Precisão (P), abrangência (R), medida F (F) e tamanho das listas organizadas por frequência de termo, disjunção de *corpora* (*tf-dcf* - eq. 4.9) obtidas por **pontos de corte absolutos**.

de Pediatria, não seria possível determinar um valor único que fosse adequado para bigramas e trigramas. Portanto, um ponto de corte absoluto único não é uma forma adequada para determinar um ponto ótimo de corte, que nesse exemplo apresentado, seria de 2.000 para bigramas e 3.300 para trigramas.

5.1.2 Pontos de Corte por Limiar

Uma forma popular de descartar termos é o uso de pontos de corte através da determinação de limiares arbitrários de ocorrências de termos no *corpus*. Por exemplo, o trabalho de Bourigault e Lame [28] sugere o uso de um número mínimo de 10 ocorrências para considerar um termo relevante. Essa forma de identificar termos relevantes, corresponde à escolha de um ponto de corte baseado em limiar, ou seja, organizar a lista de termos extraídos segundo um índice e considerar apenas os termos nos quais o seu índice possui um valor acima do limiar escolhido. No caso de Bourigault e Lame [28], o índice escolhido foi a frequência absoluta de termos, porém qualquer índice poderia ser escolhido.

O uso de pontos de corte por limiar baseados na frequência absoluta é adotado com base em um raciocínio intuitivo, que sugere uma relação direta entre o tamanho do *corpus* e o ponto de corte a escolher [202]. Esta intuição, ainda que verdadeira, não é uma relação linear, pois o número de ocorrências de termos em um *corpus* decresce exponencialmente [183].

O formato da curva de decréscimo exponencial pode variar bastante segundo o método de extração, por exemplo, para palavras extraídas segundo um processo puramente estatístico, o decréscimo segue a lei de Zipf[1] [218]. No entanto, o processo linguístico de extração de termos utilizado nessa tese, não segue esta mesma lei, como pode ser verificado pelo número de ocorrência dos 10 termos mais frequentes do *corpus* de Pediatria na Tabela 4.3.

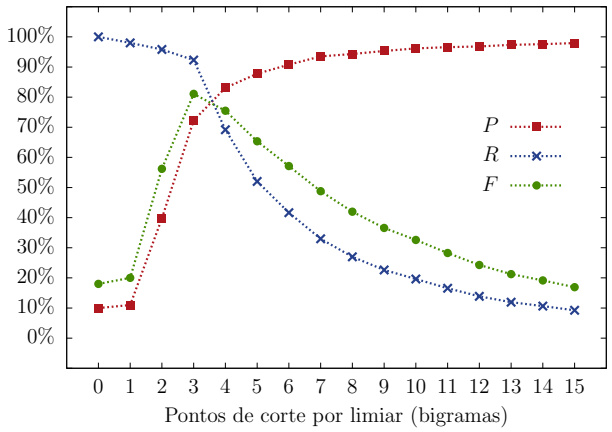
Por essa razão, é difícil propor uma fórmula que permita estimar automaticamente um limiar para ponto de corte a partir do tamanho do *corpus*. Logo, nessa seção analisam-se diversos valores de limiar escolhidos de forma arbitrária.

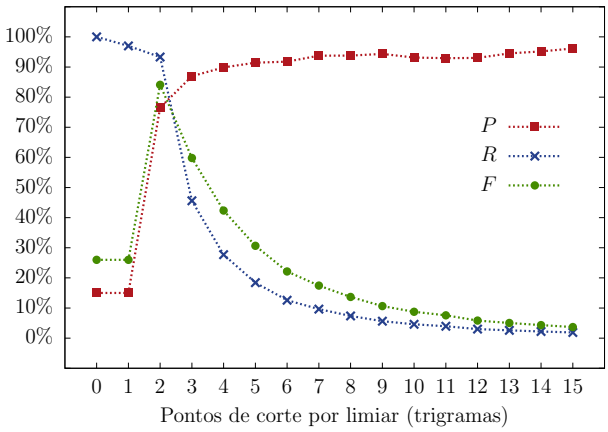
Retomando o *corpus* de Pediatria e listas de bigramas e trigramas de referência utilizados na seção anterior, podemos analisar diversos pontos de corte (de 0 a 15) segundo a frequência de termo, disjunção de *corpora* dos termos extraídos (*tf-dcf* - eq. 4.9). Cabe lembrar que o índice *tf-dcf* mantém uma semântica similar à frequência absoluta de termo (*tf*), pois um termo que possui ocorrências somente no *corpus* de domínio, ou seja, nenhuma ocorrência nos *corpora* contrastantes, terá os mesmo valores para os índices *tf* e *tf-dcf*.

A Figura 5.2 apresenta os resultados de precisão, abrangência e medida F para os pontos de corte por limiar para o índice *tf-dcf* de 0 a 15. Na parte inferior dessa figura, uma tabela indica ainda nas suas últimas duas colunas o tamanho da lista após a aplicação do ponto de corte ($| \mathcal{LE} |$), bem como o tamanho da sua intersecção com a lista de referência ($| \mathcal{LE} \cap \mathcal{LR} |$). Por exemplo, a linha central dessa tabela indica o ponto de corte pelo limiar 8, ou seja, apenas termos que tenham valor de *tf-dcf* igual ou superior a 8 são considerados. Isso resulta em uma lista com 573 bigramas, dos quais 530 estão presentes na lista de referência, assim como, uma lista de 209 trigramas, dos quais 196 estão presentes na lista de referência.

A primeira linha da tabela contida na Figura 5.2 indica um ponto de corte com limiar igual a 0 que corresponde a não desprezar nenhum dos 15.487 bigramas e 18.174 trigramas extraídos. Consequentemente, as listas representadas por essa linha incluem, cada uma delas, todos os 1.534 bigramas e 2.660 trigramas da lista de referência (LR), resultando em uma abrangência de 100% e aproximadamente 10% e 15% de precisão para bigramas e trigramas,

[1]Segundo a lei de Zipf, a frequência de uma palavra em um *corpus* é inversamente proporcional a sua posição (*rank*). Dessa forma, a palavra mais frequente de um *corpus* possui: o dobro de ocorrências do que as ocorrências da segunda palavra mais frequente; o triplo de ocorrências do que as ocorrências da terceira palavra mais frequente; e assim por diante.





limiares de pontos de de corte	bigramas					trigramas				
	P	R	F	$\mid \mathcal{LE} \mid$	$\mid \mathcal{LE} \cap \mathcal{LR} \mid$	P	R	F	$\mid \mathcal{LE} \mid$	$\mid \mathcal{LE} \cap \mathcal{LR} \mid$
0	10%	100%	18%	15.487	1.534	15%	100%	26%	18.174	2.660
1	11%	98%	20%	13.589	1.502	15%	97%	26%	17.227	2.577
2	39%	96%	56%	3.698	1.470	77%	93%	**84%**	3.245	2.483
3	72%	92%	**81%**	1.959	1.416	87%	46%	60%	1.395	1.213
4	83%	69%	75%	1.277	1.061	90%	28%	42%	820	737
5	88%	52%	65%	909	798	91%	18%	31%	536	490
6	91%	42%	57%	704	639	92%	13%	22%	365	335
7	94%	33%	49%	541	506	94%	10%	17%	273	256
8	94%	27%	42%	439	414	94%	7%	14%	209	196
9	95%	23%	37%	364	347	94%	6%	11%	159	150
10	96%	20%	33%	313	301	93%	5%	9%	131	122
11	97%	17%	28%	263	254	93%	4%	8%	113	105
12	97%	14%	24%	220	213	93%	3%	6%	86	80
13	97%	12%	21%	188	183	95%	3%	5%	73	69
14	98%	11%	19%	167	163	95%	2%	4%	62	59
15	98%	9%	17%	143	142	96%	2%	4%	52	50

Figura 5.2: Precisão (P), abrangência (R), medida F (F) e tamanho das listas organizadas por frequência de termo, disjunção de *corpora* (*tf-dcf* - eq. 4.9) obtidas por **pontos de corte por limiar**.

respectivamente. Os resultados nessa figura mostram um aumento de precisão e diminuição de abrangência conforme os pontos de corte vão ficando mais restritivos, ou seja, conforme aumenta o limiar.

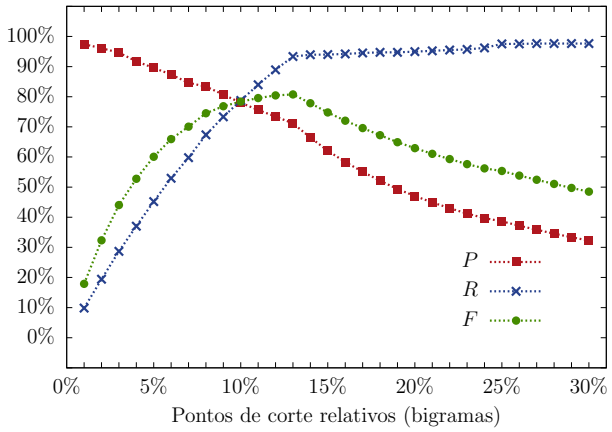
Para bigramas, o ponto que maximiza a combinação de precisão e abrangência (medida F) ocorre com um limiar de 3, que aponta para os valores de 72% de precisão e 92% de abrangência. Para trigramas, o ponto com maior valor de medida F situa-se no limiar de 2, que aponta para os valores de 77% de precisão e 93% de abrangência.

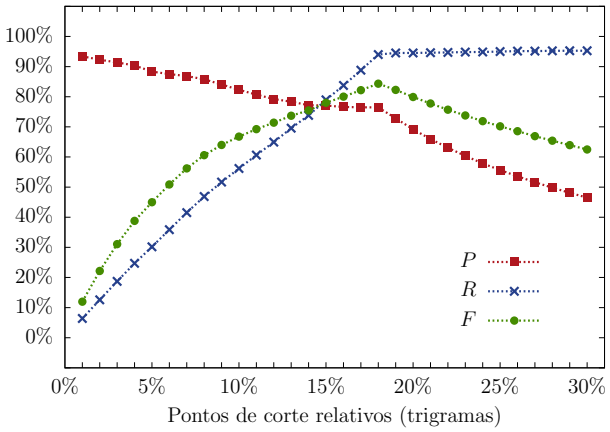
Análogo ao que foi observado com os pontos de corte absolutos, os valores da Figura 5.2 mostram que bigramas e trigramas possuem seus melhores resultados (maior medida F) para distintos valores de limiar (3 e 2, respectivamente). Dessa forma, um limiar único a ser utilizado como ponto de corte que seja adequado a todas listas extraídas não parece ser possível.

5.1.3 Pontos de Corte Relativos

Uma alternativa de pontos de corte, encontrada na literatura [134], é manter apenas um percentual da lista extraída. Essa alternativa é denominada ponto de corte relativo, pois define-se o tamanho da lista a ser considerada proporcional ao total de termos extraídos.

A Figura 5.3 apresenta os valores de precisão, abrangência e medida F para listas organizadas





pontos de corte relativos	bigramas					trigramas				
	P	R	F	$\mid \mathcal{LE} \mid$	$\mid \mathcal{LE} \cap \mathcal{LR} \mid$	P	R	F	$\mid \mathcal{LE} \mid$	$\mid \mathcal{LE} \cap \mathcal{LR} \mid$
1%	97%	10%	18%	155	151	93%	6%	12%	182	170
2%	96%	19%	32%	310	298	92%	13%	22%	363	335
3%	95%	29%	44%	465	440	91%	19%	31%	545	498
4%	92%	37%	53%	619	568	90%	25%	39%	727	657
5%	90%	45%	60%	774	693	88%	30%	45%	909	803
6%	87%	53%	66%	929	812	88%	36%	51%	1.090	954
7%	85%	60%	70%	1.084	917	87%	42%	56%	1.272	1.105
8%	83%	67%	75%	1.239	1.033	86%	47%	61%	1.454	1.247
9%	81%	73%	77%	1.394	1.125	84%	52%	64%	1.636	1.375
10%	78%	79%	78%	1.549	1.208	82%	56%	67%	1.817	1.495
11%	76%	84%	80%	1.704	1.288	81%	61%	69%	1.999	1.613
12%	73%	89%	80%	1.858	1.364	79%	65%	71%	2.181	1.728
13%	71%	93%	**81%**	2.013	1.432	78%	70%	74%	2.363	1.851
14%	66%	94%	78%	2.168	1.441	77%	74%	75%	2.544	1.964
15%	62%	94%	75%	2.323	1.442	77%	79%	78%	2.726	2.101
16%	58%	94%	72%	2.478	1.445	77%	84%	80%	2.908	2.228
17%	55%	95%	70%	2.633	1.450	76%	89%	82%	3.090	2.363
18%	52%	95%	67%	2.788	1.453	76%	94%	**84%**	3.271	2.501
19%	49%	95%	65%	2.943	1.453	73%	95%	82%	3.453	2.515
20%	47%	95%	63%	3.097	1.457	69%	95%	80%	3.635	2.515
21%	45%	95%	61%	3.252	1.461	66%	95%	78%	3.817	2.518
22%	43%	96%	59%	3.407	1.465	63%	95%	76%	3.998	2.520
23%	41%	96%	58%	3.562	1.468	60%	95%	74%	4.180	2.523
24%	40%	96%	56%	3.717	1.476	58%	95%	72%	4.362	2.524
25%	39%	98%	55%	3.872	1.496	56%	95%	70%	4.544	2.528
26%	37%	98%	54%	4.027	1.496	54%	95%	69%	4.725	2.532
27%	36%	98%	52%	4.181	1.498	52%	95%	67%	4.907	2.532
28%	35%	98%	51%	4.336	1.498	50%	95%	65%	5.089	2.534
29%	33%	98%	50%	4.491	1.498	48%	95%	64%	5.270	2.535
30%	32%	98%	48%	4.646	1.498	46%	95%	62%	5.452	2.535

Figura 5.3: Precisão (P), abrangência (R), medida F (F) e tamanho das listas organizadas por frequência de termo, disjunção de *corpora* (*tf-dcf* - eq. 4.9) obtidas por **pontos de corte relativos**.

segundo o índice *tf-dcf* (Eq. 4.9) aplicando pontos de corte relativos percentuais. Nessa figura são considerados pontos de corte variando de 1% a 30%

A primeira observação interessante dos resultados apresentados na Figura 5.3 é que os pontos de corte entre 8% e 15% para bigramas e entre 14% e 22% para trigramas possuem valores de medida F iguais ou superiores a 75%. Isso ocorre, pois conforme aumenta o ponto de corte relativo, a precisão se mantém acima de 62% para bigramas e trigramas, porém a abrangência já mostra valores superiores a 67% para bigramas e superiores a 74% para trigramas.

Em função dessas faixas, é possível observar que um ponto de corte relativo de 15% oferece um compromisso razoável de medida F não inferior a 75% para bigramas e trigramas. Ainda assim, os valores ótimos de equilíbrio entre precisão e abrangência ocorrem nos pontos de corte relativos de 13% para bigramas e de 18% para trigramas. Portanto, também não é possível determinar um único ponto de corte relativo adequado tanto a bigramas, quanto a trigramas.

5.2 Proposta de Ponto de Corte para Identificar Conceitos

Na seção anterior foram feitos experimentos sobre bigramas e trigramas extraídos do *corpus* de Pediatria, pois somente esse *corpus* possui listas de referência, e mesmo assim, essas listas só contêm termos com duas ou três palavras. No entanto, o que se busca não é descobrir uma forma adequada de aplicar pontos de corte exclusivamente a essas listas de termos, mas sim, a todas as listas extraídas de todos os *corpora*. Dessa forma, busca-se critérios que possam ser generalizados para, por exemplo, unigramas do *corpus* de Geologia, ou N-gramas do *corpus* de Modelagem estocástica, *etc.*, pois para essas listas de termos não existe referência disponível.

A análise de um ponto de corte único, seja absoluto, por limiar, ou relativo, não parece ser possível, pois mesmo para os bigramas e trigramas do *corpus* de Pediatria não foi possível estabelecer um ponto de corte único que permitisse valores equilibrados de precisão e abrangência. Se observarmos em detalhe os resultados apresentados nas tabelas das Figuras 5.1, 5.2 e 5.3, percebe-se que existem três formas distintas de escolher um ponto de corte adequado:

- buscar um ponto de corte que garanta uma alta precisão, mantendo, com menor prioridade, uma boa abrangência;
- buscar um ponto de corte que garanta uma alta abrangência, mantendo, com menor prioridade, uma boa precisão;
- buscar o maior equilíbrio possível entre precisão e abrangência, ou seja, o maior valor da medida F.

Evidentemente, se a preocupação individual é somente com precisão, ou somente com abrangência, não se trata de um ponto de interesse científico. Se quisermos maximizar apenas a precisão, basta colocar um ponto de corte muito restritivo, ou seja, com pouquíssimos termos. Por outro lado, a maior abrangência possível ocorre quando não se despreza nenhum termo. Logo, o interesse científico existe quando busca-se o equilíbrio entre precisão e abrangência.

Observando os valores nas Figuras 5.1, 5.2 e 5.3, percebe-se a maximização da medida F no ponto de inflexão da sua curva correspondente, ou seja:

- No caso de pontos de corte absolutos (Figura 5.1), trata-se de 2.000 bigramas, 3.300 trigramas;
- Para pontos de corte por limiar (Figura 5.2), trata-se dos limiares 3 para bigramas e 2 para trigramas;

- Os pontos de corte relativos (Figura 5.3), apontam para listas com 13% dos bigramas extraídos, e 18% dos trigramas extraídos.

Dessa forma, propõe-se um método híbrido para estimar um ponto de corte que seja próximo desse ótimo. Esse método é dito híbrido, pois segue alternativamente ideias vistas para ponto de corte. Especificamente, a proposta é aplicar em conjunto: um ponto de corte por limiar e um ponto de corte relativo. As seções a seguir detalham essa proposta. Cabe salientar que, o uso de ponto de corte absoluto não faz sentido nesse contexto. Um ponto de corte absoluto é a definição de um número fixo (e arbitrário) de termos, enquanto que os demais (relativo e por limiar) são naturalmente dependentes dos *corpora* e listas de termos extraídos.

5.2.1 Aplicação de um Ponto de Corte por Limiar

Aplicando um ponto de corte por limiar a todas as listas de termos extraídos, parece ser razoável descartar termos que não atingem um valor mínimo segundo o índice proposto. Dessa forma, baseado nas análises feitas na Seção 5.1.2 (Figura 5.2), sugere-se, como primeiro passo do método híbrido proposto, descartar termos que tenham um índice *tf-dcf* inferior a 2.

Essa escolha de um limiar 2 é conservadora, pois para bigramas do *corpus* de Pediatria um limiar 3 foi mais adequado (melhor medida F). Cabe lembrar que, para trigramas desse mesmo *corpus* um limiar 2 foi mais adequado, logo a escolha de um limiar 3 iria descartar trigramas relevantes para o *corpus* de Pediatria.

Apesar dessa escolha conservadora, o limiar 2 descarta um grande número de termos, como pode ser observado na Tabela 5.1, que apresenta a redução no número de termos extraídos em todos os *corpora* (Pediatria - PED, Modelagem estocástica - ME, Mineração de dados - MD, Processamento paralelo - PP, e Geologia - GEO). Nessa tabela indica-se o número de termos originalmente extraídos (O) e o número restante após a aplicação do ponto de corte com o limiar 2 ao índice *tf-dcf* (L2).

Tabela 5.1: Aplicação do ponto de corte por limiar escolhido (2) aos *corpora* utilizados.

	PED		ME		MD		PP		GEO	
	O	L2	O	L2	O	L2	O	L2	O	L2
unigramas	5.946	1.974	4.323	872	4.199	716	4.361	905	7.679	2.573
bigramas	15.485	3.696	14.107	3.438	14.804	3.121	14.301	2.938	3.0775	9.262
trigramas	18.172	3.243	18.875	4.655	19.140	4.138	19.976	4.204	3.7210	9.186
4-gramas	13.104	1.192	14.506	2.562	14.024	2.258	14.997	2.072	3.0295	4.817
5-gramas	9.223	560	12.239	2.031	12.349	1.960	12.809	1.602	23.621	3.281
6-gramas	6.676	221	8.410	1.000	8.236	949	8.484	739	17.190	1.990
7-gramas	4.516	124	6.187	690	6.348	728	6.305	458	12.045	1.267
8-gramas	3.095	70	4.210	407	4.450	450	4.404	295	8.523	855
9-gramas	2.161	52	3.061	281	3.232	309	3.216	188	5.905	562
N-gramas	5.078	113	8.077	599	8.906	705	8.726	442	15.383	1.326

Observando os resultados da Tabela 5.1, percebe-se grandes reduções, por exemplo, o descarte de 3.839 termos da lista de termos com 7 ou mais palavras extraídos do *corpus* de Pediatria, ou seja, o descarte de 97% dos termos dessa lista. Mesmo, nos casos de menor redução, como por exemplo, a redução na lista de unigramas do *corpus* de Geologia, 5.106 termos foram descartados, ou seja, descartou-se 66% dos termos dessa lista.

5.2.2 Aplicação de um Ponto de Corte Relativo

Analogamente ao ponto de corte com limiar 2 para o índice *tf-dcf*, a segunda etapa do método híbrido proposto é o descarte de termos por um ponto de corte relativo. Com base nos resultados apresentados na Tabela 5.3, onde um ponto de corte de 13% para bigramas e de 18% para trigramas, mostrou os melhores valores da medida F, escolheu-se utilizar o ponto de corte intermediário (15%) para ser aplicado a todas as listas de termos extraídos.

A Tabela 5.2 apresenta a aplicação do ponto de corte relativo de 15%. Essa tabela indica o número de termos originalmente extraídos (O) e o número de termos restantes após a aplicação do ponto de corte relativo que mantém 15% dos termos extraídos (R15), o que resulta no descarte de 85% dos termos das listas originais.

Tabela 5.2: Aplicação do ponto de corte relativo escolhido (15%) aos *corpora* utilizados.

	PED		ME		MD		PP		GEO	
	O	R15	O	R15	O	R15	O	R15	O	R15
unigramas	5.946	892	4.323	648	4.199	630	4.361	654	7.679	1.152
bigramas	15.485	2.323	14.107	2.116	14.804	2.221	14.301	2.145	30.775	4.616
trigramas	18.172	2.726	18.875	2.831	19.140	2.871	19.976	2.996	37.210	5.582
4-gramas	13.104	1.966	14.506	2.176	14.024	2.104	14.997	2.250	30.295	4.544
5-gramas	9.223	1.383	12.239	1.836	12.349	1.852	12.809	1.921	23.621	3.543
6-gramas	6.676	1.001	8.410	1.262	8.236	1.235	8.484	1.273	17.190	2.579
7-gramas	4.516	677	6.187	928	6.348	952	6.305	946	12.045	1.807
8-gramas	3.095	464	4.210	632	4.450	668	4.404	661	8.523	1.278
9-gramas	2.161	324	3.061	459	3.232	485	3.216	482	5.905	886
N-gramas	5.078	762	8.077	1.212	8.906	1.336	8.726	1.309	15.383	2.307

Observando os dados das Tabela 5.1 e 5.2, verifica-se que listas de unigramas e bigramas são mais reduzidas pelo uso de ponto de corte relativo. Por outro lado, listas de 6-gramas a N-gramas são mais reduzidas pelo uso de ponto de corte por limiar. Esse fenômeno se deve ao fato de que termos mais simples (unigramas e bigramas, por exemplo) tendem a ser mais frequentes, portanto, possuem valores mais altos de *tf-dcf* do que termos compostos por um número maior de palavras.

5.2.3 Método Híbrido para Escolha de Ponto de Corte

Frente ao apresentado nas seções anteriores (5.2.1 e 5.2.2), o método proposto para determinar automaticamente os pontos de corte das listas extraídas de um *corpus* de domínio consiste nas seguintes etapas:

- As listas extraídas e ordenadas segundo o índice *tf-dcf* são submetidas a um ponto de corte pelo limiar 2;
- As listas extraídas e ordenadas segundo o índice *tf-dcf* são submetidas a um ponto de corte relativo de 15%;
- Apenas os termos que não foram descartados por ambos os pontos de corte (limiar 2 e relativo 15%) são mantidos.

Note-se que, como ambos os pontos de corte são aplicados a listas ordenadas pelo mesmo índice (a mesma lista), esse processo equivale a determinar o tamanho da lista resultante por cada um dos dois pontos de corte e utilizar o menor deles. Por exemplo, aplicando o ponto de corte pelo limiar 2 à lista de unigramas do *corpus* de Geologia indica-se manter 2.573 termos, porém aplicando-se o ponto de corte relativo de 15% indica-se manter 1.152 termos. Por consequência, considera-se como conceitos de *corpus* de Geologia os 1.152 unigramas com os maiores valores do índice *tf-dcf*.

O resultado desse método, na forma de quantos conceitos são identificados para cada um dos *corpora* utilizados, é apresentado na Tabela 5.3. Nessa tabela, indica-se o número de termos originalmente extraídos (O) e o número de termos considerados conceitos após a aplicação dos pontos de corte (C). Indica-se ainda nessa tabela, ao lado do número de conceitos, se o ponto de corte mais restritivo foi por limiar (L) ou relativo (R).

Observando os resultados da Tabela 5.3, percebe-se que o uso combinado dos pontos de corte por limiar e relativo permite restringir as listas de termos, desde unigramas até N-gramas.

Tabela 5.3: Número de termos extraídos em cada *corpus* e número de conceitos identificados.

	PED			ME			MD			PP			GEO		
	O	C		O	C		O	C		O	C		O	C	
unigramas	5.946	892	R	4.323	648	R	4.199	630	R	4.361	654	R	7.679	1.152	R
bigramas	15.485	2.323	R	14.107	2.116	R	14.804	2.221	R	14.301	2.145	R	30.775	4.616	R
trigramas	18.172	2.726	R	18.875	2.831	R	19.140	2.871	R	19.976	2.996	R	37.210	5.582	R
4-gramas	13.104	1.192	L	14.506	2.176	R	14.024	2.104	R	14.997	2.072	L	30.295	4.544	R
5-gramas	9.223	560	L	12.239	1.836	R	12.349	1.852	R	12.809	1.602	L	23.621	3.281	L
6-gramas	6.676	221	L	8.410	1.000	L	8.236	949	L	8.484	739	L	17.190	1.990	L
7-gramas	4.516	124	L	6.187	690	L	6.348	728	L	6.305	458	L	12.045	1.267	L
8-gramas	3.095	70	L	4.210	407	L	4.450	450	L	4.404	295	L	8.523	855	L
9-gramas	2.161	52	L	3.061	281	L	3.232	309	L	3.216	188	L	5.905	562	L
N-gramas	5.078	113	L	8.077	599	L	8.906	705	L	8.726	442	L	15.383	1.326	L

Percebe-se também que, conforme dito, os pontos de corte relativos são mais restritivos para as listas de termos mais simples, enquanto que os pontos de corte por limiar são mais efetivos para termos mais longos.

5.3 Resultado Final da Identificação de Conceitos

A definição de pontos de corte conclui o processo de extração automática de conceitos de um *corpus* de domínio. Dessa forma, é possível descrever o processo completo de extração de conceitos através das seguintes etapas:

- Os termos são extraídos e tratados pelas heurísticas descritas no Capítulo 3;
- Os termos extraídos são ordenados segundo o processo de comparação com *corpora* contrastantes e cálculo do índice *tf-dcf* descrito no Capítulo 4;
- As listas de termos extraídos e ordenados são submetidas a um ponto de corte duplo (por limiar e relativo) conforme descrito nesse capítulo.

Feitas essas três etapas, os termos que não forem descartados são considerados conceitos do domínio. De um ponto de vista prático, para cada um dos cinco *corpora* utilizados nessa tese são extraídos os conceitos descritos no anexo B. De um ponto de vista numérico, a Figura 5.4 representa graficamente o número de termos extraídos e o número de conceitos identificados. Nessa figura, identifica-se com a sigla do *corpus* a barra corresponde ao número de termos extraídos e com um asterisco os conceitos identificados. Note-se que nessa figura representa-se os números de termos e conceitos sem repetição, ou seja, ao contrário do representado na Figura 3.7, não são representadas as diversas ocorrências de um mesmo termo (ou conceito) extraído do *corpus*.

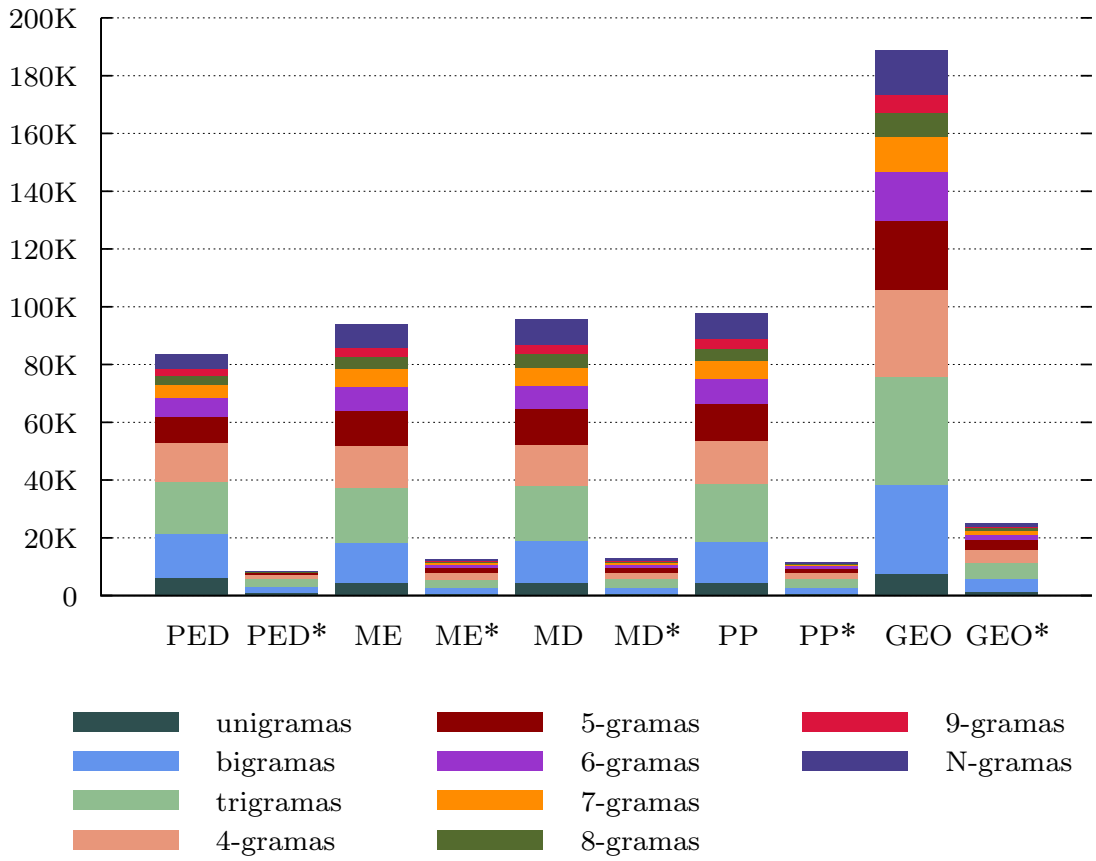


Figura 5.4: Comparativo do número de termos extraídos considerando a aplicação das heurísticas e identificação de conceitos.

6. APLICAÇÕES DOS TERMOS E CONCEITOS EXTRAÍDOS

Uma vez extraídos os conceitos, várias recursos linguísticos podem ser disponibilizados. Nesse capítulo exemplificam-se algumas dessas possíveis aplicações que foram implementadas na ferramenta EχATO$_{LP}$. Essa ferramenta de software realiza todo o processo de extração e ordenação de termos, bem como a identificação de conceitos proposta nessa tese. Os recursos linguísticos disponibilizados, ou seja, as informações detalhadas de termos extraídos (Tabela 3.10) e as listas de conceitos (anexo B), possibilitam a geração de recursos mais sofisticados pela manipulação dessas informações.

Nesse sentido, esse capítulo apresenta as seguintes aplicações:

- Geração de listas de termos e conceitos (Seção 6.1);
- Concordanciador de termos extraídos (Seção 6.2);
- Geração de nuvens de conceitos (Seção 6.3);
- Geração de hierarquia de conceitos (Seção 6.4).

Cabe salientar que, essas aplicações representam algumas utilizações dos recursos linguísticos produzidos pelo processo de extração de conceitos proposto nessa tese, mas muitas outras aplicações podem ser implementadas. No entanto, as aplicações descritas nesse capítulo representam um conjunto de funcionalidades práticas disponibilizadas com a ferramenta EχATO$_{LP}$, e que já vem sendo utilizadas por diversos grupos de pesquisa [53, 164, 67, 52].

6.1 Listas de Termos e Conceitos

A disponibilização de listas de termos e listas de conceitos dos *corpora* é a principal aplicação do processo desenvolvido nessa tese. Dados alguns *corpora* de domínio, é possível disponibilizar listas, não somente de conceitos, mas de quaisquer termos extraídos. Enquanto os conceitos tem um uso mais específico, como por exemplo, construção de hierarquias de conceitos, ontologias, glossários, *etc.* As listas de termos podem ser úteis para aplicações mais ligadas a uma análise humana detalhada, como por exemplo, análise e geração de vocabulários, dicionários de tradução, *etc.*

Adicionalmente, também é possível enriquecer a lista de termos gerada com outras informações. Essas informações adicionais, por sua vez, podem ser manipuladas por consultas que permitam ao usuário da ferramenta EχATO$_{LP}$ inferir conhecimentos sobre o uso dos termos e conceitos no *corpus* que está sendo analisado.

A Figura 6.1 mostra um exemplo de consulta aos bigramas do *corpus* de Geologia que possuem como núcleo a palavra "lago". Nesse exemplo, inclui-se os termos como foram encontrados no *corpus* (*term*), sua forma canônica (*lemma*), seu núcleo (*head*), sua etiqueta semântica (*sem_tag*) e seus índices *tf* e *tf-dcf*.

Além dessas informações, é possível gerar listas de termos e conceitos com outras informações, como, por exemplo:

	A	B	C	D	E	F
1	term	lemma	head	sem tag	tf	tf-dcf
2	lago de Recôncavo	lago de recôncavo	lago	Lwater	4	4.0
3	lagos glaciais	lago glacial	lago	Lwater	3	3.0
4	lagos antigos	lago antigo	lago	Lwater	2	2.0
5	lagos australianos	lago australiano	lago	Lwater	2	2.0
6	lago crescente	lago crescente	lago	Lwater	2	2.0
7	lago costeiro	lago europeu	lago	Lwater	2	2.0
8	lagos hipersalinos	lago hipersalino	lago	Lwater	2	2.0
9	lagos interiores	lago interior	lago	Lwater	2	2.0
10	lagos meromíticos	lago meromítico	lago	Lwater	2	2.0
11	lago profundo	lago recente	lago	Lwater	2	2.0
12	lagos tectônicos	lago tectônico	lago	Lwater	2	2.0
13	lago transgressivo	lago transgressivo	lago	Lwater	2	2.0

Figura 6.1: Exemplo de lista bigramas do *corpus* de Geologia com núcleo "lago".

- Variações morfológicas em que o termo foi encontrado no *corpus*;
- Número de vezes em que o termo foi empregado como sujeito, objeto ou complemento;
- Verbos aos quais o termo foi relacionado;
- Valor numérico dos índices *tf-idf*, *tds*, *thd* e *TF-IDF* relativos ao termo;
- Informações referente ao núcleo do termo.

As variações morfológicas nas quais o conceito foi encontrado permitem observar características de como o termo é empregado. Esse tipo de informação é útil a pesquisadores que podem, através desse recurso linguístico, observar padrões de uso de diversos termos. Por exemplo, no *corpus* de Pediatria, os termos "criança" e "bebê" têm padrões bem distintos de variações morfológicas. O termo "criança" é empregado 984 vezes no singular e 1.076 no plural. O termo "bebê" é empregado 138 vezes no singular e 64 vezes no plural.

O número de ocorrências em que o termo foi empregado como sujeito, objeto ou complemento, também pode auxiliar na detecção de padrões de uso em áreas distintas. Por exemplo, o termo "ordem" aparece em todos os *corpora*, porém ele é encontrado como sujeito 19% das vezes no *corpus* de Processamento paralelo (13 de 68 ocorrências), enquanto que no *corpus* de Geologia ele é encontrado como sujeito somente 8% das vezes (5 de 61 ocorrências).

Os verbos aos quais o termo foi relacionado podem indicar mais um aspecto das características de uso do termo. Por exemplo, no *corpus* de Pediatria, os únicos unigramas que estão relacionados com o verbo "desconhecer" são os termos "mãe", "sorologia" e "universo". Sendo que desses, apenas o termo, "mãe", foi utilizado como sujeito do verbo desconhecer.

O valor numérico dos índices relativos a cada termo extraído também permite analisar as características do termo. Esse tipo de informação permite que sejam feitas análises, ordenações e até aplicações de pontos de corte experimentais segundo outros critérios, além dos adotados nessa tese (Capítulo 5).

Finalmente, as informações relativas ao núcleo do termo possibilitam observar outros aspectos da utilização dos termos. Um exemplo do uso desse tipo de informação é a identificação das etiquetas sintáticas (*pos-tag*) dos núcleos, que permite, por exemplo, identificar quais termos possuem como núcleo substantivos comuns. Informações como essas podem permitir análises linguísticas avançadas, e até a redefinição de métodos de extração de termos e identificação de conceitos.

6.2 Concordanciador de Termos

Uma aplicação de grande utilidade para pesquisadores da área é um concordanciador [174], ou seja, uma ferramenta que localiza ocorrências de um determinado termo no *corpus* e mostra o seu contexto de utilização e outras informações adicionais. Alguns exemplos de concordanciadores são os softwares Unitex [191] e WordSmith [211]. Por contexto de utilização do termo, entende-se as frases onde o termo ocorre no *corpus* e sua posição na frase. Por exemplo, a utilização do concordanciador implementado no EχATO$_{LP}$ para o termo "parente" presente 7 vezes no *corpus* de Pediatria resulta nas informações parcialmente apresentadas na Figura 6.2.

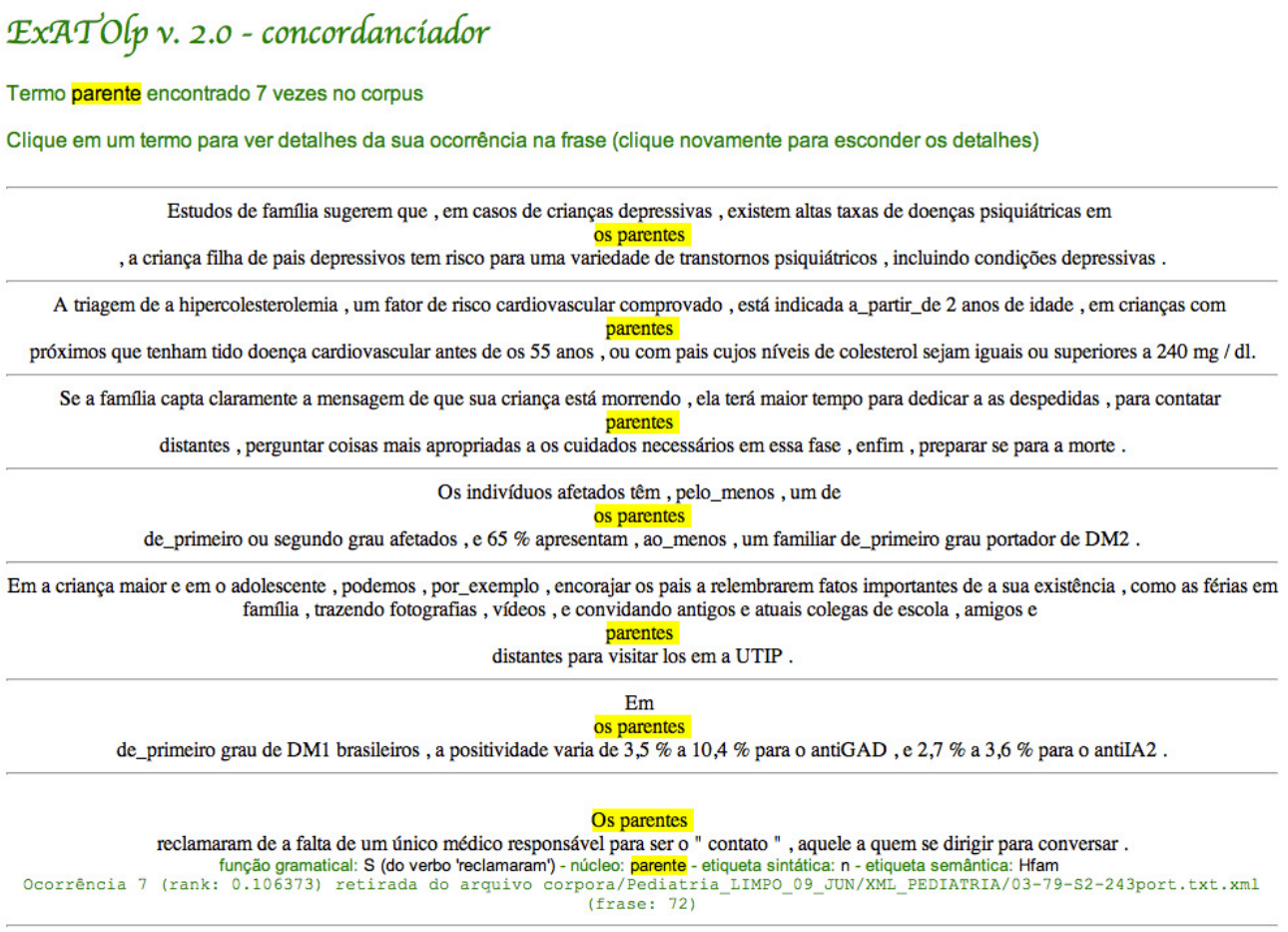


Figura 6.2: Exemplo de saída do concordanciador para o termo "parente" no *corpus* de Pediatria.

Na Figura 6.2 estão indicadas as frases do *corpus* de Pediatria onde o termo "parente" foi encontrado. Dessas sete ocorrências, está indicada na última delas a ocorrência do termo flexionado no plural ("parentes"), que aparece sendo empregado como sujeito do verbo "reclamar", precedido do artigo definido, tendo como núcleo um substantivo (etiqueta "n") e com uma etiqueta semântica "Hfam" que significa humano com relação familiar, segundo o *parser* utilizado [20].

Cabe salientar que, a qualidade das informações adicionais disponibilizadas pelo concordanciador é bastante dependente do *parser* utilizado, pois as informações de etiquetas apresentadas são originadas na anotação linguística feita no início de todo o processo. Apesar disso, ou seja, independente da qualidade do *parser*, a função principal do concordanciador é preservada, pois ele facilita a visualização em detalhe da posição onde o termo se encontra, e, por consequência, a possibilidade de análise visual do seu contexto de utilização (frases).

6.3 Nuvens de Conceitos

A aplicação de nuvens de conceitos (*tag clouds*) é um recurso de visualização que permite observar graficamente a relevância dos conceitos extraídos. A saída do processo de identificação de conceitos (lista de conceitos) é associada ao índice *tf-dcf* de cada conceito, que indica sua relevância.

A nuvem é produzida escrevendo cada um dos conceitos em posições e cores aleatórias, porém, com tamanho de fonte proporcional ao seu índice, logo a sua relevância no *corpus* de domínio. Essa forma de visualização é bastante intuitiva e seu uso em vários materias gráficos e na internet vem sendo bastante difundido recentemente [178, 112].

Exemplos de nuvens de conceitos para bigramas e trigramas do *corpus* de Pediatria são apresentados nas Figuras 6.3 e 6.4, respectivamente. Nessas duas figuras escreve-se 2.323 bigramas e 2.726 trigramas em tamanho proporcional ao seus índices *tf-dcf*.

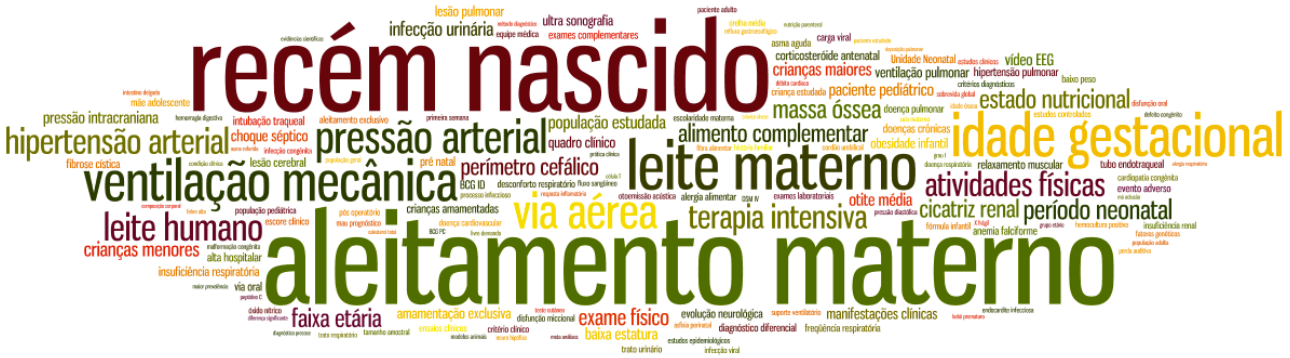


Figura 6.3: Exemplo de nuvem de conceitos para bigramas do *corpus* de Pediatria.

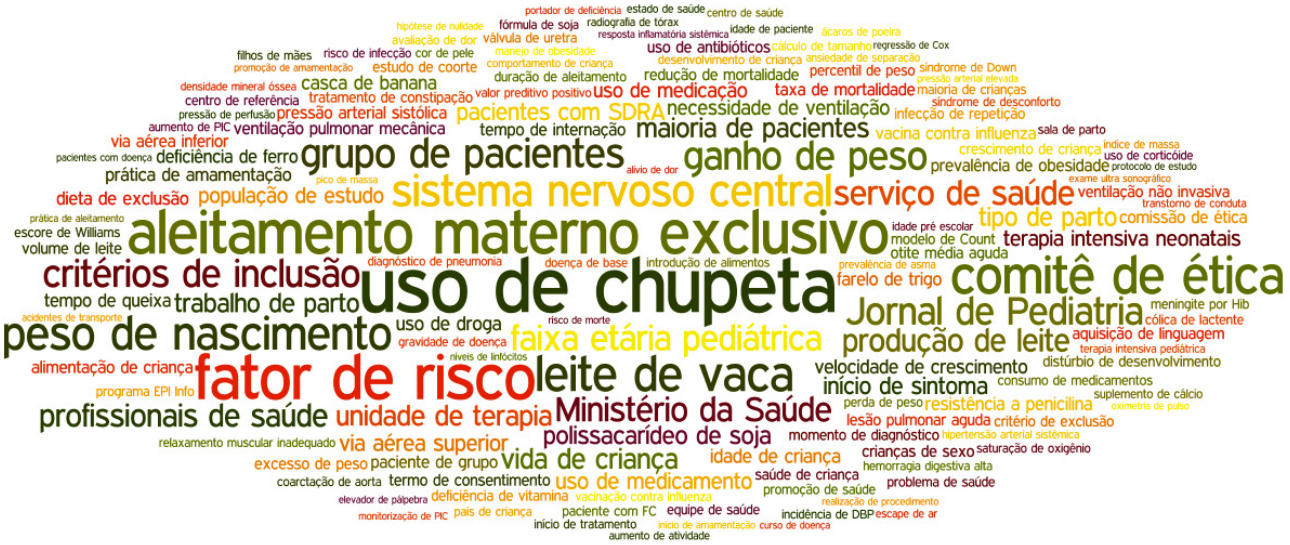


Figura 6.4: Exemplo de nuvem de conceitos para trigramas do *corpus* de Pediatria.

A observação dessas nuvens de conceitos permite visualizar, de maneira muito clara, que os bigramas “aleitamento materno” e “recém-nascido”, assim como os trigramas “uso de chupeta”, “fator de risco” e “aleitamento materno exclusivo”, são os conceitos mais relevantes, segundo o índice *tf-dcf*, dentre bigramas e trigramas do *corpus* de Pediatria.

6.4 Hierarquias de Conceitos

A aplicação mais ambiciosa dentre as apresentadas nesse capítulo é a construção de uma hierarquia de conceitos, conforme definição formal feita na Seção 2.2.3. Segundo essa definição, uma hierarquia é um conjunto de conceitos e um semi reticulado superior, ou seja, um conjunto de relações de subconceitos e superconceitos[1]. O conjunto de conceitos é disponibilizado como saída do processo descrito nos capítulos anteriores. Portanto, para construir a hierarquia é necessário definir/descobrir esse tipo de relação entre esses conceitos.

Dentre as abordagens clássicas de detecção de subconceitos e superconceitos, cita-se a busca em dicionários [113], a busca por padrões morfossintáticos proposta para o Inglês por Hearst [88], e adaptada para diversas outras línguas, como o Francês [142] e o Português [15].

Outras abordagens tradicionais são a análise de coocorrência [172, 72], a busca de similaridade distribucional [68, 87, 139, 81], a análise distribucional [28], e diversas iniciativas regrupadas na denominação genérica de Análise Formal de Conceitos (FCA - *Formal Concept Analysis*) [90, 150, 64, 35, 43, 30, 9].

Nessa seção, adota-se uma solução distinta das abordagens clássicas, que é composta por dois níveis detalhados a seguir.

6.4.1 Hierarquia por Etiquetas Semânticas

O *parser* PALAVRAS define 174 diferentes etiquetas semânticas, que são agrupadas em 16 classes. A Figura 6.5 apresenta essas classes nas folhas da árvore (representados pelos retângulos da Figura 6.5). Os demais nodos não terminais da árvore representam uma divisão por contextos de aplicação dos termos. Cada uma das 16 classes apresentadas na Figura 6.5 congrega diversas etiquetas semânticas. Maiores informações sobre as etiquetas semânticas, e suas classes, podem ser encontradas na especificação do *parser* [20], porém no anexo C o leitor encontra uma listagem dessas etiquetas, bem como sua classificação.

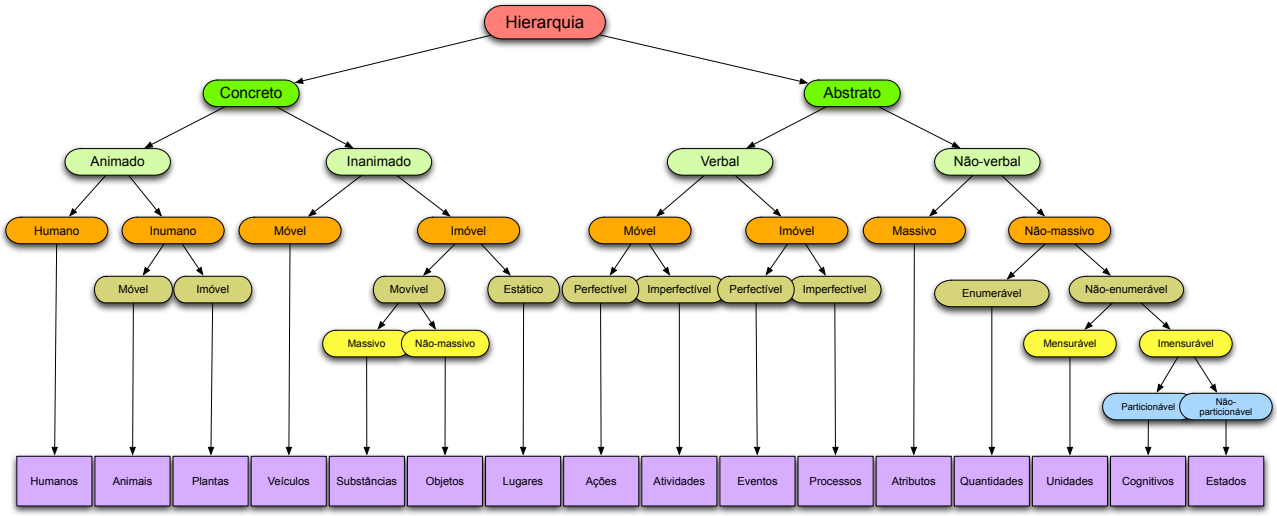


Figura 6.5: Hierarquia de classes de etiquetas semânticas encontradas no *parser*.

Inicialmente, todos os conceitos extraídos são associados, pelo *parser*, a uma das 174 etiquetas, e por consequência, a uma das 16 classes. Essa primeira hierarquização dos conceitos, feita por uma classificação segundo suas etiquetas semânticas, é denominada hierarquia semântica.

Por exemplo, os conceitos extraídos do *corpus* de Geologia: "lago", "mar", "lago profundo", "lago costeiro" e "mar aberto" possuem a etiqueta semântica "lugares aquáticos" (*Lwater*) por

[1]Um termo t_1 é considerado superconceito de um termo t_2 quando t_1 é uma generalização de t_2. Nesse caso t_2 é dito subconceito de t_1. Por exemplo, o termo "meio de locomoção" é um superconceito do termo "trem".

atribuição do *parser*. Por sua vez, essa etiqueta "lugares aquáticos" está, nos níveis semânticos, associada ao ramo "concreto" / "inanimado" / "imóvel", / "estático" / "lugares". Portanto, todos esses conceitos serão associados à classe "lugares" conforme indicado na Figura 6.6.

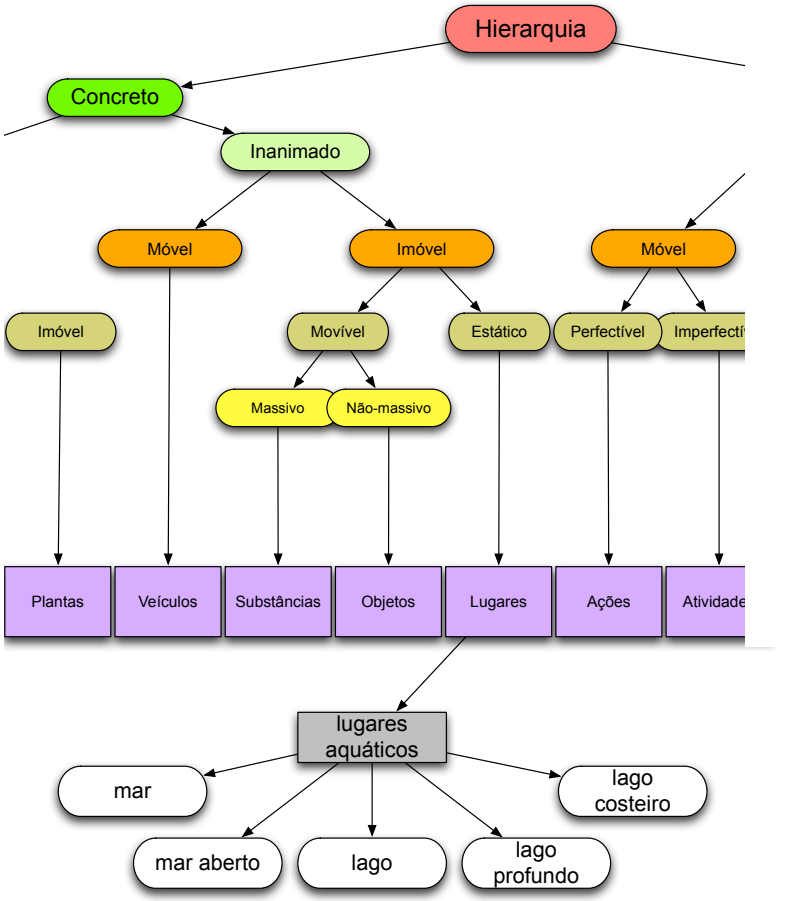


Figura 6.6: Exemplo de associação de conceitos a classes de etiquetas semânticas.

6.4.2 Hierarquia por Núcleo de Sintagmas

Após esse nível de hierarquia semântica, faz-se uma nova estruturação interna a cada subgrupo de conceitos, que possuem a mesma etiqueta semântica. Este segundo nível de hierarquização é feito em dois passos:

- agrupa-se os conceitos (sintagmas nominais) que possuem o mesmo núcleo, por exemplo, os conceitos "lago", "lago profundo" e "lago costeiro" possuem o mesmo núcleo ("lago"), e portanto são agrupados no mesmo ramo;

- dentro do grupo de conceitos com mesmo núcleo, considera-se superconceito de um conceito, o conceito que estiver contido nele, por exemplo, o conceito "lago" será considerado superconceito dos conceitos "lago profundo" e "lago costeiro".

A Figura 6.7 mostra o resultado da associação dos conceitos para esse exemplo de lugares aquáticos. Essa estruturação por núcleo guarda uma semelhança com outros trabalhos que consideram o núcleo do sintagma nominal em suas análises. Esse é o caso do método *hiperN* descrito por Freitas [74], mas também de trabalhos de Amsler baseados exclusivamente em dicionários [5], e, ainda, de iniciativas como a de Nováček [143]. No entanto, nenhuma dessas abordagens infere relações de subconceito/superconceito de conceitos extraídos através dos núcleos dos sintagmas nominais.

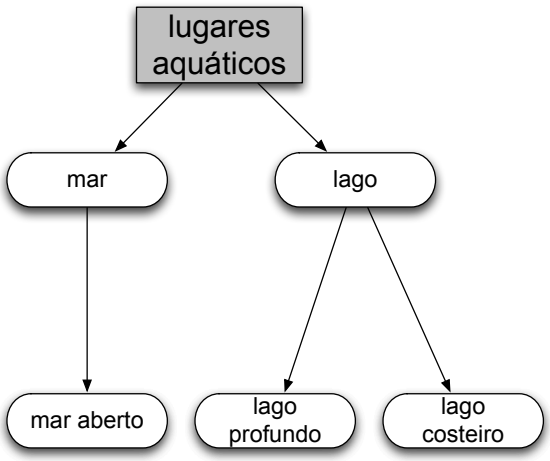


Figura 6.7: Exemplo de relações de subconceitos e superconceitos por núcleo de sintagma.

6.4.3 Exemplo Completo de Hierarquia

Para o *corpus* de Geologia utilizado anteriormente, a geração da hierarquia de conceitos resultou na estrutura apresentada nas Figuras 6.8, 6.9, 6.10 e 6.11. Nessas figuras apresenta-se representações gráficas utilizando árvores hiperbólicas [109] que permitem a visualização interativa da hierarquia.

A Figura 6.8 apresenta uma visão geral da hierarquia, mostrando a hierarquização das etiquetas semânticas. A Figura 6.9 apresenta um nível de detalhe intermediário do ramo que corresponde às etiquetas classificadas dentro do ramo "concreto", subramo "inanimado", subramo "lugares". A Figura 6.10 apresenta a subárvore dos conceitos com a etiqueta semântica "lugares aquáticos". Finalmente, a Figura 6.11 apresenta em detalhe os conceitos "mares" e "lagos".

Nessa última figura percebe-se claramente, entre outros, o conceito extraído "mares", que foi encontrado 798 vezes, e que possui como subconceitos os conceitos extraídos "mar devoniano", "mar regressivo", "mar de norte", *etc.* Percebe-se também o conceito "lago", que foi encontrado 243 vezes, e que possui como subconceitos "lagos profundos", "lagos altos", "lagos baixos", "lagos atuais", *etc.*

Outros exemplos práticos de hierarquias construídas com a ferramenta EχATO$_{LP}$, ou seja, com extração, ordenação e identificação de conceitos foram publicados no Congresso Brasileiro de Informática na Saúde - CBIS 2010 [125] e no Seminário de Pesquisa em Ontologias no Brasil / *International Workshop on Metamodels, Ontologies and Semantic Technologies* - Ontobras/MOST 2011 [67].

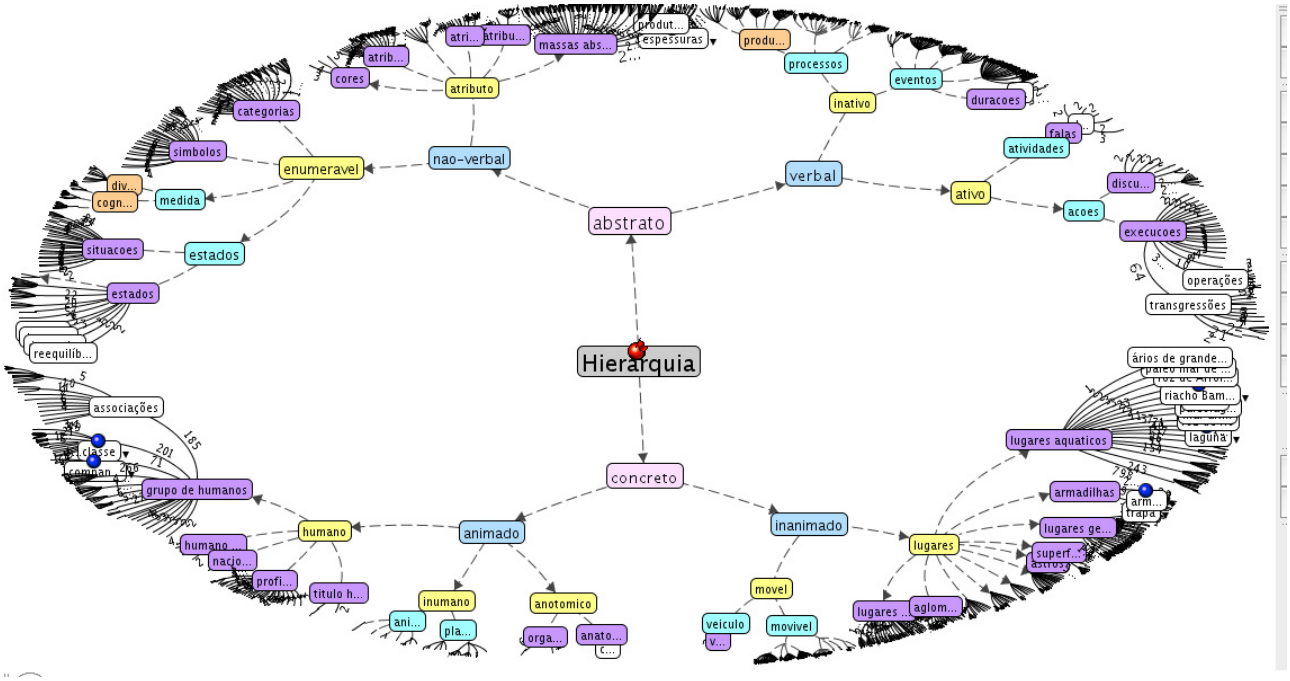


Figura 6.8: Hierarquia de conceitos para o *corpus* de Geologia - visão geral.

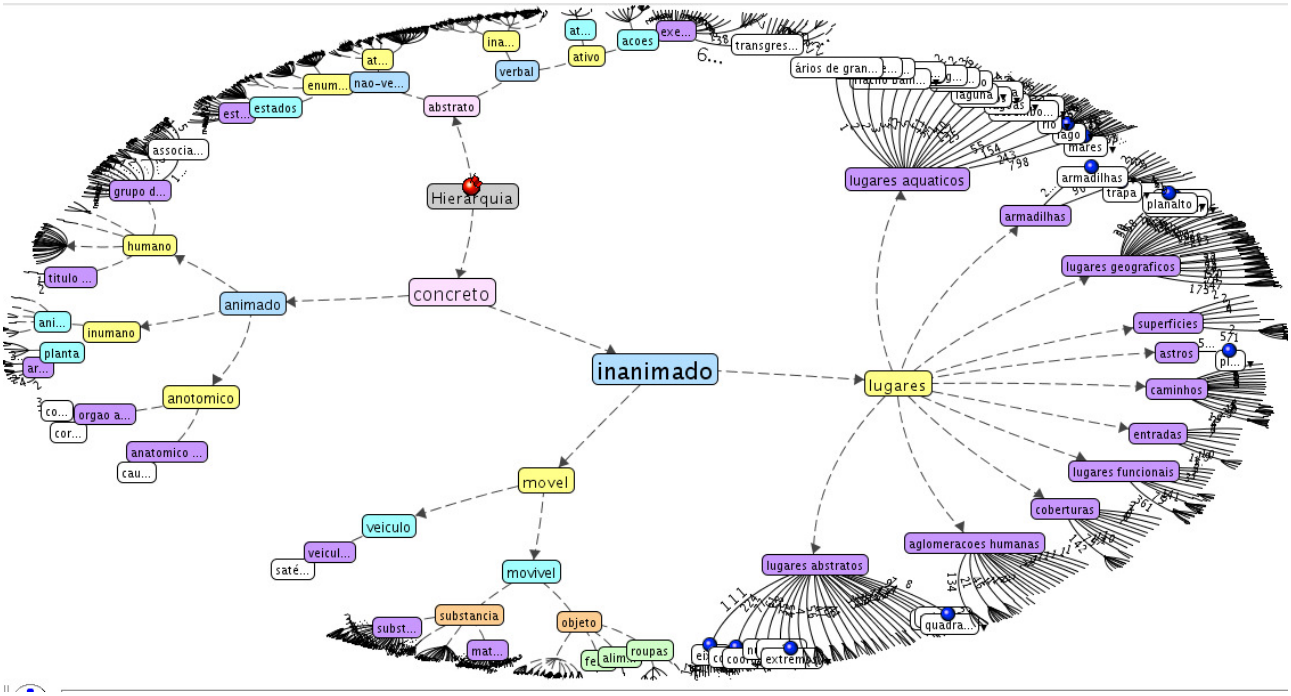


Figura 6.9: Hierarquia de conceitos para o *corpus* de Geologia - detalhe no ramo "lugares".

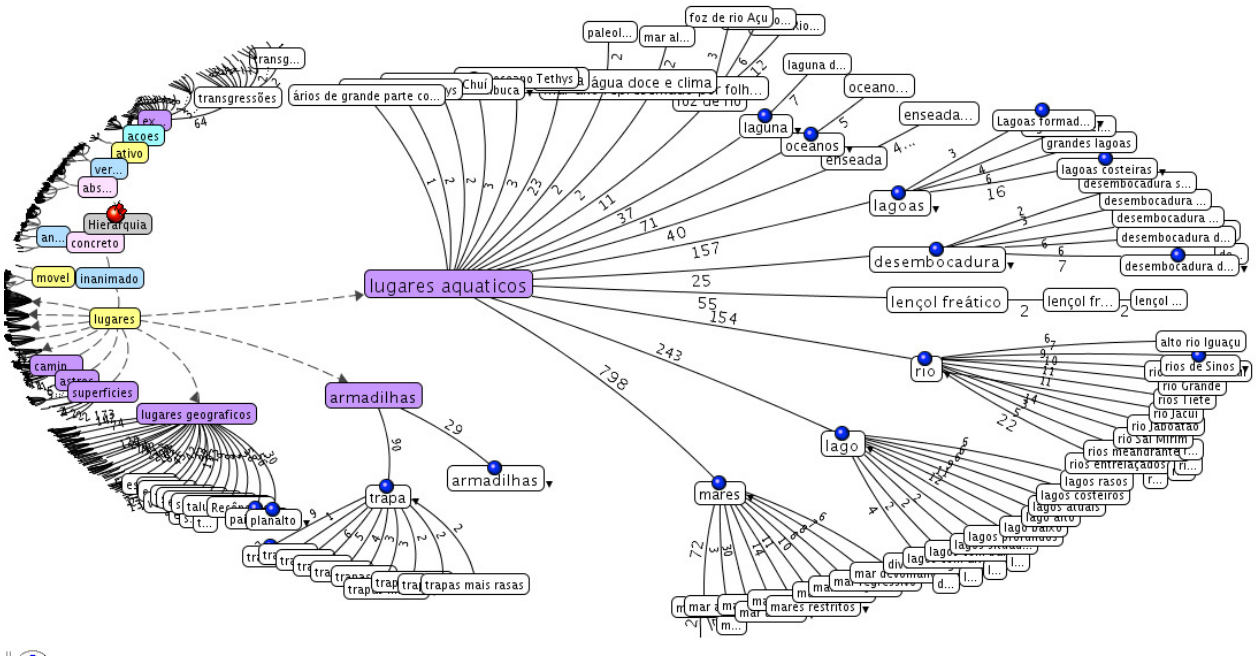


Figura 6.10: Hierarquia de conceitos para o *corpus* de Geologia - detalhe nos conceitos com etiqueta "lugares aquáticos".

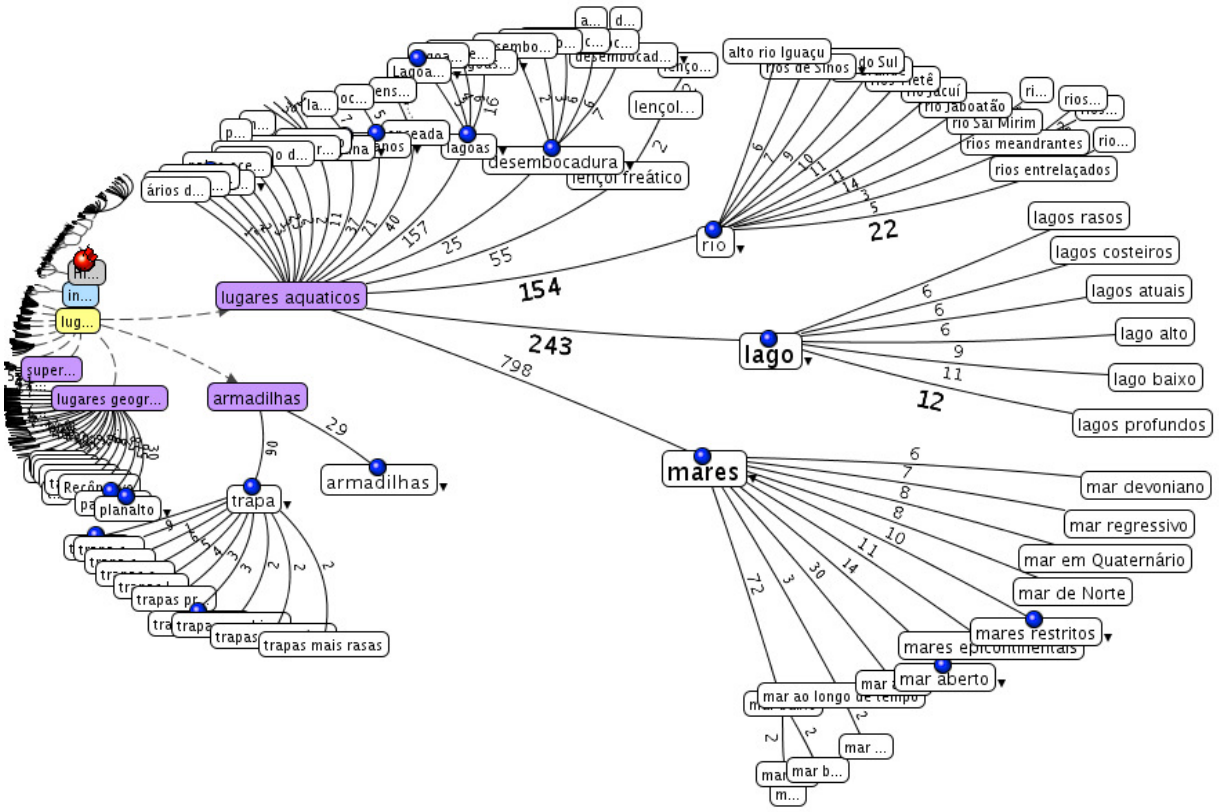


Figura 6.11: Hierarquia de conceitos para o *corpus* de Geologia - detalhe nas subárvores dos conceitos "mares" e "lagos".

7. CONCLUSÃO

O objetivo central dessa tese foi o desenvolvimento de um processo de extração de conceitos a partir de *corpora* de domínio. Dessa forma, assumiu-se como entrada *corpora* anotados linguisticamente e como saída do processo uma lista de conceitos dos domínios que cada um dos *corpus* caracteriza.

Esse objetivo foi alcançado e experimentado sobre cinco *corpora* de domínio, que juntos totalizam um conjunto de textos com quase 6 milhões de palavras. A avaliação de cada etapa do processo foi feita de forma empírica através de experimentos com bigramas e trigramas de um dos *corpus* (Pediatria), para o qual havia listas padrão de referência (*gold standard*). Cabe lembrar que, segundo a literatura [95, 153, 105], a própria natureza da extração de termos e conceitos é subjetiva, e, portanto, somente avaliações empíricas são possíveis.

Completando o objetivo, a utilidade dos conceitos extraídos foi exemplificada pela disponibilização de recursos de grande utilidade para pesquisadores e usuários de ferramentas da área de linguística computacional. Adicionalmente, todos os métodos propostos, bem como a geração automática dos recursos linguísticos, foram implementados em uma ferramenta de software, EχATO$_{LP}$, que, ao mesmo tempo, ilustra e permite avaliar empiricamente todas as propostas dessa tese feitas nos Capítulos 3, 4, 5 e 6.

7.1 Contribuições Científicas e Tecnológicas

Na busca do objetivo dessa tese, foram desenvolvidos avanços científicos expressos por:

1. uma abordagem linguística de extração de termos, que propôs um conjunto de heurísticas a aplicar a sintagmas nominais extraídos de um texto linguisticamente anotado por um *parser*, que trouxe um aumento de precisão e abrangência de cerca de 50% frente à extração tradicional;

2. um novo índice de relevância de termos, que permite, pela comparação com *corpora* contrastantes, estimar a relevância de termos para um domínio específico com precisão superior aos demais índices análogos existentes;

3. uma proposta de estimativa genérica de pontos de corte em listas de termos organizados por relevância, que permite a identificação de conceitos, resultando em bons valores de medida F;

4. um conjunto de aplicações práticas dos conceitos extraídos e seus contextos, que permite a sua compreensão, manipulação e visualização.

Em relação ao estado da arte, a proposta de extração (Capítulo 3) identificou heurísticas para transformação de sintagmas nominais em termos e conceitos, enquanto que os outros trabalhos nessa linha se limitam a fazer uma extração puramente estatística, como é o caso do NSP [11]. Mesmo trabalhos mais próximos, como o da ferramenta OntoLP [165] que também faz extração baseada em sintagmas nominais e utiliza a entrada de textos anotados linguisticamente, possuem valores de precisão e abrangência semelhantes aos conseguidos com a extração

sem o uso de heurísticas. Dessa forma, nossa contribuição ao estado da arte da extração de termos de *corpora* na língua portuguesa é de um processo, que pelo uso das heurísticas, traz, em relação aos trabalhos correlatos, um aumento de, approximadamente, 10% para mais de 60%, tanto na precisão, como na abrangência de listas de termos extraídos comparados com listas de referência.

No que diz respeito ao estado da arte no estabelecimento de um índice de relevância para termos extraídos, a proposta do índice *tf-dcf* (Capítulo 4) traz uma contribuição clara pela formalização de um índice com sólida básica matemática. O ganho de precisão trazido pelo índice *tf-dcf* frente a abordagens tradicionais, como o popular *tf-idf* [130], é de cerca de 10%. Mesmo frente a trabalhos mais recententes, com abordagens similares pelo uso de *corpora* contrastantes [148, 103, 102], o índice *tf-dcf* apresentou valores mais altos de precisão em todos os experimentos realizados.

A proposta de ponto de corte para a identificação de conceitos (Capítulo 5) traz contribuições frente ao estado da arte pela sua originalidade. Outros trabalhos similares utilizam alternativamente pontos de corte absolutos [147, 138, 202, 7], pontos de corte relativos [134], ou pontos de corte por limiar [28, 119], mas nenhuma publicação prévia cita o uso híbrido de pontos de corte. Dessa forma, a abordagem proposta pelo uso combinado de um ponto de corte por limiar do índice *tf-dcf* e de um ponto de corte relativo dos termos extraídos traz uma contribuição objetiva ao fornecer valores adequados de medida F, mas, principalmente, por propor uma forma híbrida de escolha de pontos de corte.

As aplicações desenvolvidas (Capítulo 6) trazem uma contribuição ao estado da arte de disponibilização de termos e conceitos por conjugar conceitos existentes, como listas, concordanciadores [174], nuvens de etiquetas [112] e árvores hiperbólicas [109], com a saída qualificada de termos e conceitos extraídos. Porém, uma contribuição relevante ao estado da arte é a proposta, ainda inicial, de uma forma de construir hierarquias de conceitos com uma parte semântica, e outra parte baseada em núcleo de sintagmas nominais. Essa forma de construir hierarquias pode em trabalhos futuros ser uma alternativa a outras abordagens similares encontradas na literatura [88, 81, 43, 143, 74, 9].

Além das contribuições científicas, essa tese traz três contribuições tecnológicas, que de um ponto de vista prático, se materializam nos os seguintes recursos:

1. a ferramenta EχATO$_{LP}$, que, além de implementar todo o processo de extração de conceitos descrito, oferece diversos modos de saída de termos e conceitos na forma de listas, concordanciador, nuvens de conceitos e uma hierarquia de conceitos;

2. os cinco *corpora* de domínio que serviram para todas as experiências dessa tese, e, por ser um conjunto homogêneo de *corpora*, se configura em um importante recurso linguístico para o tratamento computacional da língua portuguesa;

3. listas de conceitos (termos mais relevantes) dos *corpora* de domínio, que podem ser utilizados diretamente, ou após revisão manual por especialistas, como listas de referências para os *corpora* desenvolvidos.

Essas contribuições serão disponibilizadas imediatamente após a publicação dessa tese no site do grupo de PLN da PUCRS: `http://www.inf.pucrs.br/~linatural/` que é o grupo no qual esse trabalho de doutoramento se insere.

7.2 Difusão das Contribuições dessa Tese na Comunidade Acadêmica

Apesar de ser uma ferramenta recente, e ainda em estágio de protótipo, a comunidade acadêmica tem utilizado resultados provenientes da ferramenta EχATO$_{LP}$ para suas pesquisas. Essa rápida disseminação da ferramenta atesta um reconhecimento das contribuições propostas nessa tese, e materializadas na implementação da ferramenta.

Além de trabalhos do grupo de PLN da PUCRS [125, 119, 201, 36], resultados gerados pelo método proposto nessa tese, e implementados no EχATO$_{LP}$, vem sendo utilizados também por pesquisadores do NILC da USP-São Carlos [53, 78, 79], DIE da UFPI [52], e do NIED da UNICAMP [163, 164], como extrator de termos relevantes de *corpus* de domínio. Já o grupo do Projeto TEXTCC da UFRGS [67, 18, 156], tem utilizado, além da extração de termos, a geração de hierarquias de conceitos, que podem ser visualizadas *on-line* na forma de árvores hiperbólicas. Somam-se a esses trabalhos já publicados, trabalhos em desenvolvimento por pesquisadores do PPGIA da PUCPR, do CIn da UFPE, e do LIA da UNESP, que têm utilizado listas de termos geradas pelo EχATO$_{LP}$.

É importante ressaltar, ainda, a existência de trabalhos científicos, publicados, comparando a ferramenta EχATO$_{LP}$ com outras ferramentas com o mesmo propósito. Nesses trabalhos, verifica-se o melhor desempenho dos métodos propostos nessa tese, frente ao estado da arte da extração de termos e conceitos em textos de língua portuguesa. Dentre esses trabalhos, dois deles [119, 79] comparam o desempenho dos métodos propostos nessa tese e implementados no EχATO$_{LP}$ com a ferramenta NSP [11]. Outro trabalho [163] compara o EχATO$_{LP}$ com duas outras ferramentas similares, KEA [137] e CLUTO [96], e um último trabalho [78], compara a ferramenta OntoLP [165] ao EχATO$_{LP}$.

7.3 Trabalhos Futuros

Os trabalhos futuros dessa tese se manifestam em cinco eixos de pesquisa: um eixo experimental; um eixo de aplicações linguísticas; um eixo de desenvolvimento (programação); um eixo de extração de termos; e um eixo em construção automática de ontologias.

Dentro do eixo experimental, imagina-se a extensão dos experimentos feitos no decorrer dessa tese com outros *corpora*, mas principalmente com outras listas de referência. Posto que a avaliação do processo desenvolvido é obrigatoriamente empírica, é interessante aumentar o número de experiências para outros *corpora*. Porém, a escassez de recursos linguísticos, em especial listas de referências em português, limitou os experimentos feitos nessa tese. No entanto, experiências com o índice *tf-dcf* e com os pontos de corte propostos podem ser repetidas para *corpora* e listas de referências em outras línguas. Esse tipo de experiência pode aumentar a credibilidade dos métodos propostos nessa tese.

Dentro do eixo de aplicações linguísticas, a disponibilização da ferramenta EχATO$_{LP}$ abre a possibilidade de uma série muito grande de estudos linguísticos sobre padrões de uso da língua em diversos contextos, como foi feito no trabalho de Finatto *et al.* [67] que analisou o vocabulário empregado em jornais populares. Outros trabalhos nessa linha podem ser realizados comparando diferenças de estilo de escrita entre áreas do conhecimento, regiões do país, escolas de pensamento, *etc.* Um trabalho particularmente interessante nesse eixo é a aplicação da extração de conceitos proposta em projetos da área de Inteligência Competitiva, que já estão em desenvolvimento no grupo de PLN da PUCRS. A riqueza dos recursos linguísticos disponíveis na ferramenta EχATO$_{LP}$ permite automatizar grande parte do trabalho hoje realizado quase que manualmente por terminólogos e linguistas, ao realizarem análises profundas de vasto material textual.

Dentro do eixo de desenvolvimento (programação), se sobressai a ideia de adaptar a entrada da ferramenta EχATO$_{LP}$ a textos anotados por outros *parsers*, como é o caso da ferramenta LX-Center [177]. Esse tipo de experiência irá permitir a verificação prática de que todo o processo proposto nessa tese pode ser empregado com qualquer ferramenta de anotação linguística. Cabe salientar que essa linha de trabalho futuro implica em uma adaptação quase que exclusivamente de programação, pois o processo proposto de extração, ordenação e identificação descritos nos Capítulos de 3 a 5, não necessita de alterações teóricas em suas propostas. Apesar disso, esse trabalho futuro é de grande relevância prática, por aumentar o escopo de aplicação da ferramenta EχATO$_{LP}$.

Dentro do eixo de extração de termos, existe a ideia de adaptar todas as propostas dessa tese para a aplicação em outras línguas além do Português. Algumas das etapas devem ser diretamente aplicáveis com pouca ou nenhuma alteração necessária, como é o caso do índice *tf-dcf* e, provavelmente, da análise de pontos de corte. No que diz respeito às heurísticas utilizadas na extração de termos, a adaptação a ser feita será obrigatoriamente grande, pois as construções linguísticas são por definição dependentes do idioma. Por outro lado, toda a parte de aplicações é completamente independente da língua. Dessa forma, a extensão das propostas feitas nessa tese para tratar *corpora* de domínio em outros idiomas não deve representar um trabalho futuro muito complexo, mas que pode ampliar enormemente o escopo de aplicação das contribuições dessa tese.

Dentro do eixo de construção automática de ontologias, o caminho natural a seguir após essa tese é dar sequência ao processo de construção automática de ontologias. Essa continuação implica observar em detalhe, e com um olhar mais científico, a proposta inicial de construção de hierarquias (Seção 6.4). Também é importante sugerir novas aplicações dos conceitos extraídos, além das quatro já disponibilizadas e descritas no Capítulo 6. Em seguida, parece intuitivo buscar nas informações dos termos extraídos, os contextos verbais nos quais os conceitos apareceram de forma a deduzir relações não taxonômicas. Também parece ser viável deduzir relações e diferenciações entre conceitos e instâncias (população de ontologias), de forma a obter pelo menos uma ontologia elementar conforme definição formal feita na Seção 2.2.1 Esse eixo de trabalhos futuros, apesar de ambicioso, parece viável pela quantidade de informação disponibilizada pelo processo de extração proposto, além de promissor pela qualidade dos conceitos extraídos.

Referências Bibliográficas

[1] Muhammad Abdul-Mageed, *Automatic detection of arabic non-anaphoric pronouns for improving anaphora resolution*, ACM Transactions on Asian Language Information Processing (TALIP), 10 (2011), pp. 5:1–5:11.

[2] Otavio Costa Acosta, Aline Villavicencio, e Viviane P. Moreira, *Identification and treatment of multiword expressions applied to information retrieval*, in Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World, MWE '11, Stroudsburg, PA, USA, 2011, Association for Computational Linguistics, pp. 101–109.

[3] Alfred V. Aho e Jeffrey D. Ullman, *The theory of parsing, translation, and compiling*, Prentice-Hall, Inc., Upper Saddle River, USA, 1972.

[4] Salah. Aït-Mokhtar, Jean-Pierre Chanod, e Claude Roux, *Robustness beyond shallowness: incremental deep parsing*, Natural Language Engineering, 8 (2002), pp. 121–144.

[5] Robert A. Amsler, *A taxonomy for english nouns and verbs*, in Proceedings of the 19th Annual Meeting of the Association for Computational Linguistics, Stanford, USA, June 1981, Association for Computational Linguistics, pp. 133–138.

[6] *Ask.com - what's your question?* `http://www.ask.com`, Dezembro 2011. (último acesso em 13 dezembro 2011).

[7] Ruba Awawdeh e Terry Anderson, *Improving search in tag-based systems with automatically extracted keywords*, in Knowledge Science, Engineering and Management, Yaxin Bi e Mary-Anne Williams, eds., vol. 6291 of Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2010, pp. 378–387.

[8] *Yahoo babelfish - text translation.* `http://babelfish.yahoo.com`, Dezembro 2011. (último acesso em 13 dezembro 2011).

[9] Xin Bai e Xiang Zhen Zhou, *Development of ontology-based information system using formal concept analysis and association rules*, in Advances in Computer Science, Intelligent System and Environment, David Jin e Sally Lin, eds., vol. 106 of Advances in Intelligent and Soft Computing, Springer Berlin / Heidelberg, 2011, pp. 121–126.

[10] Timothy Baldwin e Aline Villavicencio, *Extracting the unextractable: A case study on verb-particles*, in Proceedings of CoNLL-2002, Taipei, Taiwan, 2002, pp. 98–104.

[11] Satanjeev Banerjee e Ted Pedersen, *The design, implementation and use of the ngram statistics package*, in 4th ITPCL, 2003, pp. 370–381.

[12] Jorge Baptista, Fernando Batista, e Nuno Mamede, *Building a dictionary of anthroponyms*, in Computational Processing of the Portuguese Language, Renata Vieira, Paulo Quaresma, Maria Nunes, Nuno Mamede, Cláudia Oliveira, e Maria Dias, eds., vol. 3960 of Lecture Notes in Computer Science, Springer, Berlin/Heidelberg, Germany, 2006, pp. 21–30. 10.1007/11751984-3.

[13] Yehoshua Bar-Hillel, *The present status of automatic translation of languages*, in Advances in Computers, F. L. Alt, ed., vol. I, Academic Press, New York, USA, 1960, pp. 91–163.

[14] M. Baroni e S. Bernardini, *Bootcat: Bootstrapping corpora and terms from the web*, in Proceedings of the 4th Language Resources and Evaluation Conference (LREC), Lisbon, Portugal, 2004.

[15] Tulio L. Basegio, *Uma abordagem semi-automática para identificação de estruturas ontológicas a partir de textos na língua portuguesa do Brasil*, master's thesis, PUCRS, Porto Alegre, Brazil, 2007.

[16] Nikoletta Bassiou e Constantine Kotropoulos, *Long distance bigram models applied to word clustering*, Pattern Recognition, 44 (2011), pp. 145 – 158.

[17] Daniel Emilio Beck, *Syntax-based statistical machine translation using tree automata and tree transducers*, in ACL (Student Session), The Association for Computer Linguistics, 2011, pp. 36–40.

[18] *Projeto benveniste on-line*. `http://www6.ufrgs.br/letras/benvenisteonline/`, April 2011. (último acesso em 13 dezembro 2011).

[19] Tim Berners-Lee, James Hendler, e Ora Lassila, *The semantic web*, Scientific American, 284 (2001), pp. 34–43.

[20] Eckhard Bick, *The parsing system PALAVRAS: automatic grammatical analysis of portuguese in constraint grammar framework*, PhD thesis, Arhus University, Arhus, Danemark, 2000.

[21] Chris Biemann, *Ontology learning from text: A survey of methods*, LDV Forum, 20 (2005), pp. 75–93.

[22] Simon Blackburn, *The Oxford Dictionary of Philosophy*, Oxford University Press, Oxford, UK, 1994.

[23] Francesca Bonin, Felice Dell'Orletta, Giulia Venturi, e Simonetta Montemagni, *Contrastive filtering of domain-specific multi-word terms from different types of corpora*, in Proceedings of the Workshop on Multiword Expressions: from Theory to Applications (MWE 2010), Beijing, China, August 2010, Association for Computational Linguistics, pp. 76–79.

[24] Susan Bonzi e Elizabeth DuRoss Liddy, *Testing the assumption underlying use of anaphora in natural language tests*, in Proceedings of the 51st ASIS Annual Meeting (ASIS '88), Christine L. Borgman e Edward Y. H. Pai, eds., vol. 25, Atlanta, Georgia, 1988, American Society for Information Science.

[25] J. S. Boreczky e L. A. Rowe, *Comparison of video shot boundary detection techniques*, Journal of Electronic Imaging, 5 (1996), pp. 122–128.

[26] Wauter Bosma e Piek Vossen, *Bootstrapping language neutral term extraction*, in Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, e Daniel Tapias, eds., Valletta, Malta, may 2010, European Language Resources Association (ELRA).

[27] Didier Bourigault, Marie-Paule Jacques, Cécile Fabre, Cécile Frérot, e Sylwia Ozdowska, *Syntex, analyseur syntaxique de corpus*, in Actes des 12èmes Journées sur le traitement automatique des langues naturelles, 2005.

[28] Didier Bourigault e Guiraude Lame, *Analyse distributionnelle et structuration de terminologie. application a la construction d'une ontologie documentaire du droit*, Traitement automatique des langues, 43 (2002).

[29] J. Bresnan e R. M. Kaplan, *Introduction: Grammars as mental representations of language*, in The Mental Representation of Grammatical Relations, J. Bresnan, ed., MIT Press, Cambridge, MA, 1982, pp. 27–52.

[30] Rainer Brüggemann e Ganapati P. Patil, *Formal concept analysis*, in Ranking and Prioritization for Multi-indicator Systems, G. P. Patil, ed., vol. 5 of Environmental and Ecological Statistics, Springer New York, 2011, pp. 117–133.

[31] Quoc-Chinh Bui e Peter M.A. Sloot, *Extracting biological events from text using simple syntactic patterns*, in Proceedings of BioNLP Shared Task 2011 Workshop, Association for Computational Linguistics, 2011, pp. 143–146.

[32] Paul Buitelaar, Philipp Cimiano, e Bernardo Magnini, *Ontology learning from text: An overview*, in Ontology Learning from Text: Methods, Evaluation and Applications, Paul Buitelaar, Philipp Cimiano, e Bernardo Magnini, eds., vol. 123 of Frontiers in Artificial Intelligence and Applications, IOS Press, 2005.

[33] L. Cai e T. Hofmann, *Hierarchical document categorization with support vector machines*, in 13th CIKM, ACM, 2004, pp. 78–87.

[34] Nuno Caminada, Violeta Quental, e Milena Garrão, *Linguistics tools: uma plataforma expansível de funções de consulta a corpus*, in Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web, WebMedia '08, New York, NY, USA, 2008, ACM, pp. 364–368.

[35] Sharon A. Caraballo, *Automatic construction of a hypernym-labeled noun hierarchy from text*, in Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, ACL '99, Stroudsburg, PA, USA, 1999, Association for Computational Linguistics, pp. 120–126.

[36] Fernando M.B.M. Castilho, Roger L. Granada, Renata Vieira, Tomas Sander, e Prasad Rao, *Ontology enrichment based on the mapping of knowledge resources for data privacy management*, in Proceedings of Joint IV Seminar on Ontology Research in Brazil and VI International Workshop on Metamodels, Ontologies and Semantic Technologies (ONTOBRAS-MOST 2011), CEUR, 2011, pp. 85–96.

[37] Vinay K. Chaudhri, Adam Farquhar, Richard Fikes, Peter D. Karp, e James P. Rice, *Okbc: a programmatic foundation for knowledge base interoperability*, in Proceedings of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence, AAAI '98/IAAI '98, Menlo Park, USA, 1998, American Association for Artificial Intelligence, pp. 600–607.

[38] Chaitanya Chemudugunta, America Holloway, Padhraic Smyth, e Mark Steyvers, *Modeling documents by combining semantic concepts with unsupervised statistical learning*, in The Semantic Web - ISWC 2008, Amit Sheth, Steffen Staab, Mike Dean, Massimo Paolucci, Diana Maynard, Timothy Finin, e Krishnaprasad Thirunarayan, eds., vol. 5318 of Lecture Notes in Computer Science, Springer, Berlin/Heidelberg, Germany, 2008, pp. 229–244.

[39] Noan Chomsky, *Syntactic Structures*, Mouton, The Hague, The Neederlands, 1957.

[40] Noan Chomsky, *Aspects of the theory of syntax*, MIT Press, Cambridge, USA, 1965.

[41] Teresa M. Chung, *A corpus comparison approach for terminology extraction*, Terminology, 9 (2003), pp. 221–246.

[42] Philipp Cimiano, *Ontology learning and population from text: algorithms, evaluation and applications*, Springer, London, UK, 2006.

[43] Philipp Cimiano, Andreas Hotho, e Steffen Staab, *Learning concept hierarchies from text corpora using formal concept analysis*, Journal of Artificial Intelligence Research, 24 (2005), pp. 305–339.

[44] Philipp Cimiano, Johanna Völker, e Rudi Studer, *Ontologies on demand? - a description of the state-of-the-art, applications, challenges and trends for ontology learning from text*, Information, Wissenschaft und Praxis, 57 (2006), pp. 315–320.

[45] K. M. Colby, *Simulation of belief systems*, in Computer Models of Thought and Language, R.C. Schank e K.M. Colby, eds., W. H. Freeman and Company, San Francisco, USA, 1973, pp. 251–286.

[46] Alain Colmerauer, *Total precedence relations*, Journal of the ACM, 17 (1970), pp. 14–30.

[47] Alain Colmerauer e Philippe Roussel, *The birth of Prolog*, in History of Programming Languages – II, Thomas J. Bergin Jr. e Richard G. Gibson, Jr., eds., ACM Press/Addison-Wesley, New York, USA, 1996, pp. 331–352.

[48] Mike Conway, Son Doan, Ai Kawazoe, e Nigel Collier, *Classifying disease outbreak reports using n-grams and semantic features*, International Journal of Medical Informatics, 78 (2009), pp. e47 – e58. ¡ce:title¿Mining of Clinical and Biomedical Text and Data Special Issue¡/ce:title¿.

[49] Robert James Coulthard, *The application of Corpus Methodology to Translation: the JPED parallel corpus and the Pediatrics comparable corpus*, PhD thesis, UFSC, Florianópolis, Brazil, 2005.

[50] W. Bruce Croft e David J. Harper, *Using probabilistic models of document retrieval without relevance information*, Journal of documentation, 35 (1979), pp. 285–295.

[51] Marie-Catherine De Marneffe, Bill MacCartney, e Christopher D. Manning, *Generating typed dependency parses from phrase structure parses*, in LREC 2006, 2006.

[52] Rogerio Figueredo de Sousa, Rafael Torres Anchiêta, Francisco A. Ricarte Neto, e Raimundo S. Moura, *Uso de pln com a abordagem estatística para*

identificar palavras chaves em artigos científicos, in Anais da Escola Regional de Computação Ceará – Maranhão – Piauí, ERCEMAPI 2011, Teresina, Piauí, Brasil, 2011, UFPI.

[53] Ariani Di Felippo, *The terminet project: an overview*, in Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas, YIWCALA '10, Stroudsburg, PA, USA, 2010, Association for Computational Linguistics, pp. 92–99.

[54] Jiandong Ding, Shuigeng Zhou, e Jihong Guan, *mirfam: an effective automatic mirna classification method based on n-grams and a multiclass svm*, BMC Bioinformatics, 12 (2011), p. 216.

[55] Patrick Drouin, *Detection of domain specific terminology using corpora comparison*, in Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC) 2004, Maria Teresa Lino, Maria Francisca Xavier, Fatima Ferreira, Rute Costa, e Raquel Silva, eds., Lisbon, Portugal, May 2004, ELRA, European Language Resources Association, pp. 79–82.

[56] Lucas Drumond e Rosario Girardi, *Extracting ontology concept hierarchies from text using markov logic*, in Proceedings of the 2010 ACM Symposium on Applied Computing, SAC '10, New York, USA, 2010, ACM, pp. 1354–1358.

[57] Elizabeth Du Ross Liddy, *Anaphora in natural language processing and information retrieval*, Information Processing Management, 26 (1990), pp. 39–52.

[58] Elizabeth Du Ross Liddy, *Natural Language Processing*, Encyclopedia of Library and Information Science, Marcel Dekker Inc., New York, USA, 2nd ed., 2003.

[59] Susan Dumais, John Platt, David Heckerman, e Mehran Sahami, *Inductive learning algorithms and representations for text categorization*, in Proceedings of the seventh international conference on Information and knowledge management, CIKM '98, New York, USA, 1998, ACM, pp. 148–155.

[60] Marc Ehrig, *Ontology Alignment: Bridging the Semantic Gap*, vol. 4 of Semantic Web And Beyond Computing for Human Experience, Springer, Amsterdan, The Netherlands, 2007.

[61] Jerome Euzenat e Pavel Shvaiko, *Ontology Matching*, Springer-Verlag, Berlin, Germany, 2007.

[62] Stefan Evert, *Google web 1t 5-grams made easy (but not for the computer)*, in Proceedings of the NAACL HLT 2010 Sixth Web as Corpus Workshop, WAC-6 '10, Stroudsburg, PA, USA, 2010, Association for Computational Linguistics, pp. 32–40.

[63] Carla Faria e Rosario Girardi, *An information extraction process for semi-automatic ontology population*, in Soft Computing Models in Industrial and Environmental Applications, 6th International Conference SOCO 2011, Emilio Corchado, Václav Sná?el, Javier Sedano, Aboul Hassanien, José Calvo, e Dominik Slezak, eds., vol. 87 of Advances in Intelligent and Soft Computing, Springer Berlin / Heidelberg, 2011, pp. 319–328. 10.1007/978-3-642-19644-7_34.

[64] David Faure e Claire Nédellec, *A corpus-based conceptual clustering method for verb frames and ontology acquisition*, in Proceedings of the LREC Workshop on Adapting Lexical and Corpus Resources to Sublanguages and Applications, 1998, pp. 5–12.

[65] Paulo Fernandes, Lucelene Lopes, e Duncan D. A. Ruiz, *The impact of random samples in ensemble classifiers*, in SAC'10: Proceedings of the 2010 ACM Symposium on Applied Computing, New York, USA, 2010, ACM, pp. 1002–1009.

[66] C. J. Filmore, *Lexical entries for verb*, D. Reidel, Dordrecht, Holland, 1968.

[67] Maria J. Finatto, Lucelene Lopes, Renata Vieira, e Aline Evers, *Hierarquias de conceitos para um ambiente virtual de ensino extraídas de um corpus de jornais populares*, in Proceedings of Joint IV Seminar on Ontology Research in Brazil and VI International Workshop on Metamodels, Ontologies and Semantic Technologies (ONTOBRAS-MOST 2011), CEUR, 2011, pp. 111–116.

[68] John R. Firth, *A Synopsis of Linguistic Theory, 1930-1955*, Studies in Linguistic Analysis, (1957), pp. 1–32.

[69] Felipe Flores, Viviane Moreira, e Carlos Heuser, *Assessing the impact of stemming accuracy on information retrieval*, in Computational Processing of the Portuguese Language, Thiago Pardo, António Branco, Aldebaro Klautau, Renata Vieira, e Vera de Lima, eds., vol. 6001 of Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2010, pp. 11–20. 10.1007/978-3-642-12320-7_2.

[70] Blaz Fortuna, Marko Grobelnik, e Dunja Mladenic, *Ontogen: semi-automatic ontology editor*, in Proceedings of the 2007 conference on Human interface: Part II, Berlin/Heidelberg, Germany, 2007, Springer-Verlag, pp. 309–318.

[71] Blas Fortuna, Nada Lavrac, e Paola Velardi, *Advancing topic ontology learning through term extraction*, in PRICAI 2008: Trends in Artificial Intelligence, Tu-Bao Ho e Zhi-Hua Zhou, eds., vol. 5351 of Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2008, pp. 626–635. 10.1007/978-3-540-89197-0_57.

[72] Hermine Njike Fotzo e Patrick Gallinari, *Learning "generalization/specialization" relations between concepts - application for automatically building thematic document hierarchies*, in RIAO, 2004, pp. 143–155.

[73] Katerina Frantzi, Sophia Ananiadou, e Hideki Mima, *Automatic recognition of multi-word terms:. the c-value/nc-value method*, International Journal on Digital Libraries, 3 (2000), pp. 115–130. 10.1007/s007999900023.

[74] Maria Claudia de Freitas, *Elaboração automática de ontologias de domínio: discussão e resultados*, PhD thesis, PUC-Rio, Rio de Janeiro, Brazil, 2007.

[75] R. Garside, G. Leech, e G. Sampson, *The Computational Analysis of English: A Corpus Based Approach*, Longman, London, UK, 1987.

[76] Alexander F. Gelbukh e Grigori Sidorov, *Zipf and heaps laws' coefficients depend on language*, in Proceedings of the Second International Conference on Computational Linguistics and Intelligent Text Processing, CICLing '01, London, UK, 2001, Springer-Verlag, pp. 332–335.

[77] John H. Gennari, Mark A. Musen, Ray W. Fergerson, William E. Grosso, Monica Crubézy, Henrik Eriksson, Natalya F. Noy, e Samson W. Tu, *The evolution of protégé: an environment for knowledge-based systems development*, International Journal of Human-Computer Studies, 58 (2003), pp. 89–123.

[78] Ana Catarina Gianoti e Ariani Di Felippo, *Descrição morfológica preliminar dos termos da educação a distância*, Tech. Report NILC-TR-11-02, NILC - ICMC-USP, São Carlos, SP, Brasil, 2011.

[79] Ana Catarina Gianoti e Ariani Di Felippo, *Extração de conhecimento terminológico no projeto terminet*, Tech. Report NILC-TR-11-01, NILC - ICMC-USP, São Carlos, SP, Brasil, 2011.

[80] *Language tools*. `http://www.google.com/language_tools`, Dezembro 2011. (último acesso em 13 dezembro 2011).

[81] Gregory Grefenstette, *Explorations in Automatic Thesaurus Discovery*, Kluwer Academic Publishers, Norwell, MA, USA, 1994.

[82] Barbara J. Grosz, *The representation and use of focus in a system for understanding dialogs*, in IJCAI, 1977, pp. 67–76.

[83] Barbara J. Grosz e Candace L. Sidner, *Attention, intentions, and the structure of discourse*, Comput. Linguist., 12 (1986), pp. 175–204.

[84] Thomas Gruber, *Toward principles for the design of ontologies used for knowledge sharing*, International Journal Human-Computer Studies, 43 (1993), pp. 907–928.

[85] Nicola Guarino e Luc Schneider, *Ontology-driven conceptual modelling*, in ER, Stefano Spaccapietra, Salvatore T. March, e Yahiko Kambayashi, eds., vol. 2503 of Lecture Notes in Computer Science, Springer, 2002, p. 10.

[86] *hakia.com*. `http://www.hakia.com`, Dezembro 2011. (último acesso em 13 dezembro 2011).

[87] Z. S. Harris, *Mathematical Structures of Language*, Wiley, New York, USA, 1968.

[88] Marti A. Hearst, *Automatic acquisition of hyponyms from large text corpora*, in Proceedings of the 14th conference on Computational linguistics - Volume 2, COLING '92, Stroudsburg, USA, 1992, Association for Computational Linguistics, pp. 539–545.

[89] Gary G. Herdrix, *Human engineering fcr applied natural language processing*, in Proceedings of the 5th international joint conference on Artificial intelligence - Volume 1, San Francisco, USA, 1977, Morgan Kaufmann Publishers Inc., pp. 183–191.

[90] Donald Hindle, *Noun classification from predicate-argument structures*, in Proceedings of the 28th annual meeting on Association for Computational Linguistics, ACL '90, Stroudsburg, PA, USA, 1990, Association for Computational Linguistics, pp. 268–275.

[91] Anette Hulth, *Enhancing linguistically oriented automatic keyword extraction*, in Proceedings of HLT-NAACL 2004: Short Papers, HLT/NAACL, New York, USA, 2004, ACM, pp. 17–20.

[92] Dell H. Hymes, *Competence and performance in linguistic theory*, in Language acquisition: Models and methods, R. Huxley e E. Ingrams, eds., Academic Press, London, UK, 1971, pp. 3–28.

[93] Ray Jackendoff, *Semantic Interpretation in Generative Grammar*, The MIT Press Classics, Cambridge, USA, 1972.

[94] Daniel Jurafsky e James H. Martin, *Speech and Language Processing*, Prentice-Hall, Inc., Upper Saddle River, USA, 2nd ed., 2009.

[95] Kyo Kageura e Bin Umino, *Methods of automatic term recognition - a review -*, Terminology, 3 (1996), pp. 259–289.

[96] G. Karypis, *Cluto: a clustering toolkit*, Tech. Report 02-017, University of Minnesota, 2002. `http://glaros.dtc.umn.edu/gkhome/views/cluto/`.

[97] Martin Kavalec e Vojtěch Sv Átek, *A study on automated relation labelling in ontology learning*, in Ontology Learning from Text: Methods, Evaluation and Applications, IOS Press, 2005, pp. 44–58.

[98] Paul Kay e Chad K. McDaniel, *On the logic of variable rules*, Language in Society, 8 (1979), pp. 151–187.

[99] J. Kietz, R. Volz, e A. Maedche, *Extracting a domain-specific ontology from a corporate intranet*, in Proceedings of the 2nd Workshop on Learning Language in Logic and the 4th Conference on Computational Natural Language Learning, vol. 7, Morristown, USA, 2000, Association for Computational Linguistics, pp. 167–175.

[100] A. Kilgarriff, M. Rundell, e E. U. Dhonnchadha, *Efficient corpus development for lexicography: building the new corpus for ireland*, Language Resources and Evaluation, 40 (2006), pp. 127–152.

[101] Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi, e Jun'ichi Tsujii, *Genia corpus ,Äîa semantically annotated corpus for bio-textmining*, Bioinformatics, 19 (2003), pp. i180–i182.

[102] Su Nam Kim, Timothy Baldwin, e Min-Yen Kan, *Extracting domain-specific words - a statistical approach*, in Proceedings of the 2009 Australasian Language Technology Association Workshop, Luiz Pizzato e Rolf Schwitter, eds., Sydney, Australia, December 2009, Australasian Language Technology Association, pp. 94–98.

[103] Chunyu Kit e Xiaoyue Liu, *Measuring mono-word termhood by rank difference via corpus comparison*, Terminology, 14 (2008), pp. 204–229.

[104] Alexandre Kouznetsov, Jonas B. Laurila, Christopher J. O. Baker, e Bradley Shoebottom, *Algorithm for population of object property assertions derived from telecom contact centre product support documentation*, in Proceedings of the 2011 IEEE Workshops of International Conference on Advanced Information Networking and Applications, WAINA '11, Washington, DC, USA, 2011, IEEE Computer Society, pp. 41–46.

[105] Teruo Koyama e Koichi Takeuchi, *Enhancing multi-word term extraction for designated theme embedded in a domain corpus*, in Proceedings of the 9th International Conference on Terminology and Artificial Intelligence, Kyo Kageura e Pierre Zweigenbaum, eds., Paris, France, November 2011, INALCO, pp. 73–79.

[106] H. Kucera e W. N. Francis, *Computational analysis of present-day American English*, Brown University Press, Providence, USA, 1967.

[107] Hélio Kuramoto, *Uma abordagem alternativa para o tratamento e a recuperação de informação textual : os sintagmas nominais*, Revista Ciência da Informação, 25 (1996).

[108] Hélio Kuramoto, *Nominal groups: a new purpose to information retrieval*, DataGramaZero - Revista de Ciência da Informação, 3 (2002).

[109] John Lamping, Ramana Rao, e Peter Pirolli, *A focus+context technique based on hyperbolic geometry for visualizing large hierarchies*, in Proceedings of the SIGCHI conference on Human factors in computing systems, CHI '95, New York, NY, USA, 1995, ACM Press/Addison-Wesley Publishing Co., pp. 401–408.

[110] Shalom Lappin e Herbert J. Leass, *An algorithm for pronominal anaphora resolution*, Comput. Linguist., 20 (1994), pp. 535–561.

[111] Alberto Lavelli, Fabrizio Sebastiani, e Roberto Zanoli, *Distributional term representations: an experimental comparison*, in CIKM, 2004, pp. 615–624.

[112] Stefania Leone, Matthias Geel, e Moira C. Norrie, *The use of tag clouds to support the discovery and inspection of information services*, in Proceedings of the 5th International Workshop on Web APIs and Service Mashups, Mashups '11, New York, NY, USA, 2011, ACM, pp. 10:1–10:6.

[113] M. Lesk, *Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone*, in Proceedings of the 5th annual international conference on Systems documentation, ACM, 1986, pp. 24–26.

[114] D. D. Lewis, *An evaluation of phrasal and clustered representations on a text categorization task*, in 15th SIGIR, ACM, 1992, pp. 37–50.

[115] *Lexxe search engine*. `http://www.lexxe.com`, Dezembro 2011. (último acesso em 13 dezembro 2011).

[116] *Linguateca - ferramentas para português*. `http://www.linguateca.pt/ferramentas_info.html`, May 2011. (último acesso em 13 dezembro 2011).

[117] Lucelene Lopes, Paulo Fernandes, Renata Vieira, e Guilherme Fedrizzi, *ExATO lp – An Automatic Tool for Term Extraction from Portuguese Language Corpora*, in Proceedings of the 4th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC '09), Poznan, Poland, November 2009, Faculty of Mathematics and Computer Science of Adam Mickiewicz University, Adam Mickiewicz University, pp. 427–431.

[118] Lucelene Lopes, Paulo Fernandes, Renata Vieira, Guilherme Fedrizzi, e Daniel Martins, *Exatolp - a tool for domain relevant terms extraction*, in PROPOR 2010 – International Conference on Computational Processing of Portuguese Language, 2010.

[119] Lucelene Lopes, Leandro H. Oliveira, e Renata Vieira, *Portuguese term extraction methods: Comparing linguistic and statistical approaches*, in PROPOR 2010 – International Conference on Computational Processing of Portuguese Language, 2010.

[120] Lucelene Lopes e Renata Vieira, *Building Domain Specific Corpora in Portuguese Language*, Tech. Report TR 062, Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), Porto Alegre, Brazil, Dezembro 2010.

[121] Lucelene Lopes e Renata Vieira, *Processamento de linguagem natural e o tratamento computacional de linguagens científicas*, in Linguagens Especializadas em Corpora: modos de dizer e interfaces de pesquisa, Cristina Lopes Perna, Heloísa Koch Delgado, e Maria José Finatto, eds., EDIPUCRS, Porto Alegre, Brazil, 2010, pp. 183–201.

[122] Lucelene Lopes e Renata Vieira, *Improving quality of portuguese term extraction*, in PROPOR 2012 – International Conference on Computational Processing of Portuguese Language, 2012.

[123] Lucelene Lopes, Renata Vieira, Maria José Finatto, e Daniel Martins, *Extracting compound terms from domain corpora*, Journal of the Brazilian Computer Society, 16 (2010), pp. 247–259. 10.1007/s13173-010-0020-4.

[124] Lucelene Lopes, Renata Vieira, Maria J. Finatto, Adriano Zanette, Daniel Martins, e Luis Carlos Ribeiro Jr., *Automatic extraction of composite terms for construction of ontologies: an experiment in the health care area*, RECIIS, 3 (2009), pp. 72–84.

[125] Lucelene Lopes, Renata Vieira, e Daniel Martins, *Hierarquias de conceitos extraídas automaticamente de corpus de domínio específico - um experimento sobre um corpus de pediatria*, in XII Congresso Brasileiro de Informática em Saúde (CBIS), Sociedade Brasileira de Informática em Saúde, 2010, pp. 1–6.

[126] Alexander Maedche e Steffen Staab, *Learning ontologies for the semantic web*, in SemWeb, 2001.

[127] Luiz Cláudio Maia e Renato Rocha Souza, *Uso de sintagmas nominais na classificação automática de documentos eletrônicos*, Perspectivas em Ciência da Informação, 15 (2010), pp. 154–172.

[128] Jawad Makki, Anne-Marie Alquier, e Violaine Prince, *Semi automatic ontology instantiation in the domain of risk management*, in Intelligent Information Processing IV, Zhongzhi Shi, E. Mercier-Laurent, e D. Leake, eds., vol. 288 of IFIP Advances in Information and Communication Technology, Springer Boston, 2008, pp. 254–265. 10.1007/978-0-387-87685-6_30.

[129] Mann, William C. and Thompson, Sandra A., *Rhetorical Structure Theory: Toward a functional theory of text organization*, Text, 8 (1988), pp. 243–281.

[130] Christopher D. Manning, Prabhakar Raghavan, e Hinrich Schütze, *Introduction to Information Retrieval*, Cambridge University Press, Cambridge, 2008.

[131] Christopher D. Manning e Hinrich Schütze, *Foundations of statistical natural language processing*, MIT Press, Cambridge, USA, 1999.

[132] Mitchell P. Marcus, Mary Ann Marcinkiewicz, e Beatrice Santorini, *Building a large annotated corpus of english: the penn treebank*, Computational Linguistics, 19 (1993), pp. 313–330.

[133] David L. Martin, Massimo Paolucci, Sheila A. McIlraith, Mark H. Burstein, Drew V. McDermott, Deborah L. McGuinness, Bijan Parsia, Terry R. Payne, Marta Sabou, Monika Solanki, Naveen Srinivasan, e Katia P. Sycara, *Bringing semantics to web services: The owl-s approach*, in SWSWPC, 2004, pp. 26–42.

[134] Diana Maynard, Yaoyong Li, e Wim Peters, *Nlp techniques for term extraction and ontology population*, in Proceedings of the 2008 conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge, Amsterdam, The Netherlands, The Netherlands, 2008, IOS Press, pp. 107–127.

[135] D. D. McDonald, *Natural language generation as a computational problem*, in Computational Models of Discourse, MIT Press, Cambridge, USA, 1983, pp. 209–265.

[136] K. R. McKeown, *Text generation: using discourse strategies and focus constraints to generate natural language text*, Cambridge University Press, New York, USA, 1985.

[137] O. Medelyan e Ian H. Witten, *Domain-independent automatic keyphrase indexing with small training sets*, Journal of the American Society for Information Science and Technology, 59 (2008), pp. 1026–1040.

[138] E. Milios, Y. Zhang, B. He, e L. Dong, *Automatic term extraction and document similarity in special text corpora*, in 6th Conference of the Pacific Association for Computational Linguistics, Halifax, Nova Scotia, Canada, Aug. 2003, pp. 275–284.

[139] George A. Miller e Walter G. Charles, *Contextual correlates of semantic similarity*, Language and Cognitive Processes, 6 (1991), pp. 1–28.

[140] T. Mitchell, *Machine Learning*, McGraw-Hill, 1997.

[141] A. E. Monge e C. Elkan, *The field matching problem: Algorithms and applications*, in Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), 1996, pp. 267–270.

[142] Emmanuel Morin e Christian Jacquemin, *Projecting corpus-based semantic links on a thesaurus*, in Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, ACL '99, Stroudsburg, USA, 1999, Association for Computational Linguistics, pp. 389–396.

[143] Vít Nováček, *Ontology Learning*, PhD thesis, Brno University, Brno, Czech Republic, 2005.

[144] *Ontogen - semiautomatic ontology editor*. `http://ontogen.ijs.si`, May 2011. (último acesso em 13 dezembro 2011).

[145] Jiaul H. Paik, Mandar Mitra, Swapan K. Parui, e Kalervo Järvelin, *Gras: An effective and efficient stemming algorithm for information retrieval*, ACM Transaction on Information Systems, 29 (2011), pp. 19:1–19:24.

[146] Patrick Pantel e Dekang Lin, *A statistical corpus-based term extractor*, in Proceedings of the 14th Biennial Conference of the Canadian Society on Computational Studies of Intelligence: Advances in Artificial Intelligence, New York, USA, 2001, ACM Press, pp. 36–46.

[147] Miranda Lee Pao, *Automatic text analysis based on transition phenomena of word occurrences*, Journal of the American Society for Information Science, 29 (1978), pp. 121–124.

[148] Youngja Park, Siddharth Patwardhan, Karthik Visweswariah, e Stephen C. Gates, *An empirical analysis of word error rate and keyword error rate*, in INTERSPEECH, 2008, pp. 2070–2073.

[149] Marius Pasca, *Acquisition of categorized named entities for web search*, in Proceedings of the thirteenth ACM international conference on Information and knowledge management, CIKM '04, New York, USA, 2004, ACM, pp. 137–145.

[150] Fernando Pereira, Naftali Tishby, e Lillian Lee, *Distributional clustering of english words*, in Proceedings of the 31st annual meeting on Association for Computational Linguistics, ACL '93, Stroudsburg, PA, USA, 1993, Association for Computational Linguistics, pp. 183–190.

[151] F. C. N. Pereira e D. H. D. Warren, *Definite clause grammars for language analysis - A survey of the formalism and a comparison with augmented transition networks*, Artificial Intelligence, 13 (1980), pp. 231–278.

[152] M. A. Perini, *Princípios de linguística descritiva: introdução ao pensamento gramatical*, Parábola, São Paulo, Brazil, 2007.

[153] Georgios Petasis, Vangelis Karkaletsis, Georgios Paliouras, Anastasia Krithara, e Elias Zavitsanos, *Ontology population and enrichment: State of the art*, in Knowledge-Driven Multimedia Information Extraction and Ontology Evolution, Georgios Paliouras, Constantine Spyropoulos, e George Tsatsaronis, eds., vol. 6050 of Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2011, pp. 134–166. 10.1007/978-3-642-20795-2_6.

[154] Georgios Petasis, Vangelis Karkaletsis, Georgios Paliouras, Anastasia Krithara, e Elias Zavitsanos, *Ontology population and enrichment: State of the art*, in Knowledge-Driven Multimedia Information Extraction and Ontology Evolution, Georgios Paliouras, Constantine Spyropoulos, e George Tsatsaronis, eds., vol. 6050 of Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2011, pp. 134–166. 10.1007/978-3-642-20795-2_6.

[155] Vladia Pinheiro, Vasco Furtado, Tarcisio H. C. Pequeno, e Douglas Nogueira, *Natural language processing based on semantic inferentialism for extracting crime information from text*, in ISI, Christopher C. Yang, Daniel Zeng, Ke Wang, Antonio Sanfilippo, Herbert H. Tsang, Min-Yuh Day, Uwe Glässer, Patricia L. Brantingham, e Hsinchun Chen, eds., IEEE, 2010, pp. 19–24.

[156] *Projeto porpopular*. `http://www6.ufrgs.br/textecc/porlexbras/porpopular/`, April 2011. (último acesso em 13 dezembro 2011).

[157] *Protégé ontology editor and knowledge acquisition system*. `http://protege.stanford.edu`, May 2011. (último acesso em 13 dezembro 2011).

[158] Sun Qiao, Zhang Chunhui, e Chen Zhibo, *Automatic construction of domain concept hierarchy*, in International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, Los Alamitos, USA, 2010, IEEE Computer Society, pp. 433–436.

[159] Paulo Quaresma e Teresa Gonçalves, *Using linguistic information and machine learning techniques to identify entities from juridical documents*, in Semantic Processing of Legal Texts, Enrico Francesconi, Simonetta Montemagni, Wim Peters, e Daniela Tiscornia, eds., vol. 6036 of Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2010, pp. 44–59.

[160] M. R. Quillian, *Semantic memory*, in Semantic Information Processing, M. Minsky, ed., MIT Press, Cambridge, USA, 1968, pp. 227–270.

[161] L. R. Rabiner e B. H. Juang, *An introduction to hidden Markov models*, IEEE ASSP Magazine, January. 4-15, Los Alamitos, USA, 1986.

[162] Carlos Ramisch, Aline Villavicencio, e Christian Boitet, *Multiword expressions in the wild?: the mwetoolkit comes in handy*, in Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations, COLING '10, Stroudsburg, PA, USA, 2010, Association for Computational Linguistics, pp. 57–60.

[163] Júlio Cesar Reis, Rodrigo Bonacin, e Maria Cecília Calani Baranauskas, *Identificando semântica em redes sociais inclusivas online: Um estudo sobre ferramentas e técnicas*, Tech. Report IC-10-28, IC-UNICAMP, Campinas, SP, Brasil, 2010.

[164] Júlio Cesar Reis, Rodrigo Bonacin, e Maria Cecília Calani Baranauskas, *Prospecting an inclusive search mechanism for social network services*, in Enterprise Information Systems, Joaquim Filipe, José Cordeiro, Wil Aalst, John Mylopoulos, Michael Rosemann, Michael J. Shaw, e Clemens Szyperski, eds., vol. 73 of Lecture Notes in Business Information Processing, Springer Berlin Heidelberg, 2011, pp. 555–570.

[165] Luis Carlos Ribeiro, *OntoLP: Construção semi-automática de ontologias a partir de textos da língua portuguesa*, master's thesis, Mestrado em Computação Aplicada/UNISINOS, 2008.

[166] Stephen E. Robertson, *Understanding inverse document frequency: on theoretical arguments for idf*, Journal of Documentation, 60 (2004), pp. 503–520.

[167] Stephen E. Robertson e Karen Spärck-Jones, *Relevance weighting of search terms*, Journal of American Society for Information Science, 27 (1976), pp. 129–146.

[168] Stephen E. Robertson e S. Walker, *On relevance weights with little relevance information*, SIGIR Forum, 31 (1997), pp. 16–24.

[169] Gabriella Rose, Melissa Holland, Steve Larocca, e Robert Winkler, *Semi-automated methods for refining a domain-specific terminology base*, Tech. Report ARL-RP-0311, U. S. Army Research Laboratory, Adelphi, MD, USA, 2011.

[170] Gerard Salton e Christopher Buckley, *Term-weighting approaches in automatic text retrieval*, Information Processing & Management, 24 (1988), pp. 513–523.

[171] A. Sanchez e P. Cantos, *CUMBRE – Corpus Linguistico del Español Contemporaneo – Fundamentos, Metodologia, y Aplicaciones*, SEGL, Madri, Spain, 1996.

[172] Mark Sanderson e W. Bruce Croft, *Deriving concept hierarchies from text*, in Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '99, New York, USA, 1999, ACM Press, pp. 206–213.

[173] R. C. Schank, *Conceptual dependency: A theory of natural language understanding*, Cognitive Psychology, 3 (1972), pp. 532–631.

[174] Mike Scott, *What can corpus software do?*, in Routledge Handbook of Corpus Linguistics, A. O'Keeffe e M. J. McCarthy, eds., Lecture Notes in Computer Science, Routledge, 2010, pp. 136–151.

[175] Ivo Serra e Rosario Girardi, *A process for extracting non-taxonomic relationships of ontologies from text*, Intelligent Information Management, 3 (2011), pp. 119–124.

[176] Claude Shannon, *Communication theory of secrecy systems*, Bell Systems Technical Journal, 28 (1949), pp. 656–715.

[177] João Silva, António Branco, Sérgio Castro, e Ruben Reis, *Out-of-the-box robust parsing of portuguese*, in PROPOR 2010 – International Conference on Computational Processing of Portuguese Language, 2010, pp. 75–85.

[178] Dimitrios Skoutas e Mohammad Alrifai, *Tag clouds revisited*, in Proceedings of the 20th ACM international conference on Information and knowledge management, CIKM '11, New York, NY, USA, 2011, ACM, pp. 221–230.

[179] Daniel Dominic Sleator e David Temperley, *Parsing english with a link grammar*, CoRR, abs/cmp-lg/9508004 (1995).

[180] Frank Smadja, *Retrieving collocations from text: Xtract*, Computational Linguistics, 19 (1993), pp. 143–177.

[181] E. C. Souza, A. O. Martins, e P. C. M. A. Branco, *Glossário de rochas graníticas*, DNPM-CPRM-DOCEGEO, Rio de Janeiro, Brazil, 1987.

[182] J. Sowa, *Building, sharing and merging ontologies*. `http://users.bestweb.net/~sowa/ontology/ontoshar.htm`, 1999. último acesso em 11/05/2011.

[183] Karen Spärck-Jones, *A statistical interpretation of term specificity and its application in retrieval*, Journal of Documentation, 28 (1972), pp. 11–21.

[184] K. Suguio, *Dicionário de geologia marinha*, T. A. Queiroz, São Paulo, Brazil, 1992.

[185] Jiao Tao, Evren Sirin, Jie Bao, e Deborah L. McGuinness, *Integrity constraints in owl*, in AAAI, 2010.

[186] Luís Teixeira, Gabriel Lopes, e Rita Ribeiro, *Automatic extraction of document topics*, in Technological Innovation for Sustainability, Luis Camarinha-Matos, ed., vol. 349 of IFIP Advances in Information and Communication Technology, Springer Boston, 2011, pp. 101–108. 10.1007/978-3-642-19170-1_11.

[187] *Textcc – textos técnicos e científicos*. `http://www.ufrgs.br/textecc`, April 2011. (último acesso em 13 dezembro 2011).

[188] J. Thomas, D. Milward, C. Ouzounis, S. Pulman, e M. Carroll, *Automatic extraction of protein interactions from scientific abstracts*, in Pacific Symposium on Biocomputing, vol. 5, 2000, pp. 538–549.

[189] Ivan Titov e Mikhail Kozhevnikov, *Bootstrapping semantic analyzers from non-contradictory texts*, in Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL '10, Morristown, USA, 2010, Association for Computational Linguistics, pp. 958–967.

[190] *True knowledge - the internet search engine*. `http://www.trueknowledge.com`, Dezembro 2011. (último acesso em 13 dezembro 2011).

[191] *Unitex versão 1.2*. `http://igm.univ-mlv.fr/~unitex/`, Dezembro 2011. (último acesso em 13 dezembro 2011).

[192] C. J. van Rijsbergen, *Information Retrieval*, Butterworths, London, UK, 1975.

[193] Paola Velardi, Roberto Navigli, Alessandro Cucchiarelli, e Francesca Neri, *Evaluation of OntoLearn, a methodology for automatic population of domain ontologies*, in Ontology Learning from Text: Methods, Applications and Evaluation, Paul Buitelaar, Philipp Cimiano, e Bernardo Magnini, eds., IOS Press, 2006.

[194] R. Vieira, E. Bick, J. Coelho, V. Muller, S. Collovini, J. Souza, e L. Rino, *Semantic tagging for resolution of indirect anaphora*, in Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue, SigDIAL '06, Stroudsburg, PA, USA, 2006, Association for Computational Linguistics, pp. 76–79.

[195] Renata Vieira e Massimo Poesio, *An empirically based system for processing definite descriptions*, Comput. Linguist., 26 (2000), pp. 539–593.

[196] *Floresta sintáctica - visl - visual interactive syntax learning.* `http://beta.visl.sdu.dk/visl/pt/parsing/automatic/trees.php`, May 2011. (último acesso em 13 dezembro 2011).

[197] David L. Waltz, *An english language question answering system for a large relational database*, Communication of the ACM, 21 (1978), pp. 526–539.

[198] Shi Wang, Yanan Cao, Xinyu Cao, e Cungen Cao, *Learning concepts from text based on the inner-constructive model*, in Knowledge Science, Engineering and Management, Zili Zhang e Jörg Siekmann, eds., vol. 4798 of Lecture Notes in Computer Science, Springer, Berlin/Heidelberg, Germany, 2007, pp. 255–266.

[199] W. Weaver, *Translation (1949)*, in Machine Translation of Languages, William N. Locke e Andrew D. Booth, eds., MIT Press, Cambridge, USA, 1955.

[200] Joseph Weizenbaum, *Eliza – a computer program for the study of natural language communication between man and machine*, Communications of the ACM, 9 (1966), pp. 36–45.

[201] Igor S. Wendt, Lucelene Lopes, Renata Vieira, Daniel Martins, e Vera Lúcia Strube de Lima, *Geração automática de glossários de termos específicos de um corpus de geologia*, in 3o Seminário de pesquisa em ontologia no Brasil (ONTOBRAS), UFSC, 2010, pp. 1–10.

[202] Joachim Wermter e Udo Hahn, *Paradigmatic modifiability statistics for the extraction of complex multi-word terms*, in Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT '05, Stroudsburg, PA, USA, 2005, Association for Computational Linguistics, pp. 843–850.

[203] Joachim Wermter e Udo Hahn, *You can't beat frequency (unless you use linguistic knowledge): a qualitative evaluation of association measures for collocation and term extraction*, in Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, ACL-44, Stroudsburg, USA, 2006, Association for Computational Linguistics, pp. 785–792.

[204] Rodrigo Wilkens e Aline Villavicencio, *Question answering for portuguese: How much is needed?*, in Advances in Artificial Intelligence, SBIA 2010, António da Rocha Costa, Rosa Vicari, e Flavio Tonidandel, eds., vol. 6404 of Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2011, pp. 173–182.

[205] Y. Wilks, *Preference semantics*, in Formal Semantics of Natural Language, E. L. Keenan, ed., Cambridge University Press, New York, USA, 1975, pp. 329–348.

[206] T. Winograd, *Procedures as a Representation for Data in a Computer program for Understanding Natural Language*, dissertation, MIT, Cambridge, USA, 1971.

[207] Ian H. Witten e Eibe Frank, *Data mining: practical machine learning tools and techniques*, Morgan Kaufmann, 2 ed., 2005.

[208] Ian H. Witten, Alistair Moffat, e Timothy C. Bell, *Managing Gigabytes: Compressing and Indexing Documents and Images*, Morgan Kaufmann, San Francisco, 1999.

[209] William A. Woods, *Transition network grammars for natural language analysis*, Communications of the ACM, 13 (1970), pp. 591–606.

[210] William A. Woods, *Lunar rocks in natural English: Explorations in natural language question answering*, in Linguistic Structures Processing, Antonio Zampolli, ed., North Holland, Amsterdam, The Nedderlands, 1977, pp. 521–569.

[211] *Wordsmith tools version 6*. `http://www.lexically.net/wordsmith/`, Dezembro 2011. (último acesso em 13 dezembro 2011).

[212] Ho Chung Wu, Robert Wing Pong Luk, Kam Fai Wong, e Kui Lam Kwok, *Interpreting tf-idf term weights as making relevance decisions*, ACM Transaction on Information Systems, 26 (2008), pp. 13:1–13:37.

[213] Zhang Xiaojun, *Michael w. berry and jacob kogan (eds.): Text mining: applications and theory*, Information Retrieval, 14 (2011), pp. 208–211. 10.1007/s10791-010-9153-5.

[214] Jinxi Xu e W. Bruce Croft, *Corpus-based stemming using cooccurrence of word variants*, ACM Transaction on Information Systems, 16 (1998), pp. 61–81.

[215] Nisha Yadav, Hrishikesh Joglekar, Rajesh P. N. Rao, Mayank N. Vahia, Ronojoy Adhikari, e Iravatham Mahadevan, *Statistical analysis of the indus script using ¡italic¿n¡/italic¿-grams*, PLoS ONE, 5 (2010), p. e9506.

[216] Hui Yang e Jamie Callan, *Ontology generation for large email collections*, in Proceedings of the 2008 international conference on Digital government research, dg.o '08, Digital Government Society of North America, 2008, pp. 254–261.

[217] Elias Zavitsanos, Georgios Paliouras, George A. Vouros, e Sergios Petridis, *Discovering subsumption hierarchies of ontology concepts from text corpora*, in Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, WI '07, Washington, USA, 2007, IEEE Computer Society, pp. 402–408.

[218] George K. Zipf, *The Psycho-Biology of Language - An Introduction to Dynamic Philology*, Houghton-Mifflin Company, Boston, USA, 1935.

A. Listas de Referência – *corpus* Pediatria

As listas de termos de referência para o *corpus* de Pediatria citadas na Seção 3.1 foram construídas pelo grupo TEXTECC da Universidade Federal do Rio Grande do Sul (`www.ufrgs.br/textecc`). Essas listas são fruto de um laborioso processo manual de extração de termos visando a elaboração de um glossário para apoio a estudantes de tradução. Esse processo foi feito por de estudantes de linguística, com o apoio de especialistas do domínio (Pediatria).

Como resultado, essas listas de referência possuem os seguintes 1.534 bigramas e 2.660 trigramas considerados termos relevantes para o *corpus* de Pediatria.

Bigramas da lista de referência em ordem alfabética — Tabela 1 de 6

abertura traqueal	alcalose metabólica	amamentação predominante	asma aguda
abordagem diagnóstica	alças intestinais	ambiente escolar	asma brônquica
abordagem terapêutica	álcool fetal	ambiente familiar	asma moderada
abscesso mamário	aleitamento artificial	ambiente hospitalar	asma persistente
absenteísmo escolar	aleitamento continuado	amostra estudada	asma referida
abuso sexual	aleitamento exclusivo	amostras fecais	aspectos clínicos
abusos físicos	aleitamento materno	amostras coletadas	aspectos éticos
acesso venoso	aleitamento misto	amostras genotipadas	aspectos genéticos
achados clínicos	aleitamento natural	amostras independentes	aspectos nutricionais
achados endoscópicos	aleitamento predominante	amplo espectro	assistência intensiva
achados histológicos	alérgenos testados	análise microscópica	assistência neonatal
achados radiográficos	alergia alimentar	análise morfométrica	assistência ventilatória
acidentes vasculares	alergia respiratória	análise multivariada	atenção especial
ácido acetilsalicílico	alimentação adequada	análise univariada	atenção primária
ácido fitânico	alimentação artificial	anamnese dirigida	atendimento ambulatorial
ácido fólico	alimentação complementar	anemia falciforme	atendimento médico
ácido valpróico	alimentação enteral	anemia hemolítica	atendimento pediátrico
ácidos graxos	alimentação infantil	anestesia geral	ativação imunológica
acidose metabólica	alimentos complementares	anomalias congênitas	ativação macrofágica
acometimento hepático	alimentos consumidos	anomalias cromossômicas	atividade esportiva
acompanhamento ambulatorial	alimentos saudáveis	anorexia infantil	atividade física
acompanhamento médico	alojamento conjunto	anorexia nervosa	atividade muscular
aconselhamento genético	alta hospitalar	ansiedade generalizada	atividade sexual
aconselhamento nutricional	alta morbidade	antecedentes familiares	ato cirúrgico
acth sintético	alta sensibilidade	antecedente mórbido	atresia biliar
acuidade visual	alterações auditivas	anti histamínico	átrio direito
adaptação cultural	alterações bioquímicas	anticorpos específicos	átrio esquerdo
adeno hipófise	alterações cardiovasculares	antidepressivos tricíclicos	atrofia vilositária
admissão hospitalar	alterações clínicas	antígeno polissacarídico	atuação exclusiva
adolescentes obesos	alterações fonoarticulatórias	antipsicóticos atípicos	audição normal
adolescentes pesquisados	alterações hemodinâmicas	aparelho gastrintestinal	autismo infantil
adulto autista	alterações hepáticas	aparelho locomotor	auto anticorpos
adultos jovens	alterações histológicas	aparelho respiratório	auto estima
aerossóis dosimetrados	alterações imunológicas	apresentação clínica	autoridades sanitárias
agente etiológico	alterações leves	ar ambiente	avaliação antropométrica
agente infeccioso	alterações metabólicas	arcadas dentárias	avaliação clínica
agentes antimicrobianos	alterações morfológicas	arritmias cardíacas	avaliação neurológica
agentes paralisantes	alterações neurológicas	artéria pulmonar	avaliação nutricional
agentes teratogênicos	alterações radiológicas	artrite reumatóide	avaliação oftalmológica
agitação psicomotora	altas doses	artrite séptica	baixa escolaridade
agressão física	alto risco	artrites idiopáticas	baixa estatura
albumina sérica	amamentação exclusiva	asfixia perinatal	baixa idade

Bigramas da lista de referência em ordem alfabética — Tabela 2 de 6

baixa renda	ciclos menstruais	cortisol sérico	descongestionantes tópicos
baixas doses	ciclo respiratório	coto ureteral	desenvolvimento cerebral
baixo custo	cintilografia óssea	cpap nasal	desenvolvimento cognitivo
baixo débito	cintilografia renal	crânio neonatal	desenvolvimento físico
baixo nível	circulação extracorpórea	craniotomia descompressiva	desenvolvimento infantil
baixo peso	circulação pulmonar	crescimento bacteriano	desenvolvimento motor
baixo risco	cirrose estabelecida	crescimento fetal	desenvolvimento neurológico
balanço energético	cirurgia cardíaca	crescimento linear	desenvolvimento neuropsicomotor
barreira hematoencefálica	cirurgia conservadora	crescimento somático	desenvolvimento normal
base populacional	classe social	crianças assintomáticas	desnutrição aguda
bases clínicas	clínica pediátrica	crianças doentes	desnutrição grave
bcg id	coagulação intravascular	criança internada	desnutrição leve
bcg pc	cobertura vacinal	criança obesa	despertar noturno
bebê prematuro	colangiografia transoperatória	criança ostomizada	diabetes melito
bebês chiadores	colestase crônica	crianças saudáveis	diagnóstico clínico
bebidas alcoólicas	colesterol total	crianças acometidas	diagnóstico definitivo
bexiga neurogênica	colite alérgica	crianças afetadas	diagnósticos estabelecidos
bicos artificiais	colostro materno	crianças amamentadas	diagnóstico etiológico
bilirrubinas totais	colunas líquidas	crianças autistas	diagnóstico final
bilirrubinemia total	coluna lombar	crianças brasileiras	diagnóstico preciso
biologia molecular	coluna vertebral	crianças constipadas	diagnóstico precoce
biópsia hepática	complacência pulmonar	crianças depressivas	diagnóstico prévio
bloqueador neuromuscular	complexo esfincteriano	crianças estudadas	diagnósticos diferenciais
bloqueio neuromuscular	complicações associadas	crianças febris	diagnósticos incorretos
boa evolução	complicações relacionadas	crianças infectadas	diarréia aguda
boa resposta	complicações supurativas	crianças maiores	diarréia persistente
borda esternal	comportamento alimentar	crianças nascidas	diátese hemorrágica
borda hepática	comportamento humano	crianças normais	dieta enteral
borda inferior	comportamentos sociais	crianças pequenas	dieta normal
bronquiolite aguda	comportamentos automutilantes	crianças prematuras	dificuldade diagnóstica
bronquiolite viral	comportamentos repetitivos	crianças soropositivas	dificuldade respiratória
bulimia nervosa	composição corporal	crianças vestibulopatas	dimorfismo sexual
caixa torácica	compressas frias	crise aguda	disfunção miccional
cálcio total	compressas mornas	crise asmática	disfunção miocárdica
cálculo amostral	comprometimento hepático	crises convulsivas	disfunção ventricular
camada basal	comunicação interventricular	crise dolorosa	disfunção cerebral
câmaras cardíacas	comunidade pediátrica	crises epilépticas	disfunções orais
campos pulmonares	comunidades pobres	crises hipertensivas	disfunção orgânica
canais arteriais	concentrações inibitórias	critérios clínicos	displasia broncopulmonar
cânula traqueal	condição socioeconômica	critérios diagnósticos	displasias ósseas
capacidade física	condições clínicas	cromossomo x	dispositivos inalatórios
capacidade funcional	condições socioeconômicas	curso clínico	distensão abdominal
capacidade residual	condrodisplasia puntiforme	dac prematura	distensão vesical
caracteres sexuais	condutas terapêuticas	dados epidemiológicos	distribuição universal
características clínicas	congestão pulmonar	dados clínicos	distúrbios motores
características demográficas	consentimento informado	dano oxidativo	distúrbio ventilatório
características físicas	consentimento livre	dano pulmonar	distúrbios metabólicos
características maternas	conseqüências clínicas	dano renal	distúrbios respiratórios
cardiomiopatia dilatada	constipação crônica	dano tecidual	diversos alérgenos
cardiomiopatia hipertrófica	constipação intestinal	débito cardíaco	diversos órgãos
cardiopatia congênita	consultas agendadas	débito urinário	divertículos uretrais
cardiopatias adquiridas	consultas médicas	dech aguda	dna bacteriano
carga eletrostática	consulta pediátrica	dech crônica	dobras cutâneas
carga viral	contato visual	decúbito dorsal	doença aguda
cárie dentária	conteúdo energético	defeitos congênitos	doença aterosclerótica
cascata inflamatória	conteúdo mineral	deficiência auditiva	doença avançada
casos graves	contra indicações	deficiência mental	doença bacterêmica
casos suspeitos	contraceptivos orais	déficit auditivo	doença bacteriana
categoria imunológica	contratilidade miocárdica	déficit neurológico	doença cardíaca
cateterização uretral	controle esfincteriano	déficits nutricionais	doença cardiovascular
causalidade reversa	coordenação visomotora	déficits cognitivos	doença celíaca
causa orgânica	cor branca	demanda metabólica	doença coronariana
causas respiratórias	coração esquerdo	densidade energética	doença crônica
cavidade amniótica	cordão triangular	densidade mineral	doença diarréica
cavidade oral	cordão umbilical	densidade óssea	doença hereditária
células musculares	cordas vocais	densitometria óssea	doença inflamatória
células alveolares	corpo estranho	deposição pulmonar	doença localizada
células endoteliais	corpo estriado	depressão anaclítica	doença meningocócica
células epiteliais	corpo humano	depressão infantil	doença metabólica
células progenitoras	correção cirúrgica	depressão maior	doença metastática
células t	corte transversal	depressão mascarada	doença oncológica
centros especializados	córtex cerebral	depressão miocárdica	doença orgânica
cepa utilizada	corticóides inalados	depressão respiratória	doença péptica
cepas isoladas	corticosteróide inalatório	derivação urinária	doença pneumocócica
cepas resistentes	corticóide oral	derivados imidazolínicos	doença pulmonar
cesárea eletiva	corticóide sistêmico	derivados nitroimidazólicos	doença rara
choque séptico	corticóides inalatórios	dermatite atópica	doença renal
choro inconsolável	corticosteróide antenatal	desconforto alto	doença renovascular
cicatriz renal	corticoterapia inalatória	desconforto físico	doença respiratória
ciclo circadiano	cortisol basal	desconforto respiratório	doença vascular

Bigramas da lista de referência em ordem alfabética — Tabela 3 de 6

doença viral	epilepsias generalizadas	faringe posterior	habilidades corporais
doenças alérgicas	episódio agudo	faringite aguda	habilidades sociais
doenças arteriais	episódios bulímicos	fase aguda	hábito intestinal
doenças atópicas	equipe assistencial	fator estressor	hábitos alimentares
doenças falciformes	equipe médica	fator limitante	hábitos orais
doenças graves	equipe multidisciplinar	fator predisponente	hábitos saudáveis
doenças hepáticas	esclerose tuberosa	fator protetor	haste hipofisária
doenças infecciosas	escolaridade materna	fatores ambientais	helmintíases intestinais
doenças invasivas	escores clínicos	fatores biológicos	hemoculturas positivas
doenças mentais	esforço respiratório	fatores culturais	hemoglobina s
doenças mitocondriais	esofagite eosinofílica	fatores genéticos	hemorragia digestiva
doenças neurológicas	esofagite erosiva	fator importante	hemorragia intracraniana
doenças neuromusculares	espaçador artesanal	fatores prognósticos	hemorragia intraventricular
doenças reumáticas	espaço extracelular	fatores relacionados	hepatite aguda
dor abdominal	espaço intersticial	fatores socioculturais	hepatite b
dor intensa	espaço intracelular	fatores socioeconômicos	hepatite c
dor noturna	espaço mandibular	febre alta	hepatopatia crônica
dor torácica	espécie humana	fêmur proximal	herança autossômica
dores difusas	espectro autista	fibra alimentar	herança genética
dores recorrentes	espinha bífida	fibra insolúvel	herpes simples
dose alta	espinhas dendríticas	fibra óptica	hfa dpb
doses diárias	esquema vacinal	fibra solúvel	hidrocefalias congênitas
dose inicial	esquema antimicrobiano	fibrose cística	hidrocefalias isoladas
dose terapêutica	esquemas terapêuticos	fibrose pulmonar	hiper responsividade
dose única	estado basal	fisioterapia respiratória	hipercapnia permissiva
dose utilizada	estado geral	fissuras mamilares	hiperemia conjuntival
doses baixas	estado infeccioso	flora intestinal	hiperfluxo pulmonar
doses maiores	estado nutricional	fluticasona hfa	hipermobilidade articular
doses menores	estenose aórtica	fluxo biliar	hipersensibilidade imediata
doses recomendadas	estenose pilórica	fluxo expiratório	hipersensibilidade tardia
dose tóxica	estenose pulmonar	fluxo salivar	hipertensão arterial
drenagem liquórica	estenose subglótica	fluxo sangüíneo	hipertensão endocraniana
drenagem torácica	esteróides inalados	força muscular	hipertensão essencial
drogas ototóxicas	estimulação imunológica	forças mecânicas	hipertensão intracraniana
drogas sedativas	estímulo fóbico	formação óssea	hipertensão materna
drogas usadas	estímulos dolorosos	fórmula láctea	hipertensão pulmonar
drogas utilizadas	estratégia terapêutica	fórmulas infantis	hipertensão secundária
drogas vasoativas	estudo cromossômico	fosfatase alcalina	hipertensão sistólica
ducto biliar	estudo genético	fraturas vertebrais	hipertrofia muscular
ducto hepático	estudos clínicos	freqüência cardíaca	hipertrofia ventricular
ductos lactíferos	estudos epidemiológicos	freqüência respiratória	hipoacusia condutiva
dupla mãe	esvaziamento gástrico	função esplênica	hipoplasia cerebelar
duplo placebo	etiologia bacteriana	função hepática	hipoplasia hipofisária
ecocardiograma transtorácico	etiologia viral	função pulmonar	hipotálamo hipofisária
eczema atópico	evento fisiopatológico	função renal	hipotensão arterial
edema cerebral	evento traumático	função respiratória	hipotermia moderada
edema pulmonar	eventos adversos	função surfactante	hipóxia tecidual
efeito adverso	eventos cardiovasculares	funcionamento oral	histologia normal
efeito analgésico	eventos clínicos	fundoplicatura anterior	história alimentar
efeito cumulativo	eventos paroxísticos	ganho ponderal	história clínica
efeito inotrópico	evidências científicas	gasometrias arteriais	história familiar
efeito protetor	evidências clínicas	gasto calórico	hormônios sexuais
efeito sedativo	evidências epidemiológicas	gasto energético	idade concepcional
efeito terapêutico	evolução clínica	gastrite crônica	idade corrigida
efeitos benéficos	evolução neurológica	gêmeos monozigóticos	idade cronológica
efeitos clínicos	exame citológico	germes estudados	idade escolar
efeitos colaterais	exame clínico	gestantes estudadas	idade gestacional
efeitos extrapiramidais	exame cultural	giros temporais	idade maternas
efeitos significativos	exame endoscópico	gordura corporal	idade óssea
eliminação renal	exame físico	gordura corpórea	idade pediátrica
embasamento científico	exame histológico	gordura saturada	ige específicos
emergência pediátrica	exame neurológico	gordura subcutânea	ige sérica
emissões otoacústicas	exame oftalmológico	gordura visceral	ige total
endocardite infecciosa	exame radiológico	grandes artérias	imagem corporal
endoscopia digestiva	exames complementares	grupo ambulatório	imaturidade pulmonar
endoscopia normal	exames laboratoriais	grupo controle	imc elevado
endoscopia respiratória	exames subsidiários	grupo etário	imc igual
endoscopia terapêutica	excreção renal	grupo neb	imc maior
endotélio vascular	exercício físico	grupos oligoarticular	importância clínica
ensaio imunoenzimático	exercícios orofaciais	grupos pediátricos	imprinting metabólico
ensaios clínicos	expansão intravascular	grupo poliarticular	imunidade celular
enterococcus faecalis	expressões faciais	grupo precoce	imunidade humoral
enterocolite necrosante	extubação acidental	grupo prednisona	imunofluorescência indireta
enurese noturna	fácies típica	grupo sobrevivente	inalador dosimetrado
enurese polissintomática	faixa etária	grupo supino	incontinência urinária
envolvimento hepático	faixas pediátricas	grupo total	indicação cirúrgica
envolvimento pulmonar	falência cardíaca	grupo tratado	indicadores antropométricos
enzimas hepáticas	falência respiratória	grupos estudados	índice cardíaco
enzimas pancreáticas	falha terapêutica	grupos experimentais	índice cardiotorácico
epilepsia mioclônica	falsa anorexia	habilidades cognitivas	índice ponderal

Bigramas da lista de referência em ordem alfabética — Tabela 4 de 6

indivíduos autistas	lesão térmica	miocardiopatia hipertrófica	paciente descrito
indústrias farmacêuticas	lesões cardíacas	modalidade terapêutica	pacientes estudados
infecção aguda	lesões cutâneas	modelo animal	pacientes fibrocísticos
infecção crônica	lesões glomerulares	modelos experimentais	pacientes graves
infecção hospitalar	lesões líticas	monitorização prolongada	pacientes hepatopatas
infecção materna	leucemia mielóide	moraxella catarrhalis	pacientes hospitalizados
infecções secundárias	leucemia aguda	morbidade respiratória	pacientes incluídos
infecção urinária	leucomalácia periventricular	morbimortalidade infantil	pacientes infectados
infecções bacterianas	leucometria inicial	mordida aberta	pacientes intubados
infecções congênitas	limiar convulsivo	mortalidade geral	pacientes osteopênicos
infecções graves	linfócitos citotóxicos	mortalidade infantil	pacientes pediátricos
infecções pneumocócicas	linguagem escrita	mortalidade neonatal	pacientes relatados
infecções pulmonares	linguagem oral	morte digna	padrão alimentar
infecções recorrentes	lipodistrofia generalizada	morte súbita	padrões motores
infecções respiratórias	líquido amniótico	movimentos anormais	pais obesos
infecções virais	líquido pleural	movimentos irregulares	palato duro
inflexão inferior	literatura médica	mucosa esofágica	palato mole
influências genéticas	livre demanda	mucosa gástrica	pálpebra superior
informação materna	lobo frontal	mucosa intestinal	parada cardiorrespiratória
informação visual	lobo temporal	mucosa nasal	paralisia cerebral
infra estrutura	longa duração	mulher mãe	parâmetros fisiológicos
infusão contínua	longo prazo	múltiplos órgãos	parâmetros clínicos
ingestão alimentar	má absorção	musculatura respiratória	parâmetros lineares
ingestão calórica	má nutrição	músculos respiratórios	parâmetros ventilatórios
ingestão energética	má oclusão	nascimentos prematuros	parasitose intestinal
ingurgitamento mamário	mães adolescentes	necessidades calóricas	parede abdominal
injúrias físicas	mães adultas	necessidades nutricionais	parede torácica
inspiração profunda	maior prevalência	necrose parietal	parênquima pulmonar
instrução materna	maior risco	necrose tumoral	parênquima renal
insuficiência cardíaca	malformações cerebrais	neuro hipófise	parto cesáreo
insuficiência hepática	malformações congênitas	neuropeptídeo y	parto cesariano
insuficiência pancreática	manifestações clínicas	níveis maturacionais	parto normal
insuficiência renal	manifestações cutâneas	níveis plasmáticos	parto vaginal
insuficiência respiratória	manifestações iniciais	níveis pressóricos	patologias psiquiátricas
insulina regular	manifestações sistêmicas	nível hidroaéreo	pediatra geral
interação medicamentosa	máscara facial	nível sérico	pediatra brasileiro
interação social	máscara laríngea	nível socioeconômico	peptídeo c
intercorrências clínicas	massa corporal	novas cicatrizes	pequenos pacientes
intercorrências respiratórias	massa corpórea	novas terapias	pequenos volumes
internações hospitalares	massa gorda	novos medicamentos	perda auditiva
internação prolongada	massa muscular	nutrição infantil	perda celular
intervenção cirúrgica	massa óssea	nutrição parenteral	perda óssea
intervenções terapêuticas	maturação sexual	obesidade infantil	perda urinária
intervenções tradicionais	mecânica pulmonar	obstrução biliar	perfil lipídico
intestino delgado	mecânica respiratória	obstrução infravesical	perfusão cerebral
intubação endotraqueal	mecanismo fisiopatológico	obstrução nasal	perfusão tecidual
intubação traqueal	mecanismo imunológico	obstrutiva crônica	perímetro braquial
investigação clínica	mediador inflamatório	oncologia pediátrica	perímetro cefálico
investigações laboratoriais	medicação sedativa	ordenha mamária	período crítico
irradiância espectral	medicações utilizadas	orelha direita	período estudado
isolamento social	medicina social	orelha média	período maior
isquemia cerebral	médicos assistentes	orelhas proeminentes	período neonatal
lábio inferior	médicos entrevistados	orientação alimentar	período perinatal
lábio superior	médicos residentes	orientação antecipatória	período pubertário
lactentes sibilantes	medida analgésica	orientação familiar	peroxidação lipídica
lâmina reta	medidas antropométricas	orientação nutricional	peso adequado
laringite viral	medidas preventivas	osmolalidade intracelular	peso corporal
laringoscopia direta	medula espinhal	osmolaridade plasmática	peso fecal
lavado broncoalveolar	medula óssea	osmolaridade sérica	peso normal
lavado nasal	meio ambiente	osso trabecular	pesquisa nacional
lavagem pulmonar	meia vida	ossos grandes	pessoas obesas
leishmaniose visceral	melhor oxigenação	ossos longos	pior prognóstico
leite bovino	melhor prognóstico	ossos pequenos	plasma materno
leite fraco	membrana celular	otite média	pneumologista pediatra
leite humano	membrana hialina	otoemissão acústica	pneumonia bacteriana
leite maduro	membrana timpânica	ouvido médio	pneumonia grave
leite materno	membros inferiores	óxido nítrico	pneumonia pneumocócica
leite ordenhado	membros superiores	oxigênio suplementar	pneumonias comunitárias
leites artificiais	menor escolaridade	pacientes enuréticos	pneumopatias crônicas
leites industrializados	meta análise	pacientes internados	poeira doméstica
leites modificados	metabolismo cerebral	paciente oncológico	pólo inferior
leitos intensivos	metabolismo lipídico	pacientes adolescentes	pólo superior
leitos neonatais	metabolismo ósseo	pacientes adultos	ponte nasal
lesão cerebral	metileno tetraidrofolato	pacientes alérgicos	pontos dolorosos
lesão glomerular	método diagnóstico	pacientes avaliados	população adulta
lesão grave	método invasivo	pacientes cirúrgicos	população alvo
lesão isquêmica	método sorológico	pacientes clínicos	população estudada
lesões moderadas	mielinólise pontina	pacientes colestáticos	população geral
lesão pulmonar	mímica facial	pacientes conscientes	população infantil
lesão secundária	mineralização óssea	pacientes críticos	população pediátrica

Bigramas da lista de referência em ordem alfabética — Tabela 5 de 6

pós carga	punção lombar	risco intermediário	surfactantes naturais
pós operatório	quadro clínico	riscos nutricionais	surfatante exógeno
pós parto	quadros diarréicos	risco relativo	suspeita diagnóstica
pós termo	quadro grave	ritmo cardíaco	tabagismo materno
pós tmo	quadro psicótico	ritmo circadiano	tamanho amostral
posição ereta	quadro respiratório	rm grave	tce grave
posição ortostática	quadro delirante	saco coletor	tecido adiposo
posição supina	quadros infecciosos	salário mínimo	tecidos adjacentes
postura anormal	quadro neurológico	sam associada	tecido conjuntivo
práticas alimentares	queixas principais	sangramento digestivo	tecido hipofisário
prática clínica	queixas somáticas	sangue materno	tecido ósseo
prática desportiva	quimioprofilaxia antibiótica	sangue periférico	temperatura axilar
prática diária	radicais livres	saúde infantil	temperatura corporal
prática médica	radiologista pediátrico	saúde mental	tempo inspiratório
prática pediátrica	raio x	saúde pública	tensão superficial
pré carga	ramos pulmonares	screening neonatal	terapêutica inicial
pré escolar	reabilitação vestibular	secreção nasal	terapia antimicrobiana
pré natal	reabsorção óssea	secreção nasofaríngea	terapia intensiva
pré operatório	reações alérgicas	secreções respiratórias	terceira dose
pré oxigenação	reação anafilática	segmentos renais	terrores noturnos
pré termo	reações adversas	seguimento ambulatorial	teste laboratorial
prednisona oral	reações graves	seguimento clínico	teste tuberculínico
pregas cutâneas	reações inflamatórias	seio materno	testes cutâneos
pregas vocais	recém nascidos	seios paranasais	teste diagnóstico
pressão arterial	recém natos	sepse grave	testes neuropsicológicos
pressão capilar	recrutamento alveolar	sepse neonatal	testes sorológicos
pressão coloidosmótica	recrutamento pulmonar	septo atrioventricular	tipo tensional
pressão diastólica	recuperação nutricional	seres humanos	tiques motores
pressão expiratória	recursos terapêuticos	setor público	tmo alogênico
pressão inspiratória	recusa alimentar	sibilância prévia	tmo autogênico
pressão intracraniana	rede pública	sinais clínicos	tomografia computadorizada
pressão intraventricular	reflexos orais	sinais vitais	tórax inicial
pressão positiva	refluxo gastresofágico	síndrome caracterizada	trabalho materno
pressões pulmonares	refluxo gastroesofageano	síndrome torácica	trabalho respiratório
pressão sistêmica	refluxo gastroesofágico	síndromes dismórficas	transmissão perinatal
pressão sistólica	refluxo vesicoureteral	síndromes epilépticas	transplante alogênico
pressão venosa	região cervical	síndromes genéticas	transplante autogênico
primeira avaliação	região frontal	síndromes neurocutâneas	transplante cardíaco
primeira consulta	região lombar	sintomas alvo	transplante hepático
primeira fase	região metropolitana	sintomas comportamentais	transtorno afetivo
primeira infância	região subglótica	sintomas diurnos	transtornos ansiosos
primeira internação	relações familiares	sintomas iniciais	transtorno bipolar
primeira semana	relacionamentos sociais	sintomas negativos	transtorno depressivo
primeiras mamadas	relações afetivas	sintomas presentes	transtornos alimentares
primeiro ano	relações sexuais	sintomas urinários	transtornos psiquiátricos
primeiro exame	relaxamento muscular	sintomatologia clínica	tratamento adequado
primeiro mês	relaxantes musculares	sinusite aguda	tratamento antimicrobiano
primeiro passo	remissão clínica	sinusites bacterianas	tratamento apropriado
primeiros dias	remissão completa	sistema cardiovascular	tratamento cirúrgico
primoinfecção urinária	remodelação óssea	sistema dopaminérgico	tratamento clínico
princípio ativo	renda familiar	sistema imunológico	tratamento continuado
problema clínico	rendimento escolar	sistema límbico	tratamentos convencional
problemas comportamentais	resfriados comuns	sistema nervoso	tratamento empírico
problemas emocionais	resistência antimicrobiana	sistema respiratório	tratamento endoscópico
problemas metodológicos	resistência bacteriana	sistema surfactante	tratamento específico
problema neurológico	resistência intermediária	situação clínica	tratamento farmacológico
problemas psiquiátricos	resistência vascular	situação conjugal	tratamento inicial
problemas respiratórios	respiração espontânea	sobrepeso masculino	tratamento intensivo
procedimentos cirúrgicos	respiração nasal	sobrevida global	tratamento medicamentoso
procedimentos diagnósticos	respiração oral	sódio sérico	tratamento paliativo
procedimentos dolorosos	resposta clínica	soluções hipertônicas	tratamento profilático
procedimentos invasivos	resposta imune	solução salina	tratamento tradicional
procedimentos médicos	resposta imunológica	sonda endotraqueal	tratamento proposto
processamento auditivo	resposta inflamatória	sonda nasogástrica	trato digestório
processo anabólico	resposta terapêutica	sopro cardíaco	trato respiratório
processo inflamatório	ressonância magnética	sopros diastólicos	trato urinário
processos infecciosos	ressonância nuclear	soro fisiológico	tratos gastrintestinal
produção láctea	ressuscitação volumétrica	soro glicosado	trauma craniano
proliferação ductal	resultados terapêuticos	sorologia positiva	trauma local
pronto atendimento	retardo mental	staphylococcus aureus	trauma mamilar
prontos socorros	retardo psicomotor	substância branca	triagem auditiva
prontuário médico	retorno venoso	sucção digital	triagem metabólica
proteína s	revisão sistemática	sulfato ferroso	triagem neonatal
proteínas plasmáticas	rge fisiológico	suplementos hipercalóricos	trocas gasosas
protocolo bfm	rigidez mandibular	suporte familiar	trombose venosa
pseudomonas aeruginosa	rinite alérgica	suporte nutricional	tronco cerebral
psicoses infantis	rinofaringite aguda	suporte psicológico	tubagem duodenal
psicoses reativas	risco aumentado	suporte ventilatório	tubo endotraqueal
puerpério imediato	risco básico	surfactante exógeno	tubo neural
pulmão direito	risco importante	surfactante pulmonar	tubo traqueal

Bigramas da lista de referência em ordem alfabética — Tabela 6 de 6

tumores intracranianos	vacina antipneumocócica	velocidade relativa	vias aéreas
tumor ósseo	vacina bcg	ventilação adequada	vias biliares
tumores sólidos	vacina conjugada	ventilação alveolar	vida adulta
úlcera duodenal	vacina pneumocócica	ventilação assistida	vida diária
última menstruação	vacina polissacarídica	ventilação convencional	vida saudável
última relação	vacina recombinante	ventilação espontânea	vida sexual
ultra som	valores preditivos	ventilação líquida	vídeo eeg
ultra sonografia	valores basais	ventilação mecânica	violência doméstica
umidade fecal	válvula mitral	ventilação pulmonar	vírus hiv
unidades alveolares	variáveis analisadas	ventiladores mecânicos	vírus respiratório
unidades neonatais	variáveis categóricas	ventrículo direito	vírus sincicial
unidades pediátricas	variáveis contínuas	ventrículo esquerdo	viscosidade sangüínea
uretra anterior	variáveis estudadas	vesícula biliar	volume cardíaco
uretrocistografia miccional	variáveis quantitativas	via endoscópica	volume cerebral
urina centrifugada	variáveis relacionadas	via endovenosa	volume corrente
urografia excretora	variáveis socioeconômicas	via enteral	volume intravascular
uso contínuo	vasos sangüíneos	via inalatória	volume pulmonar
uso pediátrico	vasoconstrição hipóxica	via intradérmica	volumes pequenos
uso prolongado	veia cava	via nasal	x frágil
uso tópico	veia jugular	via oral	zumbido venoso
uti neonatal	veia porta	via sistêmica	
utis pediátricas	veias pulmonares	via vaginal	

Trigramas da lista de referência em ordem alfabética — Tabela 1 de 12

abandono de amamentação	agonistas alfa adrenérgicos	animais de laboratório
abordagem cognitivo comportamental	alça de drigalsky	anos de evolução
abordagem de paciente	alcalóides de ergot	ansiedade de separação
abordagem por neurodesenvolvimento	aleitamento materno exclusivo	antagonistas de leucotrienos
abordagem terapêutica precisa	aleitamento materno predominante	antecedentes de sepse
absorção de água	alimentos semi sólidos	anticorpo anti helicobacter
absorção de cálcio	alimentação com mamadeira	anticorpos pós vacinais
absorção de ferro	alimentação complementar adequada	antígeno polissacarídico capsular
absorção de nutrientes	alimentação complementar saudável	antiinflamatórios não hormonais
abuso de drogas	alimentação de bebês	antiinflamatórios não esteróides
abuso de substâncias	alimentação de criança	antiinfluenza em pacientes
acalasia de esôfago	alimentação de filhos	aparecimento de mastite
ação de insulina	alimentação de lactente	aparecimento de sintomas
ácaros de poeira	alimentação de pacientes	apetite de criança
aceleração de crescimento	alívio de dor	aplicação de bcg
achados de exame	alívio de sintomas	aplicação de imunobiológico
achados ultra sonográficos	alta de berçário	aplicação de surfactante
acidente vascular encefálico	alteração motora oral	aplicação de vacinas
acidentes vasculares isquêmicos	alterações de comportamento	aprendizado de leitura
acidentes de transportes	alterações de sono	apoio a aleitamento
acometimento de membros	alterações ultra sonográficas	apoio a amamentação
acompanhamento de amostra	alternativas de tratamento	apoio a mãe
acompanhamento de crianças	alto fluxo pulmonar	apresentação de fármaco
acompanhamento de paciente	alto valor energético	aquisição de fala
acompanhamento de puericultura	alto valor preditivo	aquisição de infecção
acompanhamento pré natal	altura de enterócito	aquisição de linguagem
aconselhamento em amamentação	altura de indivíduo	aquisição de massa
adequação de crescimento	altura de pais	área de saúde
adequados para idade	amamentação a seio	articulações com sinovite
adesão a dieta	amamentação bem sucedida	asfixia perinatal grave
adiposidade em crianças	amamentação com leite	asma aguda grave
adiposidade em escolares	amamentação de prematuros	asma em crianças
administração de bcg	ambiente de uti	aspiração de mecônio
administração de dose	ambulatório de pediatria	assistência a paciente
administração de droga	aminotransferases em malária	assistência intensiva pediátrica
administração de medicação	amostra de conveniência	assistência pré natal
administração de noi	amostra de exames	associação de sono
administração de oxigênio	amostra de urina	atenção a saúde
administração de surfactante	amostras de colostro	atenção de criança
administração de vacina	amostras de fezes	atenção de pediatras
admissão de paciente	amostras de leite	atendimento a paciente
adoção de medidas	amostras de sangue	atendimento de pacientes
adolescentes com cirrose	amostras de secreção	atendimento de urgência
adolescentes com colestase	analgesia com opióides	ativação de neutrófilos
adolescentes com DC	análise de amamentação	atividade de proteínas
adolescentes com doença	análise de regressão	atividade de doença
adolescentes com hepatopatia	análise de sobrevida	atividade física incorporada
adolescentes com tvp	análise de urina	atividade física regular
afilamento de esôfago	análise morfométrica digitalizada	atividades físicas comuns
agente paralítico ideal	anemia por deficiência	ato de amamentar
agentes de saúde	animais de experimentação	ato de brincar
agitação de paciente	animais de grupo	atraso de desenvolvimento

Trigramas da lista de referência em ordem alfabética — Tabela 2 de 12

atraso de fala	bloqueio de ductos	cicatrização de feridas
atraso de idade	boa evolução clínica	cicatrização de lesão
atresia de vias	boa função polar	ciclo gravídico puerperal
atrofia de hipocampo	boas evidências científicas	circuito de ventilador
atrofia vilosa parcial	boca de bebê	circuitos de cec
atuação em berçário	boca de criança	cirurgia de epilepsia
aumento de morbidade	boca de rn	cirurgias de cardiopatias
aumento de PIC	bomba de coração	cistite não complicada
aumento de átrio	bombas elétricas modernas	cisto de plexo
aumento de colesterol	borda esternal esquerda	classe social alta
aumento de cortisol	borracha de manguito	classe sócio econômica
aumento de dose	bronquiolite viral aguda	classe socioeconômica baixa
aumento de gordura	cabeça de criança	classes mais altas
aumento de idade	camada de ozônio	classes menos favorecidas
aumento de massa	canais de sódio	classes sociais dominantes
aumento de obesidade	canal de crescimento	classificação de doença
aumento de osmolaridade	canal de parto	classificação de estado
aumento de permeabilidade	cânula de traqueostomia	classificação de gravidade
aumento de peso	cânula em pescoço	classificação de risco
aumento de pressão	capacidade de eliminação	classificação de tanner
aumento de resistência	capacidade de simbolização	cloreto de potássio
aumento de risco	capacidade de síntese	cloreto de sódio
aumento de secreções	capacidade de tamponamento	cloro em suor
aumento de ventrículos	capacidade residual funcional	cmo de coluna
ausência de aleitamento	capacidades de criança	coagulação intravascular disseminada
ausência de alteração	capital mineral ósseo	coarctação de aorta
ausência de amamentação	captação de radionuclídeo	colágeno tipo i
ausência de diarréia	captura de imagens	colestase extra hepática
ausência de doenças	caquexia de câncer	coleta de amostra
ausência de dor	caracteres sexuais secundários	coleta de colostro
ausência de efeito	características de amostra	coleta de dados
ausência de fala	características de domicílio	coleta de fezes
ausência de lesões	características de população	coleta de hemocultura
ausência de malformações	características de doenças	coleta de informações
ausência de secreções	características de mães	coleta de sangue
ausência de sintomas	características de pacientes	coleta de urina
autonomia de criança	caracterização de adolescentes	coleta por jato
auxiliar de enfermagem	caracterização de rgep	coleta por saco
avaliação cardiológica minuciosa	carbonato de cálcio	coletas de exames
avaliação de mamada	cardiomiopatia dilatada idiopática	coletores de drenagem
avaliação de crescimento	carência de vitamina	cólica de lactente
avaliação de crianças	carga viral secundária	colonização de nasofaringe
avaliação de desenvolvimento	casca de banana	colonização de orofaringe
avaliação de dor	casos de autismo	colonizadores de orofaringe
avaliação de imunidade	casos de crianças	coloração de fezes
avaliação de indivíduo	casos de constipação	comissão de ética
avaliação de obesidade	casos de dores	comitê de ética
avaliação de paciente	casos de hepatite	comparação de curvas
avaliação de resposta	casos de hidrocefalias	comparação de médias
avaliação de sintomas	casos de hipertensão	comparação de proporções
avaliação pré operatória	casos de avbeh	comparação de variáveis
bacterioscópico de urina	casos de dheg	comparação entre grupos
baixa atividade física	casos de infecção	compatível com amamentação
baixa condição socioeconômica	casos de insuficiência	complexo aréolo mamilar
baixa educação materna	casos de intoxicações	complexo esfincteriano uretral
baixa escolaridade materna	casos de malária	complicação de úlcera
baixo débito cardíaco	casos de morte	complicações de obesidade
baixo desempenho escolar	casos de osteomielites	comportamento de bebê
baixo ganho ponderal	casos de otite	comportamento de crianças
baixo metabolismo ósseo	casos de pneumonia	comportamento de decréscimo
baixo nível social	casos de rm	comportamento de risco
baixo nível socioeconômico	casos de tce	composição de alimentos
baixo peso gestacional	cateter de fibra	composição de vacina
baixo peso molecular	cateteres de sucção	compreensão de escrita
baixo rendimento escolar	causa de dor	compreensão de linguagem
barriga de aluguel	causa de morte	compressas com água
base de paciente	causa de rm	comprometimento de estado
base de soja	causas de delirium	comprometimento de função
base de vitamina	causas de hepatopatia	comprometimento de saúde
bases de tratamento	células musculares lisas	compulsões de verificação
bebês de grupo	células de microglia	conceito de depressão
benefícios de aleitamento	células de purkinje	concentração de cortisol
benefícios de amamentação	células progenitoras hematopoiéticas	concentrações de hemoglobina
benzodiazepínico de ação	centros de saúde	concentrações de retinol
bicarbonato de sódio	cepas bacterianas resistentes	concentração de iga
bicos de mamadeira	cepas de pneumococos	concentração de igg
binômio criança família	cepas de streptococcus	concentração de leptina
biópsia de antro	chances de reoperação	concentração de sódio
biópsia de esôfago	choro de criança	concentração de vitamina
biópsia de mandíbula	choro de lactente	concentração em leite

Trigramas da lista de referência em ordem alfabética — Tabela 3 de 12

concentrado de hemácias	crianças alto xinguanas	curva de peso
concordância entre observadores	crianças com alergia	curva de referência
condição socioeconômica desfavorável	crianças com anemia	curva de Tanner
condição socioeconômica materna	crianças com autismo	curvas de Kaplan
condições de escassez	crianças com câncer	curvas de nchs
condições de nascimento	crianças com choque	curvas de crescimento
condições de saúde	crianças com colestase	curva de sobrevida
condições de vida	crianças com constipação	curvas de velocidade
conduta para fibrilação	crianças com dbp	custo de medicação
confirmação de diagnóstico	crianças com depressão	custos de tratamento
confirmação de intubação	crianças com diarréia	cystic fibrosis foundation
conforto de paciente	crianças com dieta	dados de exames
conhecimento de criança	crianças com doenças	dados de crescimento
conhecimento de escalas	crianças com dores	dados de datasus
conjunto de fatores	crianças com enurese	dados de fase
conjunto de recomendações	crianças com fs	dados de prevalência
conseqüência de desmame	crianças com hepatopatia	dano renal crônico
conseqüências de hidrocefalias	crianças com idade	decisão de amamentar
constatação de Frost	crianças com infecção	decorrência de injúrias
constipação crônica funcional	crianças com itu	defeitos de esmalte
constipação intestinal funcional	crianças com lesão	defeitos de septo
consultas de puericultura	crianças com otite	defeitos estruturais congênitos
consultas de rotina	crianças com pielonefrite	defeitos orovalvares congênitos
consultas pré natais	crianças com risco	defesas de organismo
consultório de pediatra	crianças com rm	deficiência auditiva neonatal
consumo de água	crianças com rvu	deficiência de crescimento
consumo de alimentos	crianças com tdah	deficiência de ferro
consumo de bebidas	crianças com temperamento	deficiência de proteína
consumo de energia	crianças com transtornos	deficiência de surfactante
consumo de fibra	crianças de cor	deficiência de zinco
consumo de medicamentos	crianças de creche	deficiência de vitamina
consumo de oxigênio	crianças de escola	déficit de atenção
consumo de refrigerantes	crianças de risco	déficit de crescimento
contagem de arcos	crianças em acompanhamento	definição de fumante
contagem de blastos	crianças em aleitamento	definição de hipertensão
contagem de corpos	crianças em idade	definição de obesidade
contagem de elementos	crianças extremamente doentes	definição de tratamento
contagem de leucócitos	crianças mais jovens	deiscência parcial recente
contagem de linfócitos	crianças mais novas	demanda metabólica cerebral
contagem de plaquetas	crianças mais velhas	densidade de energia
contato com bebê	crianças não amamentadas	densidade mineral óssea
contato com doentes	crianças pré escolares	dependência de drogas
contato com hospital	crianças pré púberes	dependência de oxigenoterapia
contatos com paciente	crianças que faleceram	depleção de monoaminas
conteúdo de cálcio	crianças sem dbp	depressão de adultos
conteúdo mineral ósseo	crianças sem defeitos	depressão em criança
continuidade de amamentação	crianças sem lesões	derivação urinária temporária
contribuição de alimentos	crise de sibilância	dermatite de contato
controle de PIC	crise de asma	desaceleração de crescimento
controle de apetite	crise de dor	desaparecimento de sintomas
controle de asma	crise de sibilância	descida de leite
controle de crises	crises tônico clônicas	desconforto respiratório agudo
controle de dor	critério de classificação	desconforto respiratório neonatal
controle de peso	critério de exclusão	descontrole de impulsos
controle de hipertensão	critério de normalidade	descrições de casos
controle de infecções	critério de seleção	desenvolvimento de doença
controle de pressão	critérios de gravidade	desenvolvimento de dbp
controle de qualidade	critérios de inclusão	desenvolvimento de linguagem
controle de sintomas	critérios de infecção	desenvolvimento de asma
controle sem hepatopatia	critérios de sic	desenvolvimento de aterosclerose
coorte de crianças	critérios diagnósticos utilizados	desenvolvimento de bebê
cor de pele	critérios de wessel	desenvolvimento de caracteres
correção de rvu	critérios para diagnóstico	desenvolvimento de cárie
correção de yates	critérios pré estabelecidos	desenvolvimento de cicatrizes
correlação de spearman	cromoglicato de sódio	desenvolvimento de criança
córtex pré frontal	cuidados pré natais	desenvolvimento de especialidades
corticóide pré natal	cuidados com criança	desenvolvimento de infecção
cortisol sérico dosado	cuidados com filho	desenvolvimento de osteoporose
coto ureteral residual	cuidados de bebê	desenvolvimento de paciente
crescimento de células	cuidados intensivos neonatais	desenvolvimento de tolerância
crescimento de comprimento	cultura de linfócitos	desenvolvimento de transtorno
crescimento de criança	cultura de orofaringe	desenvolvimento motor oral
crescimento de lactente	cura de rvu	desnutrição protéico calórica
crescimento de perímetro	curso de doença	desnutrição protéico energética
crescimento de rnpt	curso de amamentação	destino de lixo
crescimento intra uterino	curso de medicina	destruição de células
crescimento pôndero estatural	curso doloroso prolongado	detecção de cicatrizes
crescimento pós natal	curva de Alexander	detecção de problemas
criança com hic	curva de aprendizado	detecção do helicobacter
criança com pais	curva de Lubchenco	detecção de lesões

Trigramas da lista de referência em ordem alfabética — Tabela 4 de 12

detecção de obesidade	diminuição de complacência	eclosão de doença
determinação de morbidade	diminuição de efeitos	ecocardiograma com doppler
determinação de susceptibilidade	diminuição de ingestão	efeitos adversos graves
determinação de pressão	diminuição de limiar	efeitos adversos sistêmicos
determinada faixa etária	diminuição de massa	efeitos colaterais graves
dia de alta	diminuição de mortalidade	efeitos de NOI
dia de corticóide	diminuição de perfusão	efeitos de tratamento
diagnóstico de aij	diminuição de reflexo	efeito de idade
diagnóstico de alergia	diminuição de resistência	eficácia de lactação
diagnóstico de atresia	diminuição de síntese	eficácia de vacina
diagnóstico de asma	dimorfismo sexual relacionado	eficácia de NOI
diagnóstico de autismo	dinâmica de crescimento	efusão de orelha
diagnóstico de avbeh	dióxido de nitrogênio	ejeção de leite
diagnóstico de bronquiolite	disfunção de órgãos	elevação de enzimas
diagnóstico de bulimia	disfunção de sistema	elevação de IF
diagnóstico de bva	disfunção de trato	elevações de fósforo
diagnóstico de cardiopatias	disfunções de cérebro	eletrólitos em suor
diagnóstico de cdpr	dislexia de desenvolvimento	elevadores de pálpebra
diagnóstico de dc	disponibilidade de alimentos	emprego de NOI
diagnóstico de depressão	disposição de criança	embriões de galinha
diagnóstico de dgc	dispositivo inalatório ideal	encefalopatia hipóxico isquêmica
diagnóstico de dgh	distensão vesical persistente	endoscopia digestiva alta
diagnóstico de disfunção	distribuição de crianças	endotélio de linfangioma
diagnóstico de doença	distribuição de pacientes	ensaio clínico prévio
diagnóstico de dsr	distrofia simpática reflexa	ensaios clínicos randomizados
diagnóstico de endocardite	distúrbio de espectro	enterocolite necrosante neonatal
diagnóstico de enxaqueca	distúrbio de linguagem	entrada de ar
diagnóstico de er	distúrbio ventilatório obstrutivo	envolvimento de membro
diagnóstico de esquizofrenia	distúrbios de alimentação	envolvimento de SNC
diagnóstico de hepatopatia	distúrbios de coagulação	enxaqueca com aura
diagnóstico de hidrocefalia	distúrbios de desenvolvimento	enxaqueca sem aura
diagnóstico de hipertensão	distúrbios de comportamento	epilepsias generalizadas idiopáticas
diagnóstico de infecção	distúrbios de sono	episódio de diarréia
diagnóstico de insuficiência	diurético de alça	episódio de infecção
diagnóstico de irab	divórcio de pais	episódios de agitação
diagnóstico de ITU	doador não aparentado	episódio de ITU
diagnóstico de lesão	doença arterial coronariana	episódio de RGE
diagnóstico de malária	doença cardíaca congênita	episódios de enurese
diagnóstico de membrana	doença de base	episódios de hematêmese
diagnóstico de miocardite	doença de chagas	episódios de regurgitações
diagnóstico de obesidade	doença de kawasaki	episódios de sepse
diagnóstico de ocmr	doença de membrana	episódios de sibilância
diagnóstico de osteomielite	doença diarréica aguda	epitélio de vaca
diagnóstico de otites	doença invasiva pneumocócica	época de diagnóstico
diagnóstico de paciente	doença de parênquima	época de internação
diagnóstico de pneumonia	doença pneumocócica invasiva	equação de Slaughter
diagnóstico de primoinfecção	doença viral prévia	equipe de pesquisa
diagnóstico de ra	doenças auto imunes	equipe de pesquisadores
diagnóstico de rge	doenças crônico degenerativas	equipe de saúde
diagnóstico de rm	doenças de tireóide	equipes de profissionais
diagnóstico de rvu	doenças sexualmente transmissíveis	erradicação de bactéria
diagnóstico de sam	domínio de linguagem	erros de interpretação
diagnóstico de sdra	dor abdominal funcional	escalas de dor
diagnóstico de sepse	dor abdominal recorrente	escape de ar
diagnóstico de sinusite	dor de neonato	escolaridade de mãe
diagnóstico de sn	dor em mamilos	escolaridade de pais
diagnóstico de st	dores de crescimento	escolas de saúde
diagnóstico de tdah	dosagem de igg	escore clínico modificado
diagnóstico pré natal	dose de ACTH	escore de gravidade
diagnóstico de transtorno	dose de antitérmico	escore de Shwachman
dias de avaliação	dose de insulina	escores de Apgar
dias de entrevista	dose de manutenção	escore de Williams
dias de internação	doses de medicações	escore Snappe II
dieta de criança	doses de medicamentos	esfíncter esofágico superior
dieta de exclusão	drenagem de liquor	esôfago de barrett
dieta de lactente	droga de escolha	espaçadores de metal
dieta de mãe	droga em plasma	espectro obsessivo compulsivo
dieta sem glúten	droga em leite	espessamento de dieta
dificuldade de diagnóstico	drogas de abuso	esquemas de tratamento
dificuldade de intubação	ducto biliar comum	esquema vacinal completo
dificuldade de leitura	duração de aij	esquizofrenia com início
dificuldade de pais	duração de aleitamento	esquizofrenia de início
dificuldade em adormecer	duração de am	esquizofrenia em homens
dificuldades de amamentação	duração de amamentação	esquizofrenia em infância
dificuldades de aprendizagem	duração de doença	estabelecimento de aleitamento
dificuldades de linguagem	duração de mamadas	estabelecimento de amamentação
dificuldades de sucção	duração de queixa	estabelecimento de diagnóstico
dificuldades durante intubação	duração de remissão	estabelecimento de lactação
dilatação de pupila	duração de tratamento	estabelecimento de lesão
diluições decimais selecionadas	duração de ventilação	estado de portador

Trigramas da lista de referência em ordem alfabética — Tabela 5 de 12

estado de remissão	existência de doença	fornecimento de oxigênio
estado de saúde	expectativa de cura	fragmentos de biópsia
estado infeccioso grave	expectativa de vida	freqüência de amamentação
estado nutricional materno	experiência com cigarro	freqüência de crises
estágio de desenvolvimento	experiência com tabaco	freqüência de mamadas
estágio de doença	exposição a agentes	freqüência de positividade
estatura de crianças	exposição a forças	freqüência de asma
estatuto da criança	exposição a agentes	freqüência de atelectasia
estenose aórtica grave	exposição a luz	freqüência de depressão
estilo de vida	exposição a sol	freqüência de alterações
estimulação de linguagem	exposições a medicamentos	friabilidade em esôfago
estratégia de abordagem	expressão de linguagem	função de células
estratégia de atenção	extensão de doença	função de eixo
estratégia de tratamento	extração de leite	função de língua
estratégias de controle	fabricantes de leites	função de músculo
estratégias de ventilação	faculdade de medicina	função de sinapses
estresse pós traumático	faixa de normalidade	função de ventrículo
estudante de medicina	faixa de peso	funcionamento de criança
estudo clínico controlado	faixa etária atendida	funcionamento de paciente
estudo com crianças	faixa etária estudada	funções corticais superiores
estudo de coagulação	faixas de idade	gânglios de base
estudo de Crowcroft	faixas etárias pediátricas	ganho de comprimento
estudo de imagem	falta de apetite	ganho de massa
estudo de SDRA	falta de autoconfiança	ganho de perímetro
estudo de Souza	falta de apetite	ganho de peso
estudo por imagem	falta de controle	ganho pôndero estatural
estudos com adolescentes	falta de evidência	gema de ovo
estudos com corticóides	falta de experiência	genes de sistema
estudos com famílias	falta de informações	gênese de hipertensão
estudos com pacientes	falta de resposta	gênese de osteopenia
estudos com adultos	falta de tempo	geração de cpap
estudos de coorte	falta de treinamento	germes mais freqüentes
estudos de genética	família de criança	giro temporal superior
estudos de neuroimagem	famílias mais carentes	glicemia de jejum
estudos de prevalência	farelo de aveia	gluconato de cálcio
estudos de seguimento	farelo de trigo	gordura em fezes
estudos em adultos	fase de consolidação	gorduras de dieta
estudos em animais	fase de crescimento	gráficos de crescimento
estudos em crianças	fase de doença	grau de desnutrição
estudos genético familiares	fase de indução	grau de disfunção
estudos in vitro	fase de manutenção	grau de esofagite
estudos não controlados	fase de preparação	grau de hp
esvaziamento de mama	fase de tratamento	grau de instrução
etiologia de AVBEH	fase de vida	grau de satisfação
etiologia de colestase	fator de ativação	grau de relaxamento
etiologia de esquizofrenia	fator de crescimento	graus de complexidade
etiologia de obesidade	fator de necrose	gravidade de doença
etiologia de rm	fator v leiden	gravidade de asma
etiologia de TDAH	fatores de coagulação	gravidade de crise
etiologia de transtorno	fatores de confusão	gravidade de bva
etiopatogenia de SAM	fator de proteção	gravidade de clpe
evidência de colestase	fatores de risco	gravidade de desconforto
evidência de doença	fatores prognósticos analisados	gravidade de problema
evidência de sinovite	fatores prognósticos desfavoráveis	gravidade de quadro
evidências de benefícios	fechamento de tubo	grupo com analgesia
evidências de literatura	fenômeno de Raynaud	grupo com enterocolite
evolução de crianças	fim de vida	grupo de adolescentes
evolução de doença	final de adolescência	grupo de cirurgia
evolução de gravidez	final de expiração	grupo de hidrocefalias
evolução de ICT	fisiologia de lactação	grupo de leite
evolução de paciente	fisiopatologia de doença	grupo de mães
evolução de rn	fisiopatologia de síndrome	grupo de médicos
evolução neurológica anormal	fluxo sangüíneo cerebral	grupo de nível
exacerbação de asma	fluxo sangüíneo hepático	grupo de pacientes
exame de eda	fome de bebê	grupo de puérperas
exame de rotina	fonte de fibra	grupo de seguimento
exame de urina	fonte de infecção	grupo de vacinas
exame físico geral	fonte de informações	grupos de risco
exame neuro oftalmológico	fonte de oxigênio	grupos de tratamento
exame ultra sonográfico	força de associação	grupo em dieta
exames de imagem	forma de aprendizado	grupo não osteopênico
exames de neuroimagem	forma de comprimidos	grupo obeso feminino
exames de triagem	forma de esquizofrenia	grupo sem analgesia
exato de Fisher	forma de gordura	grupos de crianças
excesso de peso	formação de cicatrizes	habilidades cognitivas adequadas
exclusão de alimentos	formação de edema	habilidades de aconselhamento
exercício de sexualidade	formação de osso	habilidades de comunicação
exercícios contra gravidade	formação de vínculo	habilidades de criança
exercícios de relaxamento	formas de tratamento	hábito de fumar
existência de associação	fórmula de soja	hábitos alimentares inadequados

Trigramas da lista de referência em ordem alfabética — Tabela 6 de 12

hábitos de sucção	incremento de peso	inoculação de urina
hábitos de vida	independente de idade	instituição de desmame
hábitos orais deletérios	indicação de analgesia	instituição de ensino
hábitos orais nocivos	indicação de analgésicos	instituição de saúde
hemorragia digestiva alta	indicação de antitussígenos	instituição de tratamento
hepatite aguda viral	indicação de cirurgia	instituição de ventilação
herança autossômica recessiva	indicação de ecmo	instituto da criança
hidrato de cloral	indicação de fototerapia	instituto fernandes figueira
hidratos de carbono	indicação de surfactante	instrumento de avaliação
higiene de alimentos	indicação de ventilação	instrumentos de investigação
hiper responsividade brônquica	indicação para frenectomia	insucesso de tratamento
hipersecreção de glucocorticóides	indicação para suspensão	insuficiência cardíaca congestiva
hipertensão arterial grave	indicações de intubação	insuficiência cardíaca crônica
hipertensão arterial secundária	indicações de tch	insuficiência hepática aguda
hipertensão arterial sistêmica	indicador de risco	insuficiência renal aguda
hipertensão em crianças	indicador perímetro braquial	insuficiência renal crônica
hipertensão endocraniana refratária	indicadores de aleitamento	insuficiência respiratória aguda
hipertensão intracraniana refratária	índice cardiotorácico médio	insuficiência respiratória grave
hipertensão pulmonar persistente	índice de apgar	insuficiência respiratória hipoxêmica
hipertensão pulmonar primária	índice de equilíbrio	insuficiência supra renal
hipertensão sistólica isolada	índice de massa	insulina de jejum
hipertrofia de parótidas	índice de oxigenação	integridade de haste
hipertrofia ventricular esquerda	índices de impedância	integridade de mucosa
hipoplasia de esmalte	índices de capacidade	inteligência de crianças
hipotálamo hipófise adrenal	indivíduos com anemia	intensidade de dor
hipótese de nulidade	indivíduos masculinos xyy	intensidade de exposição
hipotonia de musculatura	indução de lesão	intensidade de febre
história de alergia	indução de remissão	intensidade de rm
história de prematuridade	indução de sedação	intensidade de sintoma
história de sibilância	infecções de repetição	intercorrências de mama
história familiar positiva	infecção de criança	interesse de criança
hora de alimentação	infecção de trato	internação de rn
hora de dormir	infecção por HIV	internação em uti
hora de morte	infecção por HP	interpretação de densitometria
hormônio de crescimento	infecção por vrs	interrupção de aleitamento
idade de crianças	infecção pós natal	interrupção de amamentação
idade de lactente	infecções de orelha	intervenção de pais
idade de início	infecções respiratórias agudas	intervalo de tempo
idade de mães	infecções respiratórias virais	intoxicação por clonidina
idade de paciente	influência de sexo	introdução de alimentação
idade gestacional corrigida	informações de pacientes	introdução de alimentos
idade pós concepcional	infusão de células	introdução de chás
idade pré escolar	infusão de propofol	introdução de dieta
identificação de agente	infusão de líquidos	introdução de medicamentos
identificação de causa	infusão de solução	introdução de sulfasalazina
identificação de colapsos	ingestão de alimentos	intubação de paciente
identificação de genótipos	ingestão de cálcio	intubação sem medicação
identificação de helicobacter	ingestão de calorias	intubações de emergência
identificação de paciente	ingestão de energia	invasão de estruturas
identificação de rotavírus	ingestão de gordura	investigação de rm
identificação de vírus	ingestão de leite	irradiância espectral média
identificação etiológica viral	ingestão de sódio	itens de prescrição
ige sérica específica	ingestão de vitamina	ITU pós operatória
imagem de criança	início de adolescência	jornal de pediatria
imagens ultra sonográficas	início de aij	lábio superior fino
impacto em mortalidade	início de aleitamento	laboratório de microbiologia
importância de aleitamento	início de amamentação	lactentes com sibilância
importância de desnutrição	início de analgesia	lactentes não amamentados
importância de diagnóstico	início de antibioticoterapia	lanolina anidra modificada
importância de sucção	início de dieta	lateralidade de rvu
imunização básica completa	início de doença	leite de mãe
inalações com broncodilatador	início de estudo	leite de mama
incentivo a aleitamento	início de exantema	leite de peito
incentivo a amamentação	início de febre	leite de vaca
incidência de complicações	início de mamadas	leite humano ordenhado
incidência de doença	início de manifestações	leite materno exclusivo
incidência de meningite	início de noite	leite materno ordenhado
incidência de cardiopatias	início de processo	leitos de terapia
incidência de cicatriz	início de puberdade	leitura de cicatriz
incidência de cólica	início de quadro	leptina de cordão
incidência de dbp	início de rações	lesão cerebral isquêmica
incidência de hemorragia	início de seguimento	lesão cerebral secundária
incidências de hidrocefalias	início de sintoma	lesão cerebral traumática
incidência de infecções	início de sono	lesão de tecido
incidência de insuficiência	início de sucção	lesão glomerular mínima
incidências de meningite	início de tratamento	lesão isquêmica cerebral
incidência de otite	início de ventilação	lesão pulmonar aguda
incidência de pneumotórax	início de vida	lesão térmica grave
incidências para dbp	início em infância	lesões de gravidade
inclusão de pacientes	injeção de toxina	lesões de pele

Trigramas da lista de referência em ordem alfabética — Tabela 7 de 12

lesões não relativas	medidas preventivas eficazes	níveis de leptina
leucemia linfocítica aguda	medidas sanitárias urgentes	níveis de linfócitos
leucemia mielóide aguda	medo de doença	níveis de prolactina
liberação de histamina	megadoses de vitamina	níveis de retinol
liberação de leite	meio de contraste	níveis de vitamina
liberação de ocitocina	membro superior esquerdo	nível de bilirrubinemia
limitação de estudo	meningite por haemophylus	nível de confiança
limitação de movimentos	meningite por hib	nível de cricóide
limitações de atividades	meningites em rs	nível de escolaridade
limites de normalidade	metabolismo de ácido	nível de evidência
linfonodomegalia cervical dolorosa	metabolismo de cálcio	nível de hilo
linha axilar média	metabolismo de lactente	nível de maturação
linha de pensamento	metabolismo de repouso	nível de rejeição
lipodistrofia generalizada congênita	método de ballard	nível socioeconômico inferior
locais de assistência	método de elisa	nodularidade de borda
local de trabalho	método de escolha	número de articulações
localização de infecção	método de investigação	número de casos
localização de tubo	método de kaplan	número de consultas
lúpus eritematoso sistêmico	método de prechtl	número de crianças
má formação congênita	método de tratamento	número de desvios
má perfusão orgânica	método de triagem	número de doses
mãe com tuberculose	ministério da saúde	número de estudantes
mãe de bebê	modelo de count	número de gasometrias
mãe de criança	modelo de cox	número de hospitalizações
mãe de paciente	modo de herança	número de intubações
mãe hiv positivo	modo de ventilação	número de leitos
mães de rnpt	momento de admissão	número de leucócitos
mal de ausência	momento de alimentação	número de mamadas
malária em infância	momento de alta	número de neurônios
malformações de trato	momento de coleta	número de neutrófilos
manejo de criança	momento de diagnóstico	número de pacientes
manejo de obesidade	momento de intubação	número de procedimentos
manejo de aleitamento	momento de internação	número de rn
manejo de lactação	momento de parto	número de sinapses
manejo de sdra	monitorização de crescimento	número de sintomas
manejo de trauma	monitorização de paciente	nutrição de criança
manobra de sellick	monitorização de pic	obesidade de filhos
manutenção de amamentação	monitorização de pressão	obesidade de pais
manutenção de asma	mordida aberta anterior	obesidade em crianças
manutenção de confiança	mortalidade em pacientes	obesidade em infância
manutenção de lactação	morte de criança	óbitos por diarréia
manutenção de níveis	motivo de encaminhamento	objetivo de pesquisa
manutenção de pacientes	motivo de intubação	objetivo de quimioprofilaxia
manutenção de recrutamento	movimentos de extremidades	objetivo de tratamento
manutenção de ppc	movimentos de língua	objetivos de estudo
mapas de micção	movimentos de mastigação	objeto de estudo
marcadores de reabsorção	mudança de hábitos	observação de mãe
massa óssea relacionada	mudanças de comportamento	observação de mamadas
maternidade de caism	mutações de gene	obstáculo a amamentação
maternidade de país	nascimento de bebê	ocorrência de cólica
mau controle metabólico	nascimento de crianças	ocorrência de complicações
maus hábitos orais	nascimento de irmãos	ocorrência de diarréia
mecanismo de lesão	necessidade de assistência	ocorrência de doenças
mecanismo de ação	necessidade de cuidados	ocorrência de evento
mecanismo de defesa	necessidade de ecmo	ocorrência de fraturas
mecanismo de repulsa	necessidade de estudos	ocorrência de hemorragia
mecanismo de sucção	necessidade de internação	ocorrência de infecções
mecanismo imunológico envolvido	necessidade de intubação	ocorrência de interação
média de crianças	necessidade de leitos	ocorrência de lesões
média de escores	necessidade de oxigênio	ocorrência de obesidade
média de idade	necessidade de oxigenoterapia	ocorrência de óbito
média de permanência	necessidade de procedimentos	ocorrência de síndrome
média de peso	necessidade de reintubação	ocorrência de soroproteção
mediana de idade	necessidade de suporte	ocorrência de vômitos
mediana de amamentação	necessidade de tratamento	oferta de leitos
medicamentos entre adolescentes	necessidade de ventilação	oferta de oxigênio
medicamentos não aprovados	necessidades de ferro	opção de tratamento
medicamentos não padronizados	necrólise epidérmica tóxica	opinião de autores
medicina de emergência	necrose de pele	opinião de especialistas
medicina pré paga	neonatos com hipertensão	ordem de nascimento
medida de associação	neonatos de peso	organizações não governamentais
medidas de dobras	neuro hipófise ectópica	orientação de condutas
medidas de estatura	neurobiologia de comportamento	origem de criança
medida de perímetro	neurobiologia de tdah	osmolalidade de plasma
medidas de peso	nitroprussiato de sódio	otite média aguda
medidas de pregas	níveis de anticorpos	otite média crônica
medidas de comprimento	níveis de bilirrubina	otoemissão acústica alterada
medidas de suporte	níveis de cloro	otoemissão acústica evocada
medida de pressão	níveis de hemoglobina	óxido nítrico inalatório
medidas de prevenção	níveis de ige	oxigênio de hemoglobina

Trigramas da lista de referência em ordem alfabética — Tabela 8 de 12

oximetria de pulso	perda de apetite	postos de vacinação
pacientes com anemia	perda de calor	prática de aleitamento
paciente com câncer	perda de consciência	prática de amamentação
pacientes com respiração	perda de força	prática de consultório
pacientes com aij	perda de massa	prática de ginástica
pacientes com alergia	perda de nutrientes	predição de evolução
pacientes com anorexia	perda de peso	prejuízo de sono
pacientes com artrites	perda de pressão	preparo de alimentos
pacientes com asma	perda de seguimento	preparo de medicações
pacientes com choque	perfil de sorotipos	prescrição de antibióticos
pacientes com cicatriz	perfis de pacientes	prescrição de medicamentos
pacientes com cirrose	perfis de evolução	prescrição de pacientes
pacientes com colite	perímetro de cintura	presença de anemia
pacientes com dgh	período de acompanhamento	presença de anormalidades
pacientes com diagnóstico	período de coleta	presença de anticorpos
pacientes com doença	período de cólicas	presença de bactérias
pacientes com enterocolite	período de incubação	presença de cardiomegalia
pacientes com esquizofrenia	período de internação	presença de cateteres
pacientes com fc	período de seguimento	presença de cicatriz
pacientes com hipertensão	período de tempo	presença de cólicas
pacientes com infecção	período de ventilação	presença de débito
pacientes com insuficiência	período pós natal	presença de defeitos
pacientes com leucemia	período pós operatório	presença de doença
pacientes com malária	período pós vacinal	presença de dor
pacientes com meningite	período pré natal	presença de estridor
pacientes com metástases	períodos de sono	presença de febre
pacientes com osteopenia	períodos de vida	presença de hábitos
pacientes com refluxo	permeabilidade de membrana	presença de hemocultura
pacientes com relaxamento	peroxidação de lipídeos	presença de ige
pacientes com sam	persistência de bacteriúria	presença de infecções
pacientes com sd	persistência de canal	presença de insuficiência
pacientes com sdra	persistência de febre	presença de irmãos
pacientes com sepse	peso de crianças	presença de microrganismos
pacientes com sibilância	peso de nascimento	presença de mutação
pacientes com st	peso de pacientes	presença de obsessões
pacientes com tce	peso de placenta	presença de osteopenia
pacientes com tdah	peso fecal seco	presença de pais
pacientes com toc	peso fecal úmido	presença de processos
pacientes com transtorno	pico de incidência	presença de refluxo
pacientes com trauma	pico de massa	presença de regurgitação
pacientes com trombose	pico de pressão	presença de respostas
pacientes com tvp	pinças de biópsia	presença de sibilos
pacientes com ventilação	piora de hipertensão	presença de sintomas
pacientes hiv positivos	planejadores de saúde	presença de tiques
padrões de consumo	plicatura de diafragma	pressão arterial diastólica
padrões de crescimento	pneumonia em crianças	pressão arterial elevada
padrões de normalidade	pneumonia de aquisição	pressão arterial média
padrão de referência	pneumonias de repetição	pressão arterial normal
padrão de respostas	polissacarídeo de soja	pressão arterial pulmonar
padrão de distribuição	polissacarídeos não celulósicos	pressão arterial sistêmica
padrão de herança	políticas de saúde	pressão arterial sistólica
padrão de sensibilização	pomadas com corticóide	pressão capilar pulmonar
pais de crianças	ponta de língua	pressão de átrio
pai de pacientes	ponte nasal achatada	pressão de perfusão
palpação de pulsos	população alto xinguana	pressão intra oral
pangastrite erosiva hemorrágica	população de adultos	pressão intracraniana elevada
parada cardíaca repentina	população de crianças	pressão positiva contínua
parâmetros de mineralização	população de estudo	pressão venosa central
parâmetros de respirador	população de linfócitos	prevalência de aleitamento
parâmetros de ventilador	população de referência	prevalência de alterações
parte de investigação	população de risco	prevalência de amamentação
parte de malformações	porcentagem de gordura	prevalência de anemia
parte de tratamento	porcentagem de linfócitos	prevalência de asma
partes de cérebro	porta de entrada	prevalência de clpe
partes de corpo	portador de deficiências	prevalência de colonização
participação de ácaros	portadores de hib	prevalência de deficiência
participação de criança	portadores de doença	prevalência de diarréia
parturiente não preparada	portadores de hepatite	prevalência de distúrbios
passagem de tubo	portadores de malária	prevalência de doenças
patologia de base	portoenterostomia de kasai	prevalência de dtn
pediatra de emergências	pós cirurgia cardíaca	prevalência de ea
pega de bebê	pós parto imediato	prevalência de infecção
pequenas vias aéreas	posição de rn	prevalência de obesidade
percentil de estatura	posições de mamadas	prevalência de osteopenia
percentis de peso	positivação de rcht	prevalência de rgep
percentuais de linfócitos	possibilidade de diagnóstico	prevalência de sensibilização
percentual de admissões	possibilidade de infecção	prevalência de sobrepeso
percepção de paciente	possibilidade de tratamento	prevalência de soropositividade
percepção de pais	possibilidades de intervenção	prevalência de tabagismo
perda de aerossol	postos de saúde	prevalência de tid

Trigramas da lista de referência em ordem alfabética — Tabela 9 de 12

prevalência de transtorno	prontuário de paciente	refeições com alimentos
prevalência de tuberculose	proporção de crianças	refeições de sal
prevenção de doenças	proporção de gestantes	reflexo de busca
prevenção de dtn	proteção contra obesidade	reflexo de sucção
prevenção de eventos	proteção contra tuberculose	refluxo gastroesofágico patológico
prevenção de fissuras	proteção de leite	região de antro
prevenção de infecções	proteína de soja	região de coluna
prevenção de injúrias	proteínas de surfactante	regime de condicionamento
prevenção de morbimortalidade	protetores de mamilo	regiões de cérebro
prevenção de obesidade	protocolo de estudo	regiões de origem
prevenção de osteoporose	protocolo de kyoto	registro de medicamentos
primeira amostra fecal	protocolo de pesquisa	registro de pressão
primeiras manifestações clínicas	protocolos de investigação	regressão de cox
principais agentes etiológicos	protocolos de tratamentos	regressão de febre
principais efeitos colaterais	protrusão de língua	regressão linear múltipla
principais sinais clínicos	quadros de paralisias	regressão logística múltipla
principal manifestação clínica	qualidade de ingestão	regressão logística ordinal
probabilidade de doença	qualidades de leite	regressão logística univariada
probabilidade de sle	qualidade de vida	regressão não linear
probabilidade de sobrevida	quantidade de alimentos	regulação de balanço
problema de crescimento	quantidade de células	relação com óbito
problema de saúde	quantidade de leite	relação mãe filho
problemas com aleitamento	queixa de encaminhamento	relação médico paciente
problemas com amamentação	quelato de zinco	relatos de casos
problemas com amígdala	questões de saúde	relatos de literatura
problemas de amamentação	quimioterapia de indução	relaxamento muscular adequado
problemas de aparelho	quimioterapia pré operatória	relaxamento muscular inadequado
problemas de comportamento	quociente de inteligência	remissão clínica completa
problemas de desenvolvimento	ração com celulose	remissão de quadro
problemas de sono	radiografia de tórax	remoção de ce
problemas pós operatórios	radiografias de coluna	remoção de msv
procedimentos anti refluxo	radiograma de tórax	renda per capita
procedimentos de intubação	raiz de aorta	repetição de exames
procedimentos de reanimação	razão de prevalência	replicação de hiv
procedimentos potencialmente dolorosos	razão de probabilidade	reservas de ferro
processo de adaptação	razões de chance	reservatório de oxigênio
processo de amamentação	reação de hospedeiro	resistência a drogas
processo de avaliação	reação em cadeia	resistência a insulina
processo de cárie	reações de hipersensibilidade	resistência a penicilina
processo de crescimento	reações por ainh	resistência in vitro
processo de envelhecimento	realização de análise	resistência vascular periférica
processo de linguagem	realização de atividades	resistência vascular pulmonar
processo de luto	realização de cirurgia	resistência a fluxo
processo de maturação	realização de ecocardiograma	responsáveis por crianças
processo de pasteurização	realização de entrevistas	resposta a acetilcolina
procura de cci	realização de estudo	resposta a alterações
produção de acth	realização de exames	resposta a estímulos
produção de anticorpos	realização de procedimentos	resposta a NOI
produção de cfc	realização de rxt	resposta a terapêutica
produção de citocinas	realização de testes	resposta a tratamento
produção de fala	realização de traqueostomia	resposta a vacina
produção de gases	realização de triagem	resposta a vacinação
produção de IF	recaptação de serotonina	resposta de anticorpos
produção de leite	recém nascidos controles	resposta de hospedeiro
produtos de distorção	recém nascidos menores	resposta de soroproteção
profilaxia com antimicrobianos	receptor de angiotensina	resposta inflamatória sistêmica
profilaxia com penicilina	receptor de interferon	ressonância nuclear magnética
profilaxia de snc	receptores de melanocortina	ressuscitação com solução
profissional de enfermagem	receptor de transferrina	ressuscitação com volumes
profissionais de saúde	receptores pós sinápticos	resultado de análise
prognóstico de autismo	recidiva de doença	resultado de teste
prognóstico de avbeh	recomendação de amamentação	resultados de ensaio
prognóstico de crianças	recomendações de OMS	resultados de estudos
prognóstico de doença	recorrência de itu	resultados de exames
prognóstico de paciente	recuperação de criança	resultados de pesquisa
programa de educação	recuperação de hemoglobina	resultados de terapia
programa de estímulo	redução de apetite	resultados de USC
programas de intervenção	redução de carga	retardo de crescimento
programas de prevenção	redução de custos	retardo em diagnóstico
programas de screening	redução de efeitos	retinol em colostro
programas de treinamento	redução de hipoxemia	retinopatia de prematuridade
programas de triagem	redução de mortalidade	retirada de leite
progressão de aparelho	redução de nbt	retirada de MSV
progressão de doença	redução de pic	reversibilidade de HP
progressão de lesão	redução de pressão	revisão de prontuários
prolapso de válvula	redução de massa	rigidez de parede
promoção de alimentação	redução de peso	risco de atrasos
promoção de amamentação	redução de proteína	risco de arritmias
promoção de saúde	redução de sedentarismo	risco de aspiração
pronto atendimento pediátrico	refeição de família	risco de complicações

Trigramas da lista de referência em ordem alfabética — Tabela 10 de 12

risco de contaminação
risco de depressão
risco de doenças
risco de fraturas
risco de infecção
risco de intoxicação
risco de itu
risco de morbidade
risco de morte
risco de obesidade
risco de óbito
risco de pneumotórax
risco de prejuízo
risco de reação
risco de recorrência
risco de rm
risco de sobrepeso
risco de transmissão
risco de tromboembolismo
risco de vida
riscos de mortalidade
ritmo de crescimento
rotina de assistência
rotina de hospital
rotinas de maternidades
sais de cálcio
sais de ouro
sala de aula
sala de emergências
sala de parto
sangue de cordão
sarcoma de ewing
satisfação de famílias
saturação de hemoglobina
saturação de oxigênio
saúde de criança
saúde de mulher
saúde de paciente
saúde de população
saúde materno infantil
secreção ácido péptica
secreção de cortisol
secreção de insulina
secreção de leite
secreção de orofaringe
secreção de prolactina
secreção nasal purulenta
segurança de drogas
secretaria de saúde
segmento polar superior
segmentos renais remanescentes
seguimento de pacientes
seguimento de puericultura
seios de face
sepse neonatal precoce
septicemia por salmonela
seqüelas de intubação
seqüência de eventos
seqüência de intubação
série de casos
série de estudos
série de fatores
série de reações
serviço de medicina
serviços de referência
serviços de pediatria
serviços de emergência
serviços de neonatologia
serviços de puericultura
serviços de saúde
serviços de urgência
setor de urgência
severidade de doença
sexo de crianças
sexo de pacientes
sibilância de repetição
sinal de alerta
sinais de doença
sinais de abstinência
síndrome de abstinência
síndrome de álcool
síndrome de angelman
síndrome de asperger
síndrome de aspiração
síndrome de desconforto
síndrome de down
síndrome de hipermobilidade
síndrome de horner
síndrome de imunodeficiência
síndrome de meckel
síndrome de morte
síndrome de prader
síndrome de resposta
síndrome de ovários
síndrome de rett
síndrome de west
síndrome hemolítico urêmica
síndrome torácica aguda
síntese de leite
síntese de serotonina
sintomas de asma
sintomas de doença
sintomas de hiponatremia
sintomas de ic
sintomas de infecção
sintomas de obstrução
sintomas de resfriado
sintomas de retinoblastoma
sintomas de sepse
sintomas obsessivo compulsivos
sintomas urinários diurnos
sistema de escore
sistema de saúde
sistema nervoso autônomo
sistema nervoso central
situação de saúde
situações de emergência
situações de estresse
situação de risco
situações de violência
sn córtico sensível
sobrevida de criança
sobrevida de pacientes
sódio em dieta
soja sem fibra
solicitação de exames
solução de NaCl
soluções salinas hipertônicas
sono de crianças
sons de fala
sopro de ejeção
sopro de ramos
sopro de still
sopros cardíacos inocentes
soro de pacientes
soropositividade em gestantes
soroprevalência de infecção
sorotipos de streptococcus
subdiagnóstico de doença
subgrupo de hidrocefalias
subgrupos de pacientes
subpopulações de linfócitos
subtipos de lma
sucção de rn
sucção em seio
sucção não nutritiva
sucesso de extubação
sucesso de aleitamento
suco de frutas
sugestivas de cardiopatia
superfície de células
superfície de mucosa
supervisão de saúde
suplementação com ferro
suplementação com zinco
suplementação de oxigênio
suplemento de cálcio
suplemento de vitamina
suporte de terapia
suporte de UTI
suporte ventilatório invasivo
supressão de lactação
surgimento de cárie
suspeita de infecção
suspensão de aleitamento
suspensão de quimioterapia
suspensão de tratamento
swab de orofaringe
tabela de contingência
tabelas de nchs
tabelas de referência
tamanho de amostra
tamanho de câmara
tamanho de criança
tamanho de máscara
tamanho de osso
tamanho de partícula
taxa de crescimento
taxa de cura
taxa de fumantes
taxa de ganho
taxa de incremento
taxa de infecção
taxa de metabolismo
taxa de mortalidade
taxa de remissão
taxas de aleitamento
taxas de amamentação
taxas de hospitalização
taxas de internação
taxas de prevalência
taxas de sobrevida
taxas de soroproteção
tecido adiposo fetal
tc de crânio
técnica de aleitamento
técnica de amamentação
técnica de aplicação
técnica de hemocitômetro
técnica de inóculo
técnica de medida
técnica de microdiluição
técnica de pcr
técnica de pour
técnica de regressão
técnica de translactação
técnicas moleculares modernas
telerradiografia de tórax
tempo de acompanhamento
tempo de aleitamento
tempo de avaliação
tempo de colestase
tempo de coleta
tempo de consulta
tempo de evolução
tempo de hospitalização
tempo de internação
tempo de oxigenoterapia
tempo de queixa
tempo de sobrevida
tempo de tratamento
tempo de trombina
tempo de uso
tempo de ventilação
tempos de coagulação
tentativas de intubação
teor de vitamina
terapêutica anti retroviral
terapêutica de manutenção
terapêutica de resgate
terapia anti retroviral
terapia com noi
terapia com surfactante
terapia de sepse
terapia de suporte
terapia intensiva neonatal
terapia intensiva pediátrica
término de estudo
término de tratamento
termo de compromisso

Trigramas da lista de referência em ordem alfabética — Tabela 11 de 12

termo de consentimento
termômetro em reto
teste com acth
teste cutâneo negativo
teste cutâneo positivo
teste de coombs
teste de desenvolvimento
teste de emissões
teste de estimulação
teste de fisher
teste de mcnemar
teste de otoemissão
teste de pezinho
teste de qi
teste de resposta
teste de tolerância
teste de urease
teste in vitro
teste wppsi r
testes de metapirona
testes de triagem
testes in vivo
tipo de aleitamento
tipo de alimentação
tipo de assistência
tipo de atendimento
tipo de cardiopatia
tipo de crise
tipo de dieta
tipo de dor
tipo de doença
tipo de enurese
tipo de escola
tipo de infecção
tipo de injúria
tipo de manifestação
tipo de mecanismo
tipo de estudo
tipo de medicamento
tipo de parto
tipo de procedimento
tipo de produto
tipo de risco
tipo de terapia
tipo de profilaxia
tipo de sintomatologia
tipo de ventilação
tipo de tratamento
tipos de alimentos
tipos de enxaqueca
tipos de lesão
tipos de inaladores
tipos de injúrias
títulos de anticorpos
títulos de soroproteção
tomada de decisão
toxicidade de drogas
toxoplasmose em gestantes
trabalhadores de saúde
trabalho de annane
trabalho de parto
trabalho de revisão
trabalhos de campo
tração de epiglote
transdutor de pressão
transdutor em ponta
transfusão de granulócitos
transfusão de hemácias
transfusão de plaquetas
transmissão de doença
transmissão de sopros
transmissão de vírus
transmissão oro oral
transporte de pacientes
transtorno afetivo bipolar
transtorno de ansiedade
transtorno de conduta
transtorno de estresse
transtorno de humor
transtorno obsessivo compulsivo
transtornos de alimentação
transtornos de comportamento
transtornos de espectro
transtornos de pânico
transtornos de sono
transtorno de personalidade
tratamento com NOI
tratamento de asma
tratamento de bva
tratamento de casos
tratamento de constipação
tratamento de criança
tratamento de crise
tratamento de doença
tratamento de dor
tratamento de dsr
tratamento de enxaqueca
tratamento de epilepsia
tratamento de escolha
tratamento de fae
tratamento de fc
tratamento de fibrose
tratamento de fissuras
tratamento de hic
tratamento de hipertensão
tratamento de ic
tratamento de infecções
tratamento de linfangiomas
tratamento de mastite
tratamento de neoplasias
tratamento de obesidade
tratamento de osteoporose
tratamento de otite
tratamento de paciente
tratamento de pneumonias
tratamento de primeira
tratamento de psicoses
tratamento de resfriado
tratamento de rge
tratamento de rn
tratamento de sdr
tratamento de sdra
tratamento de sepse
tratamento de sinusite
tratamento de st
tratamento de tdah
tratamento de toc
tratamento de transtorno
tratamento do helicobacter
tratamento farmacológico específico
tratamento intensivo pediátrico
trato digestivo superior
trato digestório proximal
trato respiratório superior
trato urinário inferior
trato urinário superior
trauma de crânio
treinamento em am
tremores de frio
triagem auditiva neonatal
triagem metabólica neonatal
triagem neonatal universal
troca de gases
tromboplastina parcial ativada
tronco cerebral alterados
tronco cerebral normal
tubo de ventilação
tumor de wilms
tumores intra oculares
turno de manhã
ucm pós operatória
última relação sexual
último trimestre gestacional
ultra sonografia abdominal
ultra sonografia cerebral
ultra sonográfica sistemática
umidade de ar
umidificação de ambiente
unidade de saúde
unidades de cuidados
unidades de desempenho
unidades de internação
unidades de rede
unidades de terapia
unidades de tratamento
urina não centrifugada
usc neonatal anormal
uso de aas
uso de ácido
uso de água
uso de álcool
uso de alimentos
uso de am
uso de aminoglicosídeos
uso de analgesia
uso de anestésicos
uso de antibióticos
uso de antibioticoterapia
uso de antiinflamatório
uso de antimicrobianos
uso de antipsicóticos
uso de antitérmicos
uso de bicos
uso de chá
uso de chupeta
uso de clonidina
uso de corticóide
uso de corticosteróide
uso de corticoterapia
uso de cpap
uso de cremes
uso de dieta
uso de diuréticos
uso de dopamina
uso de drogas
uso de enzimas
uso de esteróides
uso de estimulantes
uso de fármacos
uso de fenoterol
uso de fórmula
uso de fumo
uso de imunoglobulina
uso de indometacina
uso de insulina
uso de isrs
uso de ketamina
uso de leite
uso de mamadeiras
uso de manitol
uso de material
uso de medicação
uso de medicamentos
uso de metilfenidato
uso de métodos
uso de nutrição
uso de opióides
uso de óxido
uso de oxigênio
uso de oxigenoterapia
uso de peep
uso de placebo
uso de preservativo
uso de prongas
uso de quimioprofilaxia
uso de quimioterapia
uso de rxt
uso de solução
uso de sri
uso de suplemento
uso de surfactante
uso de terapia
uso de teratógenos
uso de tocolíticos
uso de tricíclicos
uso de vacina
uso de vasopressores
uso de vc
uso de ventilação
uso em crianças

Trigramas da lista de referência em ordem alfabética		Tabela 12 de 12
usuárias de chupeta	válvula de uretra	vias de administração
usuários de prontuário	vantagens de amamentação	vias de saída
utilização de antibióticos	vantagens de aleitamento	vício de seleção
utilização de bloqueadores	variabilidade de medidas	vida de bebê
utilização de corticóides	variação de prevalência	vida de criança
utilização de critérios	variações de fatores	vida de família
utilização de dado	variações de normalidade	vida de indivíduo
utilização de fármacos	variações de técnica	vida de mulher
utilização de fórmulas	variáveis de confusão	vida de pacientes
utilização de hidrocortisona	variáveis de controle	vida de pessoas
utilização de idp	variáveis de estudo	vida de prematuros
utilização de soluções	variáveis explanatórias qualitativas	vida extra uterina
utilização de surfactante	variedade de condições	vida intra uterina
utilização de teste	vasoconstrição hipóxica pulmonar	vida pós natal
utilização de via	vasodilatador pulmonar seletivo	vídeo eeg prolongado
utis de adultos	veia cava superior	viés de causalidade
vacina bcg id	veia jugular interna	viés de informação
vacina bcg pc	velocidade de crescimento	viés de observação
vacina contra hib	velocidade de hemossedimentação	viés de seleção
vacina contra influenza	velocidade de infusão	vigência de estados
vacina contra vhb	ventilação com pressão	vigência de infecção
vacinação contra hib	ventilação líquida parcial	vigência de tratamento
vacinação contra influenza	ventilação mecânica convencional	vírus de hepatites
vacinação contra tuberculose	ventilação mecânica invasiva	vírus de imunodeficiência
vacinação em rede	ventilação não invasiva	vírus de influenza
valor de cortisol	ventilação pulmonar mecânica	vírus de sarampo
valor preditivo negativo	vesícula biliar atrófica	vírus epstein barr
valor preditivo positivo	via aérea artificial	vírus por leite
valores de freqüências	via aérea definitiva	vírus sincicial respiratório
valores de peso	via aérea difícil	visualização de cordas
valores de escore	via de exposição	vítimas de trauma
valores de ib	via de transmissão	volume de leite
valores de hemoglobina	vias aéreas inferiores	volume sangüíneo cerebral
valores de normalidade	vias aéreas periféricas	vômitos com sangue
valores de referências	vias aéreas superiores	

B. Listas de Conceitos Extraídas

Todas as listas de conceitos extraídas de todos os *corpora* utilizados nessa tese (Seção 3.1) se encontram disponíveis em formato eletrônico no material anexo a essa tese.

A título de ilustração, e para eventuais análises, nesse anexo estão impressas somente as listas de 2.323 bigramas e 2.726 trigramas considerados conceitos do *corpus* de Pediatria, conforme descrito no Capítulo 5. Essas listas estão organizadas segundo os valores dos índices *tf-dcf* de cada um dos termos.

Bigramas do *corpus* de Pediatria em ordem de frequência *tf-dcf* — Tabela 1 de 8

Aleitamento materno	Doenças crônicas	tamanho amostral	Meta análises
recém nascido	intubação traqueal	Estudos epidemiológicos	célula T
leite materno	Hipertensão pulmonar	cordão umbilical	asma referida
idade gestacional	via oral	processo infeccioso	Nutrição parenteral
Ventilação mecânica	BCG ID	História familiar	paciente estudado
via aérea	evolução neurológica	infecção congênita	maior prevalência
Pressão arterial	Unidade Neonatal	Estudos clínicos	grupo etário
leite humano	Exames complementares	Suporte ventilatório	peptídeo C
Hipertensão arterial	Anemia falciforme	paciente adulto	Má oclusão
terapia intensiva	evento adverso	Fatores genéticos	bebê prematuro
Atividades físicas	mãe adolescente	fluxo sangüíneo	febre alta
período neonatal	tubo endotraqueal	primeira semana	diagnóstico precoce
massa óssea	Diagnóstico diferencial	disfunção oral	deposição pulmonar
Estado nutricional	Doença pulmonar	DSM IV	criança obesa
faixa etária	asma aguda	asfixia perinatal	via biliar
Alimento complementar	alergia alimentar	fibra alimentar	Doenças infecciosas
cicatriz renal	carga viral	doença respiratória	colesterol total
perímetro cefálico	Freqüência respiratória	sobrevida global	grupo supino
Exame físico	criança estudada	pressão diastólica	alergia respiratória
Crianças menores	escore clínico	endocardite infecciosa	secreção nasofaríngea
infecção urinária	aleitamento exclusivo	perda auditiva	leite fraco
População estudada	desconforto respiratório	hemorragia digestiva	diferença significante
Crianças maiores	baixo peso	refluxo gastroesofágico	composição corporal
Otite média	Ensaios clínicos	trato respiratório	medula óssea
paciente pediátrico	trato urinário	malformação congênita	Úlcera duodenal
Pressão intracraniana	população pediátrica	seio materno	reabsorção óssea
vídeo EEG	fórmula infantil	população geral	trato gastrintestinal
Baixa estatura	Exames laboratoriais	BCG PC	fase aguda
choque séptico	critério clínico	idade óssea	troca gasosa
Manifestações clínicas	disfunção miccional	teste cutâneo	secreção nasal
Quadro clínico	insuficiência renal	população adulta	Refluxo vesicoureteral
Amamentação exclusiva	equipe médica	condição clínica	idade corrigida
ventilação pulmonar	orelha média	intestino delgado	sepse grave
Ultra sonografia	doença cardiovascular	Débito cardíaco	exercício físico
lesão pulmonar	escolaridade materna	neuro hipófise	Efeitos adversos
obesidade infantil	Estudos controlados	X frágil	Pré termo
corticosteróide antenatal	otoemissão acústica	defeito congênito	infecção respiratória
insuficiência respiratória	cardiopatia congênita	grau I	triagem neonatal
pré natal	pós operatório	Modelos animais	peso corporal
Crianças amamentadas	mau prognóstico	livre demanda	consentimento livre
lesão cerebral	Critérios diagnósticos	resposta inflamatória	Medidas antropométricas
fibrose cística	Infecção viral	prática clínica	Anorexia nervosa
relaxamento muscular	Óxido nítrico	Evidências científicas	tempo prolongado
alta hospitalar	hemocultura positiva	método diagnóstico	deficiência auditiva

Bigramas do *corpus* de Pediatria em ordem de frequência *tf-dcf* — Tabela 2 de 8

história clínica	interação social	Anomalias congênitas	membrana timpânica
evolução clínica	gasto energético	Corpo estranho	Estratégias terapêuticas
surfactante exógeno	crise convulsiva	população brasileira	transplante cardíaco
infusão contínua	meia vida	população alvo	formação óssea
níveis séricos	transtorno depressivo	Ingurgitamento mamário	pressão inspiratória
hipertensão endocraniana	hábito alimentar	Parâmetros clínicos	cortisol sérico
perfusão cerebral	Qui quadrado	mucosa intestinal	auto anticorpos
Hipertensão intracraniana	prática pediátrica	Transtornos alimentares	Estado infeccioso
alimentação complementar	ácido fólico	hemorragia intracraniana	obstrução nasal
artéria pulmonar	diarréia aguda	peroxidação lipídica	palato mole
artrite idiopática	membrana hialina	disfunção ventricular	átrio esquerdo
sepse neonatal	diagnóstico clínico	Tratamento farmacológico	sistema cardiovascular
reação adversa	fissuras mamilares	resposta imunológica	doses baixas
derivados imidazolínicos	agente etiológico	caracteres sexuais	sopro cardíaco
Avaliação clínica	Staphylococcus aureus	Drenagem torácica	resfriado comum
células progenitoras	congestão pulmonar	resistência vascular	USC neonatal
Biópsia hepática	Procedimentos invasivos	ventrículo esquerdo	principais efeitos
reação inflamatória	resistência intermediária	pólo superior	parênquima renal
efeito protetor	doença alérgica	linguagem oral	criança autista
Agentes infecciosos	distensão abdominal	criança febril	tique motor
bom prognóstico	sexto mês	síndrome genética	depressão maior
adulto jovem	trombose venosa	Doença falciforme	regressão logística
processo inflamatório	Risco relativo	transtorno psiquiátrico	infecção hospitalar
sinais clínicos	ambiente hospitalar	Casos graves	grupo poliarticular
necrose tumoral	parto vaginal	átrio direito	criança brasileira
maior gravidade	acompanhamento ambulatorial	espinhas dendríticas	refluxo gastresofágico
alimentação infantil	fórmula láctea	pediatra geral	grupo estudado
Exame endoscópico	mineralização óssea	injúrias físicas	evento paroxístico
cirurgia cardíaca	DAC prematura	aerossol dosimetrado	CPAP nasal
criança normal	colostro materno	grande risco	literatura médica
bulimia nervosa	Mecanismo imunológico	pronto Atendimento	Ventilação líquida
pré escolar	Métodos Este	Leucemia aguda	Cálculo amostral
coluna lombar	dermatite atópica	Aspectos éticos	sorologia positiva
Paralisia cerebral	enterocolite necrosante	infecção aguda	prednisona oral
primeira consulta	mães adultas	perímetro braquial	Consentimento informado
soro fisiológico	hipersensibilidade imediata	vacina conjugada	esquema terapêutico
grau III	leite artificial	peso fecal	espinha bífida
exame neurológico	etiologia viral	quadro infeccioso	bilirrubina total
Anti histamínicos	IgE sérica	estudo piloto	conteúdo mineral
retardo mental	dose recomendada	grupo precoce	IgE total
Baixo débito	resultado negativo	alças intestinais	mediador inflamatório
dor recorrente	ar ambiente	máscara facial	protocolo BFM
dor abdominal	hábitos orais	região subglótica	alterações bioquímicas
Soluções salinas	tronco cerebral	tubo traqueal	posição ortostática
desenvolvimento cognitivo	Suporte nutricional	vesícula biliar	grupo oligoarticular
densidade mineral	exame cultural	doença viral	doença localizada
valor preditivo	uso prolongado	crescimento fetal	IMC igual
situação clínica	TMO autogênico	gênero masculino	doença inflamatória
doença celíaca	ingestão alimentar	vida adulta	pneumonia bacteriana
grupo NEB	vacina BCG	diferença estatística	características clínicas
capacidade funcional	depressão respiratória	doença metastática	pacientes clínicos
doença invasiva	Medidas preventivas	doença arterial	líquido amniótico
fator prognóstico	ACTH sintético	gordura corporal	aspecto clínico
carga eletrostática	TCE grave	criança ostomizada	função renal
sistema imunológico	falência respiratória	Ventiladores mecânicos	achados endoscópicos
baixa idade	saúde pública	leite industrializado	Força Tarefa
uretra anterior	fibra solúvel	doença avançada	bicos artificiais
UTI neonatais	constipação crônica	colestase crônica	aleitamento natural
perfil lipídico	hepatite B	aleitamento predominante	RN prematuro
dobra cutânea	diabetes melito	remissão completa	capacidade residual
intubação endotraqueal	freqüência cardíaca	função pulmonar	haste hipofisária
risco intermediário	triagem auditiva	fêmur proximal	política pública
condição socioeconômica	IgE específica	asma persistente	distúrbio respiratório
Estudos randomizados	alojamento conjunto	ácido acetilsalicílico	Má absorção
níveis socioeconômicos	boa evolução	prega cutânea	insuficiência hepática
doença mental	alteração auditiva	sibilância prévia	dificuldade respiratória
ossos longos	modalidade terapêutica	trabalho materno	desconforto alto
complacência pulmonar	fosfatase alcalina	atendimento médico	pesquisador responsável
edema cerebral	apresentação clínica	intercorrências clínicas	Correlações significativas
Aleitamento misto	maior suscetibilidade	última menstruação	dieta normal
Volume corrente	extubação acidental	inalador dosimetrado	parâmetros ventilatórios
Constipação intestinal	mucosa gástrica	Estudo longitudinal	Evidências clínicas
enurese noturna	risco básico	doença atópica	terceiro dia
deficiência mental	grupo SO	maturação sexual	doença grave
TMO alogênico	Idade materna	grau II	auto estima
pressão sistólica	altas doses	equipe multidisciplinar	doença coronariana
infecção bacteriana	criança internada	estenose pulmonar	mecanismo fisiopatológico
morbidade respiratória	estudo transversal	canais arteriais	cânula traqueal
Contra indicações	menor escolaridade	ácido valpróico	artrite reumatóide
cavidade oral	estudo multicêntrico	tratamento medicamentoso	aleitamento artificial

Bigramas do *corpus* de Pediatria em ordem de frequência *tf-dcf* — Tabela 3 de 8

dose inicial	pacientes hospitalizados	DMSA inicial	lesão cardíaca
desenvolvimento infantil	perda óssea	ambiente escolar	aconselhamento genético
recusa alimentar	recurso terapêutico	tecido adiposo	grau leve
Abordagem terapêutica	Endoscopia digestiva	volume menor	hipertensão sistólica
hipertensão essencial	hepatite aguda	UTIs pediátricas	cavidade amniótica
transplante autogênico	síndromes dismórficas	grupo VMC	RGE fisiológico
sódio sérico	adolescente obeso	cortisol basal	ecocardiograma transtorácico
imunidade celular	histologia normal	desenvolvimento neuropsicomotor	corticóide oral
hipotensão arterial	retorno venoso	aparelho respiratório	dano renal
desnutrição grave	TNF a	idade cronológica	densidade óssea
massa corpórea	estudo brasileiro	segmento renal	forças mecânicas
displasia broncopulmonar	Descongestionantes tópicos	posição supina	endotélio vascular
Grupos experimentais	fator protetor	cordas vocais	dado epidemiológico
falha terapêutica	peso normal	efeito sedativo	associação estatística
avaliação nutricional	instrumento utilizado	Manifestações sistêmicas	indústria farmacêutica
alteração neurológica	ajuda prática	hipóxia tecidual	mucosa esofágica
dor torácica	estudo nacional	transtornos ansiosos	transplante hepático
região dependente	médico assistente	dieta enteral	alta hospitalar
déficit cognitivo	presente casuística	rendimento escolar	doença pneumocócica
avaliação neurológica	grupo controle	resposta clínica	pós parto
vacina antipneumocócica	cardiomiopatia dilatada	doença renal	índice ponderal
Gasometrias arteriais	Assistência ventilatória	espectro autista	AIJ sistêmica
habilidades sociais	interação medicamentosa	grupo tratado	cintilografia óssea
mãe filho	veia porta	seguintes sintomas	força muscular
padrões motores	borda hepática	gêmeos monozigóticos	grupo normal
Coagulação intravascular	exame citológico	ITU febril	níveis intermediários
ato cirúrgico	tratamento clínico	baixa dose	salário mínimo
parede torácica	condições socioeconômicas	Abscesso mamário	ouvido médio
estado geral	pacientes colestáticos	pressão venosa	metileno tetraidrofolato
cateterização uretral	comportamento repetitivo	Ordenha mamária	mulher mãe
ventilação convencional	caso suspeito	lábio superior	Monitorização prolongada
alterações hemodinâmicas	parto normal	causa orgânica	instrução materna
Transtorno bipolar	mortalidade neonatal	Transtorno afetivo	causalidade reversa
metabolismo cerebrais	pacientes cirúrgicos	indivíduo autista	remissão clínica
leite ordenhado	quinto minuto	dose terapêutica	leucometria inicial
antidepressivo tricíclico	temperatura axilar	Tc DMSA	características maternas
bloqueador neuromuscular	Características demográficas	doença renovascular	massa gorda
curso convencional	comunicação interventricular	dose máxima	coluna vertebral
ente querido	assistência intensiva	Máscara laríngea	primeiro minuto
bloqueio neuromuscular	densitometria óssea	borda esternal	recém nato
sintoma alvo	lesão moderada	limiar convulsivo	hospitais universitários
esclerose tuberosa	tratamento cirúrgico	ventrículo direito	doença reumática
RM grave	veia cava	álcool fetal	recrutamento alveolar
renda familiar	infecção grave	relação V	presente série
pressão positiva	lavado broncoalveolar	tipo tensional	ciclo respiratório
membro inferior	casos refratários	expansão intravascular	colite alérgica
Sexo masculino	Doenças agudas	sexo feminino	hormônio sexual
amostra estudada	escore executivo	baixa renda	população estudada
atresia biliar	queixa principal	via inalatória	grupo tradicional
causas externas	suplementos hipercalóricos	Significância estatística	Avaliação oftalmológica
doença cardíaca	investigação clínica	literatura internacional	puérpera submetida
espaçador artesanal	comportamento alimentar	Exames clínicos	hepatopatia crônica
tratamento adequado	variáveis relacionadas	estudo populacional	padrão alimentar
diagnóstico definitivo	prática diária	crises epilépticas	sintomas iniciais
Leucemia Mielóide	desenvolvimento neurológico	parasitose intestinal	Estudos retrospectivos
herança autossômica	sucção digital	teste diagnóstico	imunofluorescência indireta
escola médica	Respiração oral	mortalidade infantil	amamentação predominante
Metabolismo ósseo	relações afetivas	RM isolado	esforço respiratório
isquemia cerebral	ingestão calórica	conduta terapêutica	desnutrição aguda
risco nutricional	grupo ambulatório	desenvolvimento motor	avaliação neuropsicológica
tubo neural	oxigênio suplementar	necessidades nutricionais	espécie humana
dispositivo inalatório	habilidade cognitiva	Abordagem diagnóstica	diagnóstico etiológico
emissões otoacústicas	respiração espontânea	prática médica	pressão expiratória
princípio ativo	Intervenções terapêuticas	criança saudável	questionário padronizado
níveis elevados	bexiga neurogênica	achados clínicos	nascimento igual
teste sorológico	estenose aórtica	amostra fecal	relacionamento social
grupo total	uretrocistografia miccional	esquema vacinal	Diferenças metodológicas
trauma mamilar	menor mortalidade	hidrolisado protéico	cobertura vacinal
lesões glomerulares	sintoma urinário	medidas repetidas	grupo prednisona
esofagite erosiva	Acidose metabólica	alterações histológicas	exames subsidiários
infecções pneumocócicas	Mímica facial	lesão glomerular	forma aguda
lipodistrofia generalizada	revisão sistemática	condrodisplasia puntiforme	dinâmica familiar
uso pediátrico	risco aumentado	irradiância espectral	IMC elevado
caso relatado	transplante alogênico	realidade brasileira	maior escolaridade
capitais brasileiras	pré operatório	Somente paciente	escore verbal
grupo tardio	saúde mental	programa SPSS	intervenção tradicional
gênero feminino	recomendação anterior	importância clínica	Lesões graves
pustulose palmoplantar	estudo genético	teste negativo	sétimo dia
divertículo uretral	perda urinária	emergência pediátrica	efeito terapêutico

Bigramas do *corpus* de Pediatria em ordem de frequência *tf-dcf* Tabela 4 de 8

ambiente familiar	terror noturno	transmissão perinatal	via nasal
enzimas pancreáticas	Movimentos anormais	alterações leves	possibilidade diagnóstica
Pesquisa Nacional	comportamentos automutilantes	EDA normal	dor intensa
níveis plasmáticos	primoinfecção urinária	hipertrofia muscular	investigação complementar
Vírus respiratórios	depressão infantil	problemas metodológicos	absenteísmo escolar
relação familiar	doença hereditária	fluxo aéreo	parada cardiorrespiratória
doses menores	distúrbios metabólicos	Formas adquiridas	Antecedentes familiares
resposta terapêutica	Pré carga	desenvolvimento cerebral	hábitos saudáveis
comprometimento hepático	zumbido venoso	atividade clínica	influência genética
relação sexual	hipermobilidade articular	tumor primário	pulmão direito
pacientes fibrocísticos	eliminação renal	risco importante	uso contínuo
uso tópico	menor volume	indivíduos adultos	Medicações utilizadas
sangue materno	período estudado	antígeno polissacarídico	classe socioeconômica
prática desportiva	primeiro mês	SAM associada	infecção pulmonar
sintomas diurnos	escola pública	UNIFESP EPM	terceira dose
orientação nutricional	estudo publicado	escolares normais	segundo ano
consulta pediátrica	procedimento cirúrgico	exame histológico	paciente grave
Bronquiolite viral	Insuficiência Cardíaca	mecânica respiratória	curso clínico
doença neurológica	efeito positivo	triagem universal	diagnóstico incorreto
vida saudável	consulta médica	trato digestivo	intervenção cirúrgica
ingestão energética	alto risco	anestesia geral	população jovem
efeito cumulativo	efeito benéfico	proteína S	envolvimento hepático
múltiplos órgãos	efeito colateral	reabilitação vestibular	dose diária
causa respiratória	própria doença	leitos neonatais	terceiro ano
indicação cirúrgica	surfactante pulmonar	síndrome epiléptica	crianças prematuras
imaturidade pulmonar	atendimento ambulatorial	faixa pediátrica	gordura corpórea
Crescimento somático	pediatra brasileiro	prontuário médico	prática alimentar
DECH crônica	classe social	primeira coleta	grande morbidade
Desenvolvimento normal	Orientação dietética	diversas regiões	primeira infância
resolução espontânea	concentração inibitória	HSL PUCRS	Evidências epidemiológicas
doença neuromuscular	de tratamento	estudo internacional	índice P
exame radiológico	cepas resistentes	índice cardiotorácico	função cortical
via intradérmica	isolamento social	tórax inicial	pacientes avaliados
teste tuberculínico	crianças acometidas	grupo sobrevivente	dose alta
RCHT positiva	crianças constipadas	espessamento brônquico	posição ereta
Vacinas antipneumocócicas	estudo observacional	lesão isquêmica	experiência prévia
quadro respiratório	espaço mandibular	quarto mês	história alimentar
acesso venoso	dados clínicos	Rinofaringite aguda	evento traumático
volume intravascular	associação significativa	trabalho respiratório	volume pulmonar
remodelação óssea	diferenças étnicas	Correção cirúrgica	tempo inspiratório
osso trabecular	infecção pneumocócica	cepas isoladas	Novas terapias
dose utilizada	tratamento convencional	endoscopia normal	fibrose pulmonar
arritmia cardíaca	resistência antimicrobiana	bebês chiadores	infecções recorrentes
má nutrição	grupo étnico	paciente internado	mordida aberta
atenção primária	idade pediátrica	mau controle	parâmetro fisiológico
células endoteliais	resistência bacteriana	Avaliação funcional	modo positivo
acidente vascular	maior idade	achados radiográficos	última relação
condições sociais	umidade fecal	admissão hospitalar	problemas comportamentais
soluções hipertônicas	peso adequado	população avaliada	informação materna
orientação preventiva	fluxo biliar	rigidez mandibular	possíveis complicações
sinais vitais	Grã Bretanha	Relaxantes musculares	fisioterapia respiratória
tratamento endoscópico	Testes laboratoriais	níveis pressóricos	termo sadio
Maior concordância	déficit nutricional	tumor ósseo	pontuação global
cárie dentária	laringoscopia direta	endoscopia respiratória	esvaziamento gástrico
dose única	HFA DPB	escore maior	insuficiência pancreática
urografia excretora	Marcus Gunn	estudo cromossômico	violência doméstica
baixa especificidade	Publicações recentes	mães submetidas	artrite séptica
rinite alérgica	quadro neurológico	episódio agudo	pessoas obesas
Investigação laboratorial	esofagite eosinofílica	pacientes alérgicos	gestantes estudadas
leite bovino	centros especializados	população infantil	células epiteliais
procedimentos dolorosos	Paciente descrito	derivado nitroimidazólico	flora intestinal
maus tratos	seguimento ambulatorial	Dor noturna	sangramento digestivo
psicoses infantis	Infecção materna	falsa anorexia	ação rápida
lobo temporal	técnica padronizada	envolvimento pulmonar	sexta semana
estudo aberto	resistência absoluta	obstrução intestinal	paciente oncológico
USC normal	HUPES CPPHO	tipo coorte	doença oncológica
hipertrofia ventricular	dado interessante	hipertensão materna	criança doente
pressão sistêmica	valores basais	tecido hipofisário	doenças mitocondriais
Débito urinário	puerpério imediato	ganho ponderal	distúrbio ventilatório
agitação psicomotora	banana verde	complicações associadas	processamento auditivo
ritmo circadiano	alimentação artificial	acompanhamento médico	estresse cirúrgico
traumas mamilares	obstrução infravesical	difícil controle	confirmação diagnóstica
sangue periférico	abertura traqueal	hipercapnia permissiva	avaliação urodinâmica
célula alveolar	hospitais públicos	alta morbidade	alteração clínica
imprinting metabólico	limites propostos	aleitamento continuado	fácil execução
Alfa MSH	corticóide inalado	primeira internação	região cervical
ductos lactíferos	ensaio imunoenzimático	população específica	pacientes intubados
punção suprapúbica	necrose parietal	acometimento hepático	recuperação nutricional
coluna líquida	Avaliação antropométrica	eventos cardiovasculares	dano pulmonar
disfunção orgânica	crianças Ikpeng	medicação sedativa	agressão física

Bigramas do *corpus* de Pediatria em ordem de frequência *tf-dcf* — Tabela 5 de 8

medida analgésica	MG anormais	pneumonia comunitária	incapacitação funcional
drogas vasoativas	giro temporal	screening neonatal	bilirrubinemia total
pais obesos	sistema surfactante	diversos órgãos	agentes teratogênicos
aspectos nutricionais	membrana celular	receptores adrenérgicos	esteróides inalados
alterações imunológicas	faringite aguda	crise asmática	corticóides sistêmicos
pós TMO	anticorpos específicos	Prezado Editor	lactentes sibilantes
DECH aguda	casos selecionados	habilidades corporais	diagnóstico tardio
doença péptica	atividade esportiva	pacientes relatados	Drogas usadas
ansiedade generalizada	procedimento diagnóstico	aquisição comunitária	ressonância nuclear
curvas preditas	Temperatura corporal	pronto socorro	fatores socioculturais
avaliação física	crianças afetadas	principalmente crianças	ácido fitânico
caixa torácica	Hipotermia moderada	colangiografia transoperatória	fluticasona HFA
crianças americanas	Craniotomia descompressiva	comunidade pobre	Calcificações puntiformes
complicações perinatais	segunda semana	elevada taxa	pálpebra superior
incontinência urinária	saco coletor	níveis educacionais	nervo oculomotor
oferta hídrica	secreções respiratórias	ossos pequenos	suplementação vitamínica
leite maduro	programa educacional	ossos grandes	orientação alimentar
seios paranasais	barreira hematoencefálica	Alcalose metabólica	melhor deposição
problemas perinatais	sonda nasogástrica	controles normais	processo anabólico
Estenose subglótica	Tratamento geral	soro glicosado	células adiposas
imunidade humoral	Sinusite aguda	comunidade pediátrica	cardiomiopatia hipertrófica
início abrupto	mortalidade geral	intervenção educativa	método de
grau IV	recomendações específicas	Problemas neurológicos	fundoplicatura anterior
mecânica pulmonar	Compressas mornas	ativação imunológica	cálcio total
célula muscular	ressuscitação volumétrica	DNA bacteriano	vacina recombinante
drogas sedativas	lesão secundária	baixo rendimento	pequenos pacientes
sonda endotraqueal	produção láctea	adaptação cultural	malformação cardíaca
deficiência visual	própria vida	infecção crônica	maior fermentação
médico entrevistado	anorexia infantil	controle sadio	infecção neonatal
decúbito dorsal	anemia hemolítica	Moraxella catarrhalis	doença bacterêmica
drogas utilizadas	pré oxigenação	pneumonia pneumocócica	duplo placebo
proliferação ductal	sociedade civil	serviço social	resposta parcial
ducto hepático	volume cerebral	dificuldade diagnóstica	desenvolvimento físico
ducto biliar	substância branca	clínica pediátrica	função respiratória
Cordões triangulares	obstrutiva crônica	agente causal	doença diarréica
terapêutica inicial	canal familiar	bases clínicas	diluições selecionadas
desenvolvimento emocional	terapia antimicrobiana	hidrocefalias congênitas	diluições decimais
alterações metabólicas	perda celular	ressonância magnética	líquido pleural
início precoce	curso curto	evidência II	quadro diarréico
atividade muscular	sintomas comportamentais	doença vascular	lesões líticas
alimentação enteral	rituais religiosos	duplo cego	câmara hiperbárica
disfunção cerebral	Pós carga	ressuscitação cardiopulmonar	primeiras mamadas
altura final	Triagem metabólica	hiperbilirrubinemia indireta	dupla mãe
palato duro	herança genética	elevada prevalência	vacina pneumocócica
recrutamento pulmonar	antipsicóticos atípicos	profissionais treinados	maior precocidade
primeira dose	válvula mitral	baixa condição	distensão vesical
crianças nascidas	ramos pulmonares	amostras genotipadas	dilatação uretral
massa muscular	campos pulmonares	pacientes infectados	derivação urinária
musculatura respiratória	inflexão inferior	padrão ouro	altos volumes
função cognitiva	infecção secundária	obstrução biliar	melhor desenvolvimento
surfactante natural	via enteral	ultra som	íleo meconial
Sintomas gastrointestinais	Epilepsia mioclônica	desfecho clínico	Lesões ósseas
função surfactante	defeito estrutural	diversas publicações	partes moles
alimentos saudáveis	Malformações cerebrais	alta prevalência	sobrepeso masculino
coto ureteral	circulação pulmonar	menor atividade	luz ultravioleta
shunt intrapulmonar	crise dolorosa	método sorológico	secreção purulenta
manifestações iniciais	vasoconstrição hipóxica	parada cardíaca	melhor sensibilidade
vírus sincicial	vasculatura pulmonar	regiões pendentes	crescimento bacteriano
suporte familiar	pulmão normal	gasto calórico	alteração macroscópica
menor dose	presente estudo	metabolismo lipídico	efeitos clínicos
estímulo fóbico	Alto Xingu	pediatra treinado	coluna total
necessidades especiais	Recente estudo	face humana	rações experimentais
necessidade calórica	Pseudomonas aeruginosa	escola privada	unidades pediátricas
função esplênica	pressão pulmonar	fibra insolúvel	leitos intensivos
novas cicatrizes	massa corporal	incisivos superiores	drogas ototóxicas
RVU primário	associação positiva	serviço privado	população Ikpeng
funcionamento oral	idade escolar	malformação anorretal	indicadores P
Exercícios orofaciais	metade de	gastrite antral	unidades alveolares
ventilação espontânea	período prolongado	tecido adjacente	episódios bulímicos
circulação extracorpórea	embasamento científico	corpo estriado	hidrocefalias isoladas
quadro delirante	fator ambiental	Abuso físico	diagnóstico prévio
autismo infantil	possíveis associações	regressão total	fácil administração
mucosa nasal	Biologia Molecular	doença metabólica	ponto doloroso
doença orgânica	lobo frontal	tratamento específico	hiper responsividade
reação grave	Grupo I	aparelho locomotor	lesões cutâneas
reação alérgica	internação hospitalar	doença bacteriana	vacina polissacarídica
alterações cardiovasculares	rede NEOCOSUR	triagem inicial	cirurgia conservadora
pós termo	referido estudo	hipertensão diastólica	maior especificidade
movimentos irregulares	helmintíases intestinais	alto fluxo	ativação macrofágica
crânio neonatal	conhecimento atual	ponte nasal	pneumonia grave

Bigramas do *corpus* de Pediatria em ordem de frequência *tf-dcf* Tabela 6 de 8

inspiração profunda	corticóide sistêmico	nascimento prematuro	pacientes críticos
tratamento dietético	queixa clínica	extratos sociais	internação prolongada
IG menor	período crítico	alérgenos testados	etnia e
atual estudo	acuidade visual	RAST negativo	UTI pediátrica
corrente sangüínea	diagnóstico final	expressão verbal	atividade sexual
predisposição genética	parede celular	parênquima pulmonar	Orientação antecipatória
doença rara	gastrite crônica	edema pulmonar	hipersensibilidade tardia
imagem corporal	estenose pilórica	morfometria computadorizada	último trimestre
hepatites virais	parede abdominal	via sistêmica	criança desnutrida
fluxo expiratório	estratégia protetora	NRS III	bronquiolite aguda
líquido cefalorraquidiano	ciclo menstrual	tecidos moles	atuação exclusiva
transmissão vertical	escores z	cápsula polissacarídica	toxoplasmose congênita
serviço especializado	consulta agendada	Suporte respiratório	idade adulta
sibilância desencadeada	atrofia vilositária	paciente asmático	diagnóstico estabelecido
olhos abertos	forma inadequada	pior evolução	gordura subcutânea
criança vestibulopata	novas modalidades	tabagismo materno	terceira semana
nível hidroaéreo	sela vazia	disfunção miocárdica	resposta favorável
evento clínico	sela túrcica	postura anormal	dose cumulativa
capacidade física	hipotálamo hipofisária	pacientes adolescentes	músculos respiratórios
trauma local	hipoplasia hipofisária	primeiro atendimento	Avaliação laboratorial
expressão facial	adeno hipófise	tratamento tradicional	déficit auditivo
síndrome caracterizada	análise ajustada	idade posterior	suscetibilidade genética
cateterismo umbilical	Hemograma completo	ventilação alveolar	lesão renal
função hepática	ciclo circadiano	borda inferior	MOV FHEMIG
gordura saturada	Doses tóxicas	comparações múltiplas	problemas emocionais
Métodos Foram	amostra casual	aparelho gastrintestinal	proteínas plasmáticas
parâmetro estudado	base populacional	avaliação imunológica	idade precoce
recursos tecnológicos	doses variáveis	nível terciário	osmolalidade intracelular
qualidade técnica	situação conjugal	período pubertário	lesão térmica
uretra posterior	segunda consulta	níveis maturacionais	espaço extracelular
radiologista pediátrico	formação continuada	respiração nasal	fator materno
solução colóide	equipe assistencial	arcada dentária	política estatal
adolescentes masculinos	seguimento clínico	alterações fonoarticulatórias	análise compreensiva
sintoma presente	sintomatologia clínica	própria criança	via percutânea
e mail	eczema atópico	retardo psicomotor	vacina percutânea
alterações hepáticas	ganho ponderal	adolescentes pesquisados	linfócitos citotóxicos
treinamento específico	grande controvérsia	área cerebral	imunidade protetora
problemas psicológicos	somente leite	crianças usuárias	cepa utilizada
Cf III	resposta imune	resultado normal	quadro grave
estudo histológico	variáveis estudadas	Estudos prospectivos	punção venosa
diagnóstico endoscópico	variáveis maternas	tratamento inicial	recorrência familiar
tratamento empírico	saúde infantil	controle hemodinâmico	surfatante exógeno
dimorfismo sexual	assistência hospitalar	manifestações cutâneas	SNAPPE II
supervisão médica	escore total	Métodos Foi	aspecto cognitivo
expressão manual	percepção materna	distribuição universal	regime terapêutico
presente amostra	nutrição infantil	germe estudado	mama puerperal
achados histológicos	processo diagnóstico	diagnóstico específico	intercorrência respiratória
baixa escolaridade	intensidade variável	plasma materno	peso atual
pacientes osteopênicos	pacientes incluídos	vitaminas lipossolúveis	assistência neonatal
insulina regular	vida sexual	esquema antimicrobiano	punção lombar
câmara cardíaca	achado comum	suplementação oral	ventilação invasiva
atual pesquisa	antecedente mórbido	vida diária	máscara nasal
medula espinhal	Estudos etnográficos	menor lesão	permanência hospitalar
fator socioeconômico	indicador antropométrico	grupo social	faixa normal
reflexo vermelho	próprio indivíduo	cromossomo X	fraturas vertebrais
crianças alérgicas	atendimento primário	autoridade sanitária	início agudo
diversos alérgenos	diátese hemorrágica	origem multifatorial	acuidade auditiva
consentimento esclarecido	teste Kappa	tratamento continuado	atividades habituais
gastrite alérgica	conseqüências clínicas	asma brônquica	maior predisposição
crianças soropositivas	hemorragia intraventricular	orelha direita	cuidados convencionais
displasia óssea	diarréia persistente	hipoacusia condutiva	veias pulmonares
trato digestório	imunodeficiência humana	audição normal	lábio inferior
enzimas hepáticas	alimentação adequada	principais sinais	reação local
lavado nasal	corticóide inalatório	plexo coróide	problema clínico
assistência médica	tratamento antimicrobiano	orelha interna	controle esfincteriano
alterações radiográficas	relatos iniciais	melhor tratamento	porta hepatis
via vaginal	áreas endêmicas	fenda palatina	intra hepática
escolares adolescentes	leishmaniose visceral	paciente enurético	Tubagem duodenal
população local	crianças brancas	enurese polissintomática	comportamento social
menor calibre	taquicardia ventricular	complexo esfincteriano	córtex visual
alterações morfológicas	doença meningocócica	anamnese dirigida	intervenção precoce
significância menor	alterações cognitivas	metodologias diferentes	estimulação imunológica
grupo pediátrico	prejuízo funcional	criança assintomática	categoria imunológica
hiperemia conjuntival	Orientação familiar	asma moderada	problemas respiratórios
obstrução alta	Ventilação assistida	ductos biliares	via endoscópica
achados obtidos	pega correta	crianças infectadas	necessidades energéticas
doença aterosclerótica	absorção intestinal	forma fidedigna	evento fisiopatológico
comunidade estudada	desnutrição leve	índios Pima	período avaliado
região lombar	Aconselhamento nutricional	fator estressor	material escrito
cesárea eletiva	albumina sérica	avanços importantes	alterações radiológicas

Bigramas do *corpus* de Pediatria em ordem de frequência *tf-dcf* — Tabela 7 de 8

subseqüentes comparações	único estudo	último efeito	alimentos excluídos
estado basal	efeito rebote	função ventricular	verdadeira densidade
PEEP alta	contraceptivos orais	Distúrbios hemorrágicos	patologias crônicas
terapêutica convencional	veia jugular	retirada gradual	detector esofágico
sulfato ferroso	consultório pediátrico	epilepsias generalizadas	cepas sensíveis
raio X	Enterococcus faecalis	crises mioclônicas	anoxia perinatal
situações especiais	dano tecidual	Drogas antiepilépticas	cultura positiva
pesquisa clínica	técnicas moleculares	novo alimento	parto cesariano
choro inconsolável	retraimento social	densidade energética	único fato
hábito intestinal	osmolaridade sérica	corpo caloso	critérios uniformes
analgesia sistêmica	osmolaridade plasmática	vaso sangüíneo	manifestações gastrintestinais
rim contralateral	Drenagem liquórica	dor crônica	pezinho ampliado
quimioprofilaxia antibiótica	tratamento comportamental	falência cardíaca	anedotário folclórico
pólo inferior	leite produzido	dores difusas	trabalhos epidemiológicos
inferior ipsilateral	não despolarizante	doença específica	problemas infecciosos
endoscopia terapêutica	demanda metabólica	Considerações gerais	bactérias viáveis
pielonefrite aguda	herpes simples	tumores intracranianos	TCH alogênico
RVU bilateral	HTLV I	desvio intracelular	PCR positiva
orelhas proeminentes	refluxo gastroesofageano	ventilação protetora	antígenos bacterianos
método invasivo	anorexia verdadeira	espaço morto	controles saudáveis
interesses restritos	funcionamento global	VC reduzido	escolas pediátricas
espaço intracelular	alimento consumido	Membros superiores	resistência plena
morbimortalidade infantil	transdutor externo	rede pública	Crianças alimentadas
dose habitual	pressão intraventricular	variáveis estudadas	instrumento genérico
fluxo salivar	tratamento apropriado	córtex cerebral	forma poliarticular
efeito analgésico	características anatômicas	ácidos graxos	sorotipos isolados
amamentação natural	Progressos consideráveis	Modelos experimentais	importante viés
tratamento proposto	viscosidade sangüínea	Sistema nervoso	origem fetal
exame inicial	potencial evocado	alta sensibilidade	incapacidade física
via endovenosa	reforço positivo	forte intensidade	fracasso terapêutico
situações sociais	acidose lática	primeira hora	parâmetro avaliado
intervenções familiares	efeito inotrópico	Grupo B	eficácia semelhante
peso menor	mielinólise pontina	menor renda	artigos publicados
etiologia bacteriana	PEEP adequada	sociedade brasileira	classe médica
regiões cerebrais	adulto autista	morte súbita	gastrite hemorrágica
indivíduos afetados	urina centrifugada	diagnóstico preciso	metodologia científica
síndrome torácica	ITU recorrente	própria paciente	sabedoria popular
infarto ósseo	ITU baixa	linguagem escrita	comportamentos maternos
hemoglobina S	queixa somática	luz solar	membrana basal
neuropeptídeo Y	depressão mascarada	Considerações finais	Oncologia Pediátrica
reflexos orais	crianças depressivas	primeiro ano	componente essencial
Sistema respiratório	Depressão anaclítica	variáveis socioeconômicas	adolescentes celíacos
distúrbio motor	sistema límbico	menor efeito	nível populacional
quadro psicótico	hipertensão secundária	atividade diária	tratamento eficaz
psicoses reativas	Comportamentos ritualísticos	Ensino Médio	artigo experimental
expansão volumétrica	curta ação	tecido ósseo	indicação clínica
estímulo doloroso	tratamento paliativo	Balanço Energético	idade fértil
necrólise epidérmica	não intervenção	Primeira avaliação	sofrimento materno
médico paciente	morte digna	Grupo II	sintomas apresentados
movimentos espontâneos	agente paralisante	Estudo recente	especialidade médica
leucomalácia periventricular	efeitos extrapiramidais	menor idade	síndrome metabólica
idade concepcional	volume controlado	avaliação inicial	origem bacteriana
USC anormais	sistema dopaminérgico	tomografia computadorizada	replicação viral
sinusite bacteriana	depressão miocárdica	países industrializados	déficit imunológico
resultados terapêuticos	Casos leves	país desenvolvido	cepas invasivas
antibióticos profiláticos	terapia convencional	receptores cardíacos	cirurgia infantil
problemas psiquiátricos	Crises parciais	capital mineral	Colangiografia operatória
abuso sexual	Suporte psicológico	internação prolongada	medicamento administrado
perfusão tecidual	dor persistente	transtornos paroxísticos	análise morfométrica
gordura visceral	sedativo ideal	possíveis asmáticos	qualidade metodológica
compressas frias	pacientes conscientes	resultados discordantes	ato operatório
superfície alveolar	lâmina reta	base explicativa	maior eficácia
jato médio	faringe posterior	melhores efeitos	fatores biológicos
medicina social	agente paralítico	estadiamento puberal	desempenho cardiovascular
leites modificados	Posicionamento adequado	problema epidêmico	dor lombar
única dose	crises hipertensivas	leitos hospitalares	Centro Nacional
resultados controversos	parede torácica	Achados radiológicos	doença hepática
ventilação adequada	contratilidade miocárdica	processos virais	alta especificidade
índice cardíaco	centro cirúrgico	quimioterapia convencional	principais quadros
complicações supurativas	dor moderada	artigo original	conclusão diagnóstica
tiques vocais	déficit neurológico	revista eletrônica	Centros brasileiros
tecido conjuntivo	septo atrioventricular	menino autista	violência urbana
complicações relacionadas	hipoplasia cerebelar	Sb clássica	consulta pública
má pega	hiperfluxo pulmonar	distúrbios psiquiátricos	instituição privada
tratamento profilático	grande shunt	situação endotraqueal	cólica verdadeira
entidade clínica	segunda bulha	terapêutica adequada	PA sistólica
tumores sólidos	miocardiopatia hipertrófica	pneumonia aguda	classe IgG
sinais aferentes	grandes artérias	líderes espirituais	função auditiva
gasto total	coração esquerdo	critério científico	recém nascido
IMC maior	cardiopatias adquiridas	prescrição médica	sorotipos presentes

Bigramas do *corpus* de Pediatria em ordem de frequência *tf-dcf* — Tabela 8 de 8

alteração significante	trajeto colônico	achado interessante	esofagite histológica
observação experimental	g ingerida	evocadas transitórias	malformações isoladas
fibra ideal	dietas enterais	trabalho médico	diagnóstico sindrômico
apenas fibra	cólon proximal	Serviços brasileiros	crianças anencéfalas
assunto controverso	suspeita diagnóstica	primeira pesquisa	achados de
doenças pneumocócicas	intestino primitivo	prevalência estimada	colônias típicas
cuidados neonatais	hidrocefalia severa	envelope fechado	carga microbiana
trabalho brasileiro	fístula entérica	ensino público	Standard Methods
saúde profissionais	Motilidade intestinal	adultos fumantes	Bactérias lácticas
política nacional	estímulos inflamatórios	adolescentes argentinos	dias alternados
avaliação epidemiológica	droga segura	estruturas adjacentes	dieta habitual
fato conhecido	análise descritiva	ácido linoleico	única ocasião
líquido duodenal	última aplicação	módulo eczema	escolas sorteadas
colestase neonatal	regressão espontânea	peso seco	valva protética
níveis iguais	diarréia crônica	importante problema	entidade heterogênea
fototerapia profilática	triagem populacional	aparência acromegálica	deiscência parcial
bilirrubina sérica	indicador perímetro	repercussão negativa	único trabalho
contagem leucocitária	interpretação radiológica	tempo único	sociedade ocidental
profissionais experientes	complicações esofágicas	medidas seriadas	vidro despolido
Crianças asmáticas	EDA pediátrica	pH esofágico	tumoração mandibular
causa importante	uso endovenoso	esofagite crônica	osteomielite crônica
futuras gestações	seqüelas emocionais	células ganglionares	osteomielite aguda
apenas leite	infecções gastrintestinais	Eletromanometria esofágica	inflamação óssea
estratégias vacinais	limites imprecisos	DMO reduzida	espessamento cortical
pós cirurgia	distrofia simpática	fenômenos biológicos	curso doloroso
parto cesárea	Alterações tróficas	pneumopatia crônica	Paciente feminina
asfixia neonatal	droga administrada	grandes metrópoles	hormônios tireoidianos
região metropolitana	budesonida inalatória	vigilância epidemiológica	referência internacional
investigação metabólica	Abordagem interessante	meningites bacterianas	leites infantis
valor normal	tratamento seco	impacto maior	insulina NPH
elevado nível	tecidos danificados	cuidados médicos	edema local
doença residual	plástico pequeno	forma suave	paredes espessas
programa terapêutico	lanolina anidra	espaçador plástico	jato miccional
hospitais gerais	importante complicação	resíduo gástrico	fulguração endoscópica
consenso internacional	importante causa	maternidades privadas	esforço miccional
medidas sanitárias	hidrocorticóide sintético	maiores complicações	derivação temporária
exame parasitológico	somente criança	alta morbimortalidade	causa desconhecida
esquema profilático	tio materno	motivo aparente	dose elevada
desenvolvimento mental	recaídas freqüentes	também benefícios	materiais estranhos
redes multicêntricas	proteinúria significativa	implantação baixa	brônquio principal
corticóides antenatais	biópsia renal	unidades mistas	broncoscópio rígido
confundimento residual	pacientes estáveis	imaturidade fisiológica	Pinça endoscópica
evolução favorável	múltiplos fatores	Hemorragia intraperiventricular	CE traqueobrônquico
tipo Iia	malformações somáticas	esofagite péptica	lado comprometido
anomalia cromossômica	malformações complexas	principalmente televisão	uso internacional
ajustes freqüentes	cirurgia retardada	ensino privado	quinto lugar
práticas assistenciais	ICr HCFMUSP	suscetibilidade antimicrobiana	mucosa esofagiana
sistema NADPH	uso diário	crianças indigentes	área óssea
pneumonia tratada	profissionais habilitados	bacteremia oculta	fragilidade óssea
imunização básica	mobilidade articular	Coombs direto	crianças sadias
deleção GT	lúpus neonatal	artrite crônica	quinto percentil
abscesso hepático	forma rizomélica	educação médica	médias encontradas
fatores culturais	corpos vertebrais	encaminhamentos excessivos	média total
enorme importância	CIV subaórtica	linha média	comprimento médio
invasão bacteriana	CFC DPB	mãe HIV	nascimentos hospitalares
achados perinatais	miastenia gravis	forma espontânea	medicina fetal
volume fecal	alteração oftalmológica	esôfago distal	

Trigramas do *corpus* de Pediatria em ordem de frequência *tf-dcf* — Tabela 1 de 13

uso de chupeta	vida de criança	dieta de exclusão
Fator de risco	maioria de pacientes	farelo de trigo
aleitamento materno exclusivo	trabalho de parto	resistência a penicilina
Comitê de Ética	tipo de parto	prática de amamentação
leite de vaca	polissacarídeo de soja	deficiência de ferro
peso de nascimento	início de sintoma	tempo de queixa
sistema nervoso central	Pacientes com SDRA	taxa de mortalidade
grupo de pacientes	via aérea superior	tempo de internação
ganho de peso	uso de medicamento	Comissão de Ética
Jornal de Pediatria	terapia intensiva neonatais	paciente de grupo
Critérios de inclusão	População de estudo	via aérea inferior
faixa etária pediátrica	Necessidade de ventilação	pressão arterial sistólica
Ministério da Saúde	uso de medicação	alimentação de criança
serviço de saúde	casca de banana	prevalência de obesidade
produção de leite	uso de droga	vacina contra influenza
profissionais de saúde	idade de criança	ventilação pulmonar mecânica
unidade de terapia	Velocidade de crescimento	redução de mortalidade

Trigramas do *corpus* de Pediatria em ordem de frequência *tf-dcf* — Tabela 2 de 13

excesso de peso	início de amamentação	promoção de aleitamento
termo de consentimento	prática de aleitamento	apoio a amamentação
crianças de sexo	elevador de pálpebra	atividade de doença
Uso de antibióticos	acidentes de transporte	prevenção de infecção
ventilação não invasiva	tratamento de fissuras	ensaio clínico randomizado
aquisição de linguagem	Pressão arterial elevada	prevenção de doença
Otite média aguda	taxas de soroproteção	tratamento de infecção
volume de leite	hipótese de nulidade	gravidade de asma
lesão pulmonar aguda	unidade de tratamento	crise de sibilância
tratamento de constipação	telerradiografia de tórax	velocidade de hemossedimentação
estudo de coorte	Protocolo de estudo	prevalência de doença
critério de exclusão	exame ultra sonográfico	Evidências de benefícios
momento de diagnóstico	seguimento de paciente	incidência de hidrocefalia
percentil de peso	regressão de Cox	deficiência de zinco
saúde de criança	promoção de amamentação	Infecção por HIV
infecção de repetição	curso de doença	incidência de doença
crescimento de criança	risco de morte	ato de amamentar
modelo de Count	realização de procedimento	resultado falso positivo
deficiência de vitamina	oximetria de pulso	taxa de sobrevida
distúrbio de desenvolvimento	uso de fármacos	leite materno exclusivo
centro de referência	alívio de dor	fator de confusão
meningite por Hib	níveis de linfócitos	protocolo de tratamento
válvula de uretra	Transtorno de ansiedade	Curva de sobrevida
duração de aleitamento	portador de deficiência	capacidade residual funcional
problema de saúde	monitorização de PIC	z de peso
cólica de lactente	resposta inflamatória sistêmica	neuro hipófise ectópica
escore de Williams	grupo de risco	administração de surfactante
consumo de medicamentos	esquizofrenia com início	vida intra uterina
programa EPI Info	crise de asma	evolução de paciente
equipe de saúde	acompanhamento pré natal	resposta a tratamento
paciente com FC	diagnóstico de paciente	medicamentos não aprovados
maioria de crianças	fator de proteção	crianças não amamentadas
promoção de saúde	crescimento intra uterino	escore de Shwachman
perda de peso	acalasia de esôfago	história familiar positiva
fórmula de soja	incidência de meningite	criança mais jovem
diagnóstico de pneumonia	uso de surfactante	Desenvolvimento motor oral
gravidade de doença	início de doença	término de tratamento
início de tratamento	avaliação de mamada	situação de estresse
cálculo de tamanho	período pré natal	problemas de comportamento
hemorragia digestiva alta	tratamento de paciente	baixo nível socioeconômico
incidência de DBP	vídeo EEG prolongada	mapa de micção
cor de pele	comparação entre grupos	som de fala
relaxamento muscular inadequado	índice de Apgar	células progenitoras hematopoiéticas
sala de parto	Estudos de neuroimagem	sucção não nutritiva
valor preditivo positivo	sucesso de aleitamento	Dificuldades de aprendizagem
uso de corticóide	pacientes de estudo	tratamento de sepse
suplemento de cálcio	pico de incidência	dificuldade de leitura
doença de base	QV de criança	início de quadro
risco de infecção	crianças de estudo	presença de cicatriz
filhos de mães	uso de fórmula	diagnóstico de AVBEH
pais de criança	síndrome de abstinência	pneumonia de repetição
Saturação de oxigênio	amostra de sangue	teste de triagem
Síndrome de desconforto	avaliação de estado	uso de diurético
desenvolvimento de criança	falta de apetite	principais efeitos colaterais
introdução de alimentos	média de escore	Coagulação intravascular disseminada
avaliação de dor	ausência de aleitamento	controle de PIC
vacinação contra influenza	Instituto Fernandes Figueira	Transtorno de humor
síndrome de Down	IgE sérica específica	diagnóstico de TDAH
aumento de PIC	uso de suplemento	fluxo sangüíneo cerebral
coarctação de aorta	manutenção de amamentação	depressão em criança
densidade mineral óssea	crianças com infecção	valor de referência
prevalência de asma	diagnóstico de sepse	tratamento de pneumonia
manejo de obesidade	tubo de ventilação	serviço de emergência
centro de saúde	presença de sintomas	número de consultas
estado de saúde	caracteres sexuais secundários	déficit de crescimento
escape de ar	crescimento de perímetro	taxa de aleitamento
ácaros de poeira	animais de experimentação	Avaliação de paciente
idade pré escolar	produção de IF	Indivíduos masculinos XYY
idade de paciente	prevenção de obesidade	uso de ventilação
índice de massa	hipotálamo hipófise adrenal	dieta sem glúten
terapia intensiva pediátrica	ingestão de alimento	época de diagnóstico
aumento de atividade	início de ventilação	relaxamento muscular adequado
Radiografia de tórax	distúrbio de sono	taxa de prevalência
comportamento de criança	pacientes com sepse	Sn córtico sensível
Hipertensão arterial sistêmica	baixo débito cardíaco	camada de ozônio
ansiedade de separação	criança com câncer	prevalência de EA
Pressão de perfusão	curva de crescimento	risco de doença
transtorno de conduta	grupo de crianças	doença cardíaca congênita
Pacientes com doença	artigo de revisão	crianças de Grupo
pico de massa	causa de RM	dados de estudo

Trigramas do *corpus* de Pediatria em ordem de frequência *tf-dcf* — Tabela 3 de 13

média de idade	teste de pezinho	altura de enterócito
média de peso	crianças com constipação	tratamento de criança
tipo de injúria	constipação crônica funcional	ingestão de cálcio
crianças pré escolares	doença de membrana	tempo de oxigenoterapia
uso em criança	herança autossômica recessiva	história de sibilância
situação de emergência	critérios de Wessel	infecção respiratória viral
quelato de zinco	tratamento de doença	período pós operatório
paciente com infecção	leucemia mielóide aguda	pacientes com SD
número de leucócitos	sobrevida de paciente	diagnóstico de hipertensão
crianças com idade	resistência a insulina	Metade de pacientes
redução de massa	valor preditivo negativo	peso de criança
doença arterial coronariana	níveis de anticorpos	taxa de amamentação
pacientes com asma	prática de consultório	uso de preservativo
infecção por Hp	prevalência de infecção	correlação de Spearman
Student para amostra	indicação de ECMO	agente de saúde
encefalopatia hipóxico isquêmica	sistema de saúde	presença de insuficiência
período de janeiro	Meningite por Haemophylus	antiinflamatório não hormonal
coleta de colostro	obesidade em crianças	uso de corticosteróide
Pesquisa de instituição	mecanismo imunológico envolvido	teste de otoemissão
prognóstico de paciente	desenho de estudo	uso de CPAP
diagnóstico de DGH	administração de vacina	caso de infecções
momento de parto	diferença entre médias	fator de necrose
tipo de alimento	Endoscopia digestiva alta	crescimento de comprimento
viés de seleção	exposição a forças	necessidade de oxigênio
Relato de caso	Metade de crianças	crescimento de RNPT
tempo de hospitalização	tratamento de crise	transtorno de comportamento
magnitude de problema	Coleta de sangue	vida pós natal
regressão logística múltipla	hepatite aguda viral	diagnóstico de bronquiolite
sala de emergência	risco de fratura	RVU de grau
hormônio de crescimento	realização de cirurgia	condição de nascimento
maioria de meninas	limites de normalidade	solução de NaCl
baixo nível social	duração de amamentação	período pós natal
presença de IgE	crianças alto xinguanas	admissão de paciente
tratamento de obesidade	diagnóstico de SAM	absorção de nutrientes
contagem de leucócitos	ingestão de vitamina	diagnóstico de infecção
aumento de peso	pai de paciente	desenvolvimento de doença
critério de normalidade	diagnóstico pré natal	prontuário de paciente
diminuição de massa	acompanhamento de paciente	casos de hipertensão
via aérea artificial	fase de indução	Indução de lesão
período de ventilação	pacientes com relaxamento	diagnóstico de insuficiência
consulta pré natal	taxa de remissão	trauma de crânio
término de estudo	radiograma de tórax	soluções salinas hipertônicas
uso de mamadeira	problema de crescimento	insuficiência respiratória hipoxêmica
curva de referência	vacinação contra Hib	coleta de urina
escala de dor	metabolismo de ácido	terapia com surfactante
curvas de velocidade	transtornos de alimentação	deficiência de surfactante
atresia de vias	presença de mutação	grupo com analgesia
momento de alta	relatos de literatura	dislexia de desenvolvimento
crianças com DBP	criança com doença	Analgesia com opióides
portoenterostomia de Kasai	vantagens de amamentação	aspiração de mecônio
casos de AVBEH	maior em crianças	início em infância
idade pós concepcional	pacientes com alergia	comprometimento de estado
presença de febre	delineamento de estudo	profilaxia com penicilina
teste de estimulação	níveis de hemoglobina	qualidade de MG
insuficiência supra renal	adequado para idade	ingestão de leite
região não dependente	Características de pacientes	presença de pais
variáveis de confusão	teste de Fisher	duração de ventilação
índice de oxigenação	concentração de IgA	Hidrato de cloral
Indivíduos com anemia	leite de peito	Transtorno obsessivo compulsivo
transtorno de personalidade	sintoma mais freqüente	uso de antimicrobiano
diagnóstico de esquizofrenia	ventilação mecânica convencional	Hipertensão pulmonar persistente
avaliação de crescimento	tempo de colestase	droga em leite
alterações de comportamento	participação em estudo	Monitorização de crescimento
pós termo e	contagem de linfócitos	redução de PIC
estudo duplo cego	grupo de mães	ejeção de leite
diagnóstico de autismo	início de dieta	final de tarde
início de ação	veia cava superior	transdutor em ponta
sinais de alerta	ausência de lesão	lesão cerebral secundária
transmissão de vírus	Serviço de Referência	transfusão de plaquetas
episódio de ITU	conteúdo mineral ósseo	recorrência de ITU
remoção de MSV	administração de droga	diagnóstico de ITU
Síndrome de álcool	pico de pressão	amostra de urina
país em desenvolvimento	achado de estudo	doença de parênquima
Dados de literatura	técnica de amamentação	criança com TDAH
uso de oxigênio	base de soja	uso de corticosteróides
crianças mais velhas	tratamento de escolha	etiologia de TDAH
aplicação de vacina	saturação de hemoglobina	tratamento de transtorno
uso de fenoterol	pressão arterial diastólica	Seqüência de intubação
adolescentes de sexo	teste cutâneo positivo	sopro de Still
evolução de doença	refeição de família	Modo de ventilação

Trigramas do *corpus* de Pediatria em ordem de frequência *tf-dcf* — Tabela 4 de 13

dose de manutenção	indução de remissão	tempo de ventilação
tratamento de SDRA	dupla de observadores	uso de leite
uso de posição	avaliação de resposta	assistente de pesquisa
estilo de vida	faixa etária estudada	bronquiolite viral aguda
t de Student	vida extra uterina	critério diagnóstico utilizado
risco de óbito	auxiliar de enfermagem	desenvolvimento de filho
tamanho de osso	correção de Yates	grupo de comparação
bairro de Pedregal	diferença estatisticamente significativa	contagem de plaquetas
uso de indometacina	grupo de leite	incentivo a aleitamento
adesão a dieta	mau controle metabólico	controle de peso
estudos de genética	expectativa de vida	necessidade de reintubação
exclusão de alimento	início de antibioticoterapia	óxido nítrico inalatório
estabelecimento de diagnóstico	fragmentos de biópsia	sais de cálcio
nível de evidência	necessidade de intubação	cloro em suor
déficit de atenção	tipo de alimentação	Cystic Fibrosis Foundation
tratamento de asma	prevalência de RGEP	período de incubação
manejo de aleitamento	fisiopatogenia de BA	hábito de sucção
dificuldades de amamentação	exato de Fisher	alimentos semi sólidos
casos de intoxicação	mediana de amamentação	sintoma urinário diurno
diagnóstico de DGC	risco de aspiração	episódios de enurese
prevenção de injúrias	escolaridade de mãe	apoio a mãe
colonização de orofaringe	diagnóstico de asma	curva de Lubchenco
polissacarídeos não celulósicos	dilatação de pupila	secreção de insulina
sinais de doença	diagnóstico de miocardite	glicemia de jejum
razão de prevalência	receptor de transferrina	ação de insulina
detecção de obesidade	ausência de diarréia	Pacientes com St
diagnóstico de DSR	duração de queixa	baixa auto estima
protetores de mamilo	mãe de criança	número de neutrófilos
período de acompanhamento	Sistema de escore	sintomas de infecção
prognóstico de criança	teste in vitro	boca de criança
crianças de cor	curva de peso	utilização de via
concordância entre observadores	teste de urease	compreensão de linguagem
terapia anti retroviral	Complicações de úlcera	incremento de peso
final de expiração	posto de saúde	passagem de tubo
muito baixo peso	BVA por VRS	dinâmica de crescimento
casos mais graves	escore de Apgar	recém nascidos menores
período de seguimento	necessidade de suporte	erradicação de H
produto de distorção	idade de mãe	aumento de resistência
escore de QI	evidência de doença	vacina BCG PC
diagnóstico de RM	pacientes com DGH	proteção contra tuberculose
experiência com tabaco	interrupção de amamentação	elevação de IF
crianças de idade	estágio de doença	estratégia de atenção
método de triagem	animais de grupo	Monitorização de pressão
idade gestacional corrigida	conforto de paciente	não corticosteróide antenatal
Desenvolvimento de linguagem	maior em grupo	família de criança
presença de doença	período de maio	consulta de rotina
pequenas vias aéreas	Regulação de balanço	absorção de ferro
remoção de CE	sucesso de extubação	diurético de alça
paciente de faixa	falta de treinamento	vida de prematuros
estabelecimento de aleitamento	formação de cicatrizes	incidência de cicatriz
prognóstico de doença	comparação de variáveis	distúrbio de linguagem
saúde materno infantil	medicamentos não padronizados	avaliação de desenvolvimento
retinol em colostro	mediana de idade	Risco de recorrência
níveis de vitamina	Hipertensão arterial grave	mecanismo de defesa
níveis de retinol	pelo menos lesão	intubação sem medicação
crianças em aleitamento	amostras de leite	insuficiência renal aguda
via aérea difícil	sintoma mais comum	porcentagem de linfócitos
portadores de malária	fase de vida	RN de termo
boa condição socioeconômica	trato urinário inferior	alto valor preditivo
pressão positiva contínua	suplemento de vitamina	extração de leite
protrusão de língua	crianças com dieta	função de eixo
quimioterapia pré operatória	qualidade de leite	tremores de frio
inclusão de paciente	duração de AM	liberação de histamina
duração de AIJ	caso de criança	programas de intervenção
Ultra sonografia abdominal	gravidade de quadro	curva de NCHS
trato urinário superior	avaliação de criança	incentivo a AM
desconforto respiratório agudo	número de sintomas	distúrbio de comportamento
cânula de traqueostomia	função cortical superior	segmento renal remanescente
fórmula pré espessada	dose de medicação	RVU pré operatório
recém nascidos controles	número de crianças	tratamento de dor
controle sem hepatopatia	parte de rotina	abuso de drogas
adolescentes com TVP	níveis de leptina	surgimento de cárie
preenchimento de questionário	avaliação de imunidade	pré termo e
leitos de terapia	risco de atraso	aprendizado de leitura
paciente em grupo	ocorrência de infecção	alterações de linguagem
pacientes com enterocolite	parâmetros de ventilador	pacientes com insuficiência
enterocolite necrosante neonatal	ambulatório de pediatria	tipo de profilaxia
nível de escolaridade	protocolo de investigação	paciente com anemia
remissão clínica completa	Meta análises recentes	comportamento de bebê
probabilidade de sobrevida	opção de tratamento	alimentos de origem

Trigramas do *corpus* de Pediatria em ordem de frequência *tf-dcf* — Tabela 5 de 13

molécula de Fas	níveis de PA	criança de creche
Ventilação com pressão	ocorrência de fratura	Valores de escores
avaliação de MG	cálculo de poder	metade de profissionais
uso de clonidina	tempo de aleitamento	crianças de risco
boca de bebê	artigo de Silva	portadores de hepatite
leite de mãe	controle de dor	pacientes com malária
síndrome de Angelman	desenvolvimento de infecção	diagnóstico de malária
processo de amamentação	cepas de Streptococcus	casos de hepatite
desenvolvimento de cicatriz	amamentação bem sucedida	gordura em fezes
paciente com hipertensão	casos de pneumonia	soja sem fibra
criança com transtorno	dose de medicamento	Unidade de Saúde
secreção de prolactina	replicação de HIV	índias de PIX
manutenção de lactação	elevação de carga	natureza de variáveis
receptor de melanocortina	resultados de ensaio	recomendação de fabricante
síndrome de Asperger	diagnóstico de DC	títulos de soroproteção
redução de peso	período de junho	resposta a vacinação
alimentação de filho	maioria de doenças	ocorrência de soroproteção
hora de dormir	Nascimento de crianças	momento de internação
transtorno de pânico	exacerbação de asma	Toxicidade de drogas
leite materno ordenhado	grupo de adolescentes	caso de hidrocefalia
concentrado de hemácias	consumo de fibra	medida de dobras
tratamento de HIC	resistência a drogas	hiper responsividade brônquica
pressão arterial média	introdução de dieta	diminuição de mortalidade
urina não centrifugada	gráfico de percentis	extensão de doença
diagnóstico de primoinfecção	cidade de Maceió	trato respiratório superior
dano renal crônico	indicador perímetro braquial	associação com aumento
ajustamento com humor	boas evidências científicas	vantagens de aleitamento
animais de laboratório	envolvimento de membro	tempos de coagulação
morte de criança	alimentação de bebês	estado de remissão
supervisão de saúde	ressonância nuclear magnética	classificação de risco
projeto Bright Futures	Estudos in vitro	acometimento de SNC
síndrome de West	medicina de emergência	Curvas de Kaplan
distúrbio de coagulação	diagnóstico de CDPR	população de adultos
palpação de pulsos	dia de alta	desaparecimento de sintomas
infecção de trato	pneumonia em crianças	via de transmissão
sopro cardíaco inocente	lipodistrofia generalizada congênita	proporção de crianças
borda esternal esquerda	imaturidade de sistema	fase de consolidação
Promoção de alimentação	sistema nervoso autônomo	diretor de escola
Higiene de alimentos	estudos de prevalência	redução de NBT
teste de QI	adolescentes com DC	tratamento de RGE
síndrome de Rett	vacina contra VHB	regressão logística ordinal
etiologia de RM	vacina contra Hib	episódio de RGE
defeito estrutural congênito	CCDTA SS RS	pacientes com trombose
resposta a NOi	número de desvios	pacientes com TVP
administração de NOi	hidratos de carbono	fator V Leiden
casos de dor	infecção pós natal	atividade de proteína
paciente com câncer	início de aleitamento	limitação de atividades
local de trabalho	diversas faixas etárias	avaliação pré operatória
grupo de estudo	pacientes aqui apresentados	articulações com sinovite
padrão de resposta	peso fecal úmido	método de Kaplan
aumento de pressão	poder de estudo	presença de cardiomegalia
qualidade de vida	tempo de tratamento	pediatra de emergência
Características de população	irradiância espectral média	pacientes com cirrose
revisão de literatura	prejuízo de sono	habilidades de aconselhamento
protocolo de pesquisa	disponíveis em Brasil	análise de sobrevida
asma aguda grave	exames de triagem	diagnóstico de alergia
vez em vida	série de casos	atendimento de urgência
utilização de corticóide	ano de estudo	crianças de países
crianças que faleceram	tipo de medicamento	desenvolvimento de DBP
achados de autores	serviço de pediatria	RN com DBP
células de microglia	síndrome de Meckel	exame de rotina
uso de teratógenos	infecção respiratória aguda	radiografias de coluna
período de vida	técnica de pour	esquema de tratamento
presença de sibilos	solidificação de meio	diagnóstico de Er
diagnóstico de atresia	diluições decimais selecionadas	Distribuição de crianças
atividade física incorporada	medida de perímetro	percentual de creme
resposta a estímulo	gráficos de crescimento	amamentação em Brasil
referida meta análise	amamentação com leite	Idade de início
pacientes mais graves	intubação de emergência	menor poder aquisitivo
resultado de exame	incidência de pneumotórax	estatura de criança
profissionais de enfermagem	vacina contra vírus	malformações de trato
forma de tratamento	momento de coleta	sintomas de IC
crescimento pós natal	vasodilatador pulmonar seletivo	saúde de país
dificuldade de diagnóstico	uso de terapia	crianças com alergia
medida de prevenção	fisiopatologia de doença	soroprevalência de infecção
associação em questão	condição socioeconômica materna	morador de domicílio
apenas leite materno	abordagem de paciente	trato digestório proximal
Instituto da Criança	perímetro de cintura	refluxo gastroesofágico patológico
grupo de Campinas	orientação de conduta	diagnóstico de hepatopatia
ganho pôndero estatural	momento de acidente	alterações ultra sonográficas

Trigramas do *corpus* de Pediatria em ordem de frequência *tf-dcf* — Tabela 6 de 13

achados ultra sonográficos	eficácia de vacina	tempo de acompanhamento
Critério de Roma	pacote estatístico SPSS	rara em crianças
Colaterais de sistema	maior entre crianças	importância de diagnóstico
tipo de infecção	acometimento de membros	início de puberdade
cepas de pneumococo	intensidade de sintoma	diagnóstico de otite
padrão de herança	duração de doença	distúrbio de alimentação
suporte ventilatório invasivo	prevalência de soropositividade	dieta de criança
Validação de EQVC	freqüência de crises	início de seguimento
recomendação de OMS	meninas de EB	manutenção de níveis
coleta de exames	utilização de critérios	razão de probabilidade
escolaridade de pais	subdiagnóstico de doença	Tc de crânio
estado de portador	número de pacientes	menor idade gestacional
número de doses	título de HAC	ganho de comprimento
Variáveis explanatórias qualitativas	aleitamento em alta	uso de SRI
adiposidade em crianças	identificação de agente	atendimento de pacientes
prevalência de diarréia	momento de avaliação	vírus de hepatite
ocorrência de diarréia	grau de desnutrição	quantidade de células
rotina de hospital	análise morfométrica digitalizada	cabeça de criança
prevenção de eventos	criança com lesão	prevalência de aleitamento
hábito de fumar	risco de complicações	aplicação de surfactante
HC de FMUSP	finais de semana	fome de bebê
influência de paridade	faixa etária atendida	teste de desenvolvimento
renda per capita	escassez de estudos	prevalência de transtorno
títulos de anticorpos	programa de educação	superfície de células
termo de compromisso	utilização de vacina	Descrição de casos
início de manifestação	diagnóstico de BVA	maior idade gestacional
erradicação de bactéria	indicador de risco	diagnóstico de lesão
Tratamento do Helicobacter	shunt esquerdo direito	ressuscitação com volumes
escore clínico modificado	uso de insulina	ressuscitação com solução
Aleitamento materno predominante	prevalência de deficiência	peso de paciente
sangue de cordão	crianças com risco	aplicação de BCG
resposta a vacina	exames de neuroimagem	modelo após regressão
produção de citocinas	confirmação de diagnóstico	início de mamada
crianças com colestase	proteção de leite	incidência de otite
adolescentes com colestase	estudo piloto realizado	abuso de substâncias
risco de mortalidade	transporte de pacientes	sibilância de repetição
lactentes com sibilância	cultura de linfócitos	diagnóstico de cicatriz
freqüência de alterações	Alteração motora oral	crianças sem DBP
total de crianças	última relação sexual	RVU de pacientes
amostra de secreção	comportamento de risco	banco de leite
cálculo de proporção	unidade de cuidados	cuidado com criança
amostra casual simples	reservas de ferro	concentração de hemoglobina
problemas de amamentação	uso de alimentos	incidência de desfechos
insuficiência respiratória grave	interrupção de aleitamento	terapia de suporte
saúde de população	progressão de doença	droga de escolha
concentração de IgG	amostra de conveniência	final de adolescência
equipe de pesquisa	Métodos Foi realizado	prevalência de alteração
retardo em diagnóstico	Prevalência de distúrbios	bebês de grupo
treinamento em AM	utilização de surfactante	obesidade em infância
grupo de seguimento	teste de McNemar	administração de dose
equipe de AM	integridade de mucosa	diagnóstico de transtorno
painel de peixe	risco de obesidade	custo de tratamentos
freqüência de positividade	responsáveis por criança	etiologia de AVBEH
desenvolvimento de asma	Complicações de obesidade	colestase extra hepática
programa estatístico utilizado	Características de amostra	dificuldades de linguagem
somente leite materno	cuidado pré natal	teste cutâneo negativo
dose de insulina	forma de gordura	atenção de criança
disfunção de sistema	resultado de terapia	desenvolvimento de caracteres
aplicação de PEEP	sentimento de culpa	altura de pai
época de internação	trato respiratório inferior	risco de intoxicação
características de domicílio	necessidade de internação	pressão capilar pulmonar
tipo de lesão	pega de bebê	solução salina isotônica
atendimento a paciente	mecanismo de ação	técnica de aleitamento
tratamento de primeira	fibrocísticos de países	diversos grupos experimentais
sono de criança	suco de frutas	subgrupo de pacientes
ocorrência de cólica	otoemissão acústica evocada	teste com ACTH
incidência de cólica	diagnóstico de disfunção	secreção de cortisol
Presença de cólicas	benefício de aleitamento	paciente com choque
nutrição de criança	episódio de infecção	concentração de cortisol
perda de nutrientes	indicação de ventilação	fator de crescimento
via de administração	introdução de chás	lesão cerebral traumática
distribuição de pacientes	duração de mamadas	infusão de solução
força de associação	percentil de estatura	repetição de exame
Mecanismos de lesão	decisão de amamentar	crianças pré púberes
medida de peso	etiologia de transtorno	suplementação com ferro
curso de Medicina	fluxo sangüíneo hepático	prevalência de anemia
necessidade de oxigenoterapia	síndrome de Prader	Desnutrição protéico calórica
história de prematuridade	técnica de translactação	doença auto imune
incidência de infecções	aumento de morbidade	atraso de desenvolvimento
crises de sibilância	uso de oxigenoterapia	sarcoma de Ewing

Trigramas do *corpus* de Pediatria em ordem de frequência *tf-dcf* — Tabela 7 de 13

saúde de filho	mãe de RNPT	CIM de penicilina
indicação de analgesia	Transtorno afetivo bipolar	estudo recentemente publicado
grupo sem analgesia	efeito de drogas	artigo de Anderson
prolapso de ureterocele	eficácia de conduta	proteína de soja
problemas pós operatórios	velocidade de infusão	validação de CHAQ
ITU pós operatória	aumento de ventrículo	manejo de pacientes
crianças com RVU	sintomas de sepse	esquizofrenia de início
vida de bebê	freqüência de mamadas	aumento de massa
kg de solução	gênese de sintoma	experiência em país
prevalência de CLPE	prognóstico de autismo	instituição de terapêutica
hipoplasia de esmalte	células de Purkinje	tubo em situação
risco de depressão	ganho de peso	autores de artigo
necessidades de paciente	indicação de cirurgia	redução de comportamentos
coorte de crianças	taxa de cura	condução de casos
suspeita de infecção	cura de RVU	dificuldade de pais
sintomas de doença	cistite não complicada	constatação de Frost
uso de corticoesteróides	UTIs de adulto	maternidade de país
sintoma de obstrução	internação em UTI	tratamento de fibrose
síndrome torácica aguda	crianças com depressão	exercícios contra gravidade
Septicemia por salmonela	hipertensão arterial secundária	presença de débito
ritmo de crescimento	controle de sintomas	pneumonia de aquisição
resposta a terapêutica	retirada de MSV	uso de nutrição
satisfação de famílias	hora de morte	neurobiologia de comportamento
ponta de língua	final de vida	vida de paciente
veia jugular interna	pressão arterial sistêmica	uso de álcool
tipo de cardiopatia	Controle de pressão	único fato aceito
mecanismo fisiopatológico envolvido	nitroprussiato de sódio	base de método
autonomia de criança	dependência de drogas	anticorpo anti Helicobacter
tratamento de psicoses	alterações de sono	grande experiência clínica
esquizofrenia em infância	perda de consciência	verdade de hoje
vírus sincicial respiratório	tratamento de TDAH	Controvérsias a parte
consultas de puericultura	uso de metilfenidato	efusão de orelha
pacientes com transtorno	fornecimento de oxigênio	indicações de TCH
relação médico paciente	uso de imunoglobulina	criação de centro
necrólise epidérmica tóxica	diagnóstico de cardiopatias	doador não aparentado
maioria de reações	uso de ácido	autores de estudo
resultados de USC	transmissão de sopro	escolas de saúde
evolução neurológica anormal	prolapso de válvula	reação em cadeia
mortalidade de pacientes	insuficiência cardíaca congestiva	baixo rendimento escolar
complexo aréolo mamilar	exame físico geral	resposta a questionário
transtorno depressivo maior	uso de antipsicótico	impacto de mortes
insuficiência respiratória aguda	tipo de crise	casos de constipação
utilização de antibióticos	canais de sódio	cepa de S
uso de estimulantes	fatores de coagulação	artrite idiopática poliarticular
córtex pré frontal	bicarbonato de sódio	probabilidade de doença
aumento de gordura	ambiente de UTI	síndrome de morte
consumo de oxigênio	pressão de átrio	manejo de criança
novas cicatrizes renais	vasoconstrição hipóxica pulmonar	mulheres com problemas
resistência vascular periférica	tratamento com NOi	redução de efeitos
habilidades de criança	terapia com NOi	tratamento de otite
fabricantes de leites	necessidade de ECMO	controle de hipertensão
ingestão de sódio	efeitos de NOi	causa de morte
diagnóstico de sinusite	célula muscular lisa	subtipos de LMA
tratamento de St	enxaqueca sem aura	alternativas de tratamento
pacientes com TOC	crise de dor	duplo cego controlado
coréia de Sydenham	casos de Dar	maior umidade fecal
compulsões de verificação	Critérios de SIC	autora de editorial
Estudos genético familiares	suporte de terapia	prevenção de DTN
aumento de freqüência	administração de oxigênio	prevalência de DTN
casos de TCE	instituição de ventilação	mulheres em idade
acidente vascular encefálico	uso de PEEP	conseqüências de hidrocefalias
paciente com esquizofrenia	neonatos com hipertensão	apoio a aleitamento
parte de corpo	número de casos	utilização de IDP
supressão de lactação	período de estudo	manejo de trauma
idade de lactente	grupo de nível	efeitos adversos sistêmicos
concentração em leite	aumento de idade	dispositivo inalatório ideal
aparecimento de sintomas	limitação de estudo	cálculo amostral prévio
etiologia de obesidade	ausência de alteração	países de mundo
estudos com famílias	aplicação de questionário	detecção de antígenos
vírus de influenza	hábito de vida	isolamento de vírus
síntese de leite	maioria de estudos	ensaio clínico prévio
níveis de prolactina	tempo de atendimento	cidade de Pelotas
esvaziamento de mama	uso de fumo	presença de processo
Bloqueio de ductos	evolução de gravidez	período pós vacinal
problemas de desenvolvimento	corticóide pré natal	carga viral secundária
alívio de sintomas	capital mineral ósseo	antiinfluenza em pacientes
transmissão de doença	artigo em foco	grupo de cirurgia
mãe com tuberculose	prática de ginástica	casos de RM
medicação de primeira	sintomas de asma	asfixia perinatal grave
kg de peso	assistência a paciente	troca de gases

Trigramas do *corpus* de Pediatria em ordem de frequência *tf-dcf* — Tabela 8 de 13

risco de prejuízo	exposição a sol	Estudos epidemiológicos recentes
baixa condição socioeconômica	receptor de interferon	aplicabilidade de método
maioria de centros	membro superior esquerdo	método de ELISA
diferenças entre pacientes	situação de saúde	aumento de volume
ocorrência de lesões	via de exposição	manejo de síndrome
possível associação causal	procura de CCI	hipertrofia de parótidas
atenção para fato	intoxicação por clonidina	peroxidação de lipídeos
orientação de autor	maioria de mulheres	progressão de aparelho
desenvolvimento de especialidades	tratamento de linfangioma	procedimento anti refluxo
criativo em tema	pacientes com regressão	avaliação de DMO
suspensão de quimioterapia	endotélio de linfangioma	afilamento de esôfago
triagem auditiva neonatal	alterações em função	saúde de mulher
desenvolvimento de função	papel de pediatra	coleta de fezes
característica de doença	alunos de escola	história de alergia
programas de screening	paciente aqui descrito	freqüência de asma
RN que apresentam	mãe de paciente	famílias mais carentes
dificuldade de intubação	especificidade de escores	crianças com diarréia
confirmação de intubação	uso de RXT	vacinação em rede
lesão cerebral isquêmica	diminuição de reflexo	vacinação contra agente
recomendação de amamentação	Tipo de aleitamento	meningites em RS
realização de RXT	lacunas de conhecimento	artigo de suplemento
Grupo de Vacinas	esôfago de Barrett	espaçador de metal
sexo de paciente	forma de comprimidos	z de indicadores
utilização de fórmula	lugar de destaque	população de referência
conhecimento de enfermidade	final de estudo	medida de prega
ativação de sistema	tratamento de DSR	anos de idade
risco de reação	distrofia simpática reflexa	participação de criança
amamentação de prematuro	origem de criança	início de noite
metabolismo de cálcio	hipertensão em criança	redução de incidência
indicação de fototerapia	posições de mamadas	comprometimento de saúde
faixa de peso	pomadas com corticóide	redução de carga
maioria de recomendações	lanolina anidra modificada	atenção a saúde
continuidade de amamentação	dor em mamilos	boa evolução clínica
Estatuto da Criança	compressas com água	casos mais leves
observação de mamada	coador de plástico	gênese de hipertensão
locais de Brasil	cicatrização de feridas	Bogalusa Heart Study
primeira amostra fecal	base de vitamina	população de crianças
identificação de rotavírus	rotinas de maternidades	coleta de hemocultura
identificação de genótipos	estabelecimento de amamentação	trabalho anteriormente publicado
pós cirurgia cardíaca	lesão glomerular mínima	teste de Coombs
nascimento de bebê	diagnóstico de Sn	possibilidade de intervenção
pós parto imediato	Sn com lesões	tamanho de câmara
inserção de mulher	capacitação de profissionais	menor escolaridade materna
classes menos favorecidas	títulos mais elevados	mãe HIV positivo
oferta de leitos	piora de hipertensão	detecção de alterações
estudo de Souza	nascidos em HCFMUSP	secreção ácido péptica
interpretação de densitometria	mortalidade em subgrupo	graus mais graves
tipo de inalador	fatores prognósticos analisados	friabilidade em esôfago
medidas sanitárias urgentes	diagnóstico de HDC	tempo após nascimento
aquisição de infecção	vias aéreas periféricas	lesões de pele
anticorpos pós vacinais	produção de CFC	Estudo de seguimento
morbidade de doença	ponte nasal achatada	mecanismos fisiopatológicos relacionados
adequação de rotinas	dia de corticóide	ocorrência de vômitos
tipo de risco	IPDM com CFC	técnica de inóculo
população de risco	Dificuldades durante intubação	número mais provável
período de manhã	CIV subaórtica perimembranosa	aquisição de amostras
Fundação Oswaldo Cruz	síndrome de Horner	aparecimento de fissuras
retardo de crescimento	função de músculo	alça de Drigalsky
parada cardíaca repentina	exame neuro oftalmológico	quadro de paralisia
método de tratamento	fechamento de tubo	diagnóstico mais preciso
ecocardiograma com Doppler	manejo de lactação	risco de sobrepeso
alto fluxo pulmonar	planejadores de saúde	diagnóstico de endocardite
estratégias de ventilação	programa de triagem	deiscência parcial recente
resultados muito semelhantes	ausência de estimulação	controle de asma
imunização básica completa	lesão pré existente	amamentação em situações
realização de entrevista	persistência de febre	diagnóstico de OCMR
maioria de asmáticos	episódio de diarréia	curso doloroso prolongado
uso de placebo	doença diarréica aguda	biópsia de mandíbula
sepse neonatal precoce	crianças em faixa	crescimento de bebês
mediana de grupo	taxa de fumantes	anos de evolução
grande volume fecal	rede de ensino	diversas regiões geográficas
farelo de aveia	programas de prevenção	vacina contra pneumococo
absorção de água	prevalência de tabagismo	distensão vesical persistente
níveis de bilirrubinemia	panorama de prevalência	derivação urinária temporária
incentivo a amamentação	linha de pensamento	vício de seleção
teoria mais aceita	indústria de tabaco	exposição a agente
obstáculo a amamentação	experiência com cigarro	retirada de traqueal
unidades de rede	definição de fumante	cânula em pescoço
uso de óxido	QE de ISAAC	aspiração de CE
esquema vacinal completo	Formas mais brandas	tratamento de RN

Trigramas do *corpus* de Pediatria em ordem de frequência *tf-dcf* — Tabela 9 de 13

tipo de ventilação	RN com IG	doença pneumocócica invasiva
período neonatal conclusão	RNPT com PN	determinação de susceptibilidade
constante de tempo	perfil de pacientes	antígeno polissacarídico capsular
altos volumes correntes	subgrupo de hidrocefalias	perfil de evolução
aspectos técnicos necessários	perímetro cefálico maior	fórmula anti regurgitação
benefícios de amamentação	hidrocefalias em CAISM	espessamento de dieta
provas de função	grupo de hidrocefalias	diagnóstico de RGE
regressão de febre	diagnóstico de hidrocefalia	amido de milho
episódios de sibilância	aqueduto de Sylvius	recursos de terapia
vigência de infecção	Maternidade de CAISM	modelo de Cox
hospitais universitários brasileiros	teórico prático específico	variação de prevalência
umidade de ar	tentativas de intubação	momento de estudo
adolescentes com doença	movimentos de extremidades	acompanhamento de criança
swab de orofaringe	indução de SRI	ultra sonografia cerebral
prevalência de colonização	dados de fase	obtenção de liquor
doença invasiva pneumocócica	variabilidade de medidas	lesão isquêmica cerebral
composição de vacina	entendimento de fisiopatologia	idade de grupo
colonização de nasofaringe	abordagem terapêutica precisa	coagulação em pacientes
colonizadores de orofaringe	momento de análise	adultos com TVP
teor de vitamina	uso de prongas	adolescentes com cirrose
estado nutricional materno	porção de pronga	Deficiência de proteína
concentração de vitamina	perda de pressão	vida de família
concentração de retinol	geração de pressão	índice de equilíbrio
pessoas de sexo	geração de CPAP	tipo de terapia
segurança em trânsito	curva de Alexander	quantificação de resultados
decorrência de injúrias	padrão de média	maior atividade física
dados de Datasus	uso de quimioterapia	diminuição de gasto
mudança em perfil	pacientes com metástases	avaliação de obesidade
obesidade em sexos	fatores prognósticos desfavoráveis	remissão de quadro
grupo obeso feminino	adaptação a vida	Evidências de literatura
doenças arteriais coronarianas	estimativa de prevalência	exposição a medicamentos
resistência a antimicrobianos	ausência de marcador	presença de resposta
pacientes com meningite	risco de população	índice de capacidade
presença de refluxo	peso de placenta	quality of my
biópsia de esôfago	leptina de cordão	percepção de pais
medidas preventivas eficazes	dimorfismo sexual relacionado	pacientes com artrite
estudos com corticóide	concentração de leptina	número de articulação
dado de prevalência	reação de hospedeiro	limitação de movimentos
padrão de crescimento	complexidade de relações	evidência de sinovite
inoculação de urina	Relatos em literatura	Comissões de Ética
embriões de galinha	síndrome de ativação	estratégia de tratamento
crescimento pôndero estatural	sais de ouro	doença alérgica respiratória
contato com responsáveis	pacientes com SAM	D Blomia tropicalis
uso de antibioticoterapia	pacientes com AIJ	diminuição de resistência
regular estado geral	início de AIJ	tratamento farmacológico específico
pesquisa de plasmódio	introdução de sulfasalazina	estudos com pacientes
malária em infância	insuficiência hepática aguda	detecção de lesões
casos de malária	etiopatogenia de SAM	predomínio de Klebsiella
aminotransferases em malária	diagnóstico de AIJ	grupo com enterocolite
usuárias de chupeta	Elevação de enzimas	asma em crianças
Hospital das Clínicas	recombinação BCR ABL	indicadores de estado
comprometimento de função	profilaxia de SNC	tipo de dieta
teste de tolerância	leucemia linfocítica aguda	resultados de casuística
má formação congênita	importante fator prognóstico	recidiva de doença
função de célula	envolvimento de SNC	quimioterapia de indução
manutenção de recrutamento	contagem de blastos	probabilidade de SLE
manobras de recrutamento	alto risco RR	posição de RN
ração com celulose	semelhante em grupos	idade em início
kg de ração	obesidade de filhos	freqüência de atelectasia
início de rações	adiposidade em escolares	efeito de posição
fonte de fibra	Hospital Municipal Jesus	duração de remissão
fezes de grupo	recuperação de criança	crianças com LMA
amostras de fezes	ocorrência de distúrbios	tratamento mais adequado
dados norte americanos	trabalhadores de saúde	índice cardiotorácico médio
grandes centros urbanos	lesão de tecido	identificação de colapsos
índice de impedância	intensidade de exposição	evolução de ICT
valores de Mm	leite de mama	estatística de Kappa
valores de IB	fisiologia de lactação	diagnóstico de IRAB
população alto xinguana	suplementação com zinco	análise de ICT
equação de Slaughter	refeições de sal	distribuição por sexo
problemas com aleitamento	teste de emissões	desenvolvimento de bebê
sorotipos de Streptococcus	adoção de medidas	setor de urgência
diagnóstico de bulimia	observado em estudos	serviços de urgência
problema com amamentação	homogeneidade de população	parturiente não preparada
realização de triagem	termo com encefalopatia	aconselhamento em amamentação
movimentos de mastigação	técnica de microdiluição	Pacientes de gênero
soroproteção mais elevadas	resistência in vitro	indivíduos de idades
resposta de soroproteção	resistente de nível	classe socioeconômica baixa
produção de anticorpos	perfil de sorotipos	trato digestivo alto
dobro de dose	halo de inibição	episódio de regurgitação

Trigramas do *corpus* de Pediatria em ordem de frequência *tf-dcf* — Tabela 10 de 13

Faculdade de Medicina	desenvolvimento de tolerância	oxigênio de hemoglobina
Ventilação líquida parcial	curva de soroprevalência	necessidade de VM
administração de leite	Comentários Hemorragia digestiva	infecção por VRS
recuperação de paciente	Biópsia de antro	gravidade de BVA
faixa de normalidade	vesícula biliar atrófica	ensino de aleitamento
período de desenvolvimento	presença de regurgitação	análise de amamentação
informação de paciente	nível de linha	população aqui estudada
Exames de imagem	nodularidade de borda	prevalência de osteopenia
estratégias de controle	colaterais porto sistêmicas	pacientes colestáticos estudados
suplementação de oxigênio	causas de hepatopatia	evidência de colestase
internação de RN	caracterização de RGEP	crianças com hepatopatia
incidências para DBP	alteração de conduta	adolescentes com hepatopatia
evolução de RN	ausência de sintomas	injeção de toxina
efeito de idade	número de estudantes	efeitos adversos graves
casos de DHEG	precipitado de amostra	síndrome de imunodeficiência
RN sem DBP	lactentes com BA	hipertensão pulmonar primária
postos de vacinação	amostras de ANF	parte de investigação
avanço de conhecimento	ANF de pacientes	presença de anormalidades
segundo faixa etária	turno de manhã	possíveis mutações genéticas
proteção contra obesidade	consenso sobre significado	integridade de haste
investigação mais aprofundada	binômio criança família	idade gestacional menor
ingestão de energia	amostra São Paulo	hábitos alimentares inadequados
grau de esofagite	pinças de biópsia	terapêutica de resgate
exame de EDA	insucesso de tratamento	mudança em vidas
esofagite não erosiva	controle de infecção	expectativa de cura
efeito de tratamento	turno de estudo	diagnóstico de St
diferença estatisticamente significante	questões de saúde	diferença entre proporções
correlação de Pearson	pesquisa em Holanda	valores de normalidade
tratamento medicamentoso empregado	medicamentos entre adolescentes	doença de Chagas
tratamento de IC	filhos de pai	fato de haver
relação com óbito	falta de autoconfiança	tipo de atendimento
lesão isquêmica extensão	menor em grupo	seguimento de puericultura
grau de disfunção	falta de resposta	Freqüência de Ame
defeitos orovalvares congênitos	pacientes com teste	parcela de pacientes
contato com hospital	tempo de evolução	direção de escola
cardiomiopatia dilatada idiopática	importância de desnutrição	avaliação de tratamento
bomba de coração	idade em apresentação	pacientes com idade
atividades físicas comuns	grupo que sobreviveu	condições de vida
acompanhamento de amostra	doença viral prévia	realização de ecocardiograma
Statsoft programa Statistica	lesão de natureza	prevenção de morbimortalidade
IC em repouso	valores de hemoglobina	novas modalidades terapêuticas
Redução de proteína	regressão logística univariada	prevenção de osteoporose
processo de pasteurização	recuperação de hemoglobina	classificação de gravidade
hora de oferta	nível de variáveis	inclusão em pesquisa
cálculos matemáticos específicos	efeito de variável	gênese de problema
amostras de LHP	vida de pessoas	inteligência de crianças
presença de osteopenia	unidade de internação	avaliação de lobo
perda óssea relacionada	unidades de desempenho	diagnóstico de osteomielite
pacientes com osteopenia	locais de assistência	casos de osteomielite
massa óssea relacionada	interrupção de AM	medida de associação
insulina regular utilizada	uso de dieta	políticas de saúde
gênese de osteopenia	uso de esteróides	indivíduos com idade
grupo não osteopênico	tumor intra ocular	óbitos por diarréia
determinação de calciúria	sintomas de retinoblastoma	dosagem de IgG
baixo metabolismo ósseo	retinoblastoma em meio	baixo peso gestacional
Pesquisa de HCPA	estádio de doença	níveis de bilirrubina
Secretaria de Saúde	Classificação de Tanner	pacientes com leucemia
baixa atividade física	evidências de que	critérios de Tanner
aumento de DMO	avaliação de composição	tempo de sobrevida
técnica de medida	desenvolvimento de aterosclerose	referentes a alimentação
suspensão de agentes	portadores de Hib	estudo de Crowcroft
definição de obesidade	ausência de secreções	choro de lactente
existência de doença	diminuição de efeitos	associação entre cólica
percepção de paciente	grupo de puérperas	secreção de orofaringe
densidade de incidência	amostras de colostro	necessidade de identificação
American Heart Association	possibilidade de tratamento	comparação de proporções
integrante de grupos	resultado de RAST	número de RN
critérios para diagnóstico	região de origem	efeitos colaterais graves
cepas bacterianas resistentes	queixa de encaminhamento	Conseqüências de desmame
vômitos com sangue	participação de ácaros	duração de sintomas
vaca de dieta	nível de rejeição	ventilação mecânica invasiva
transmissão oro oral	hipersensibilidade imediata negativo	ordem de nascimento
taxa de infecção	permeabilidade de membrana	serviço de neonatologia
região de antro	grupo mais favorecido	ocorrência de óbito
pangastrite erosiva hemorrágica	procedimentos de reanimação	determinada faixa etária
pacientes com colite	altura de indivíduo	resistência vascular pulmonar
identificação de Helicobacter	momento de consulta	tabelas de NCHS
gastrite por ALV	dor abdominal funcional	risco de uso
episódio de hematêmese	Erradicação do H	viés de causalidade
destino de lixo	Detecção do Helicobacter	registro de medicamentos

Trigramas do *corpus* de Pediatria em ordem de frequência *tf-dcf* — Tabela 11 de 13

pacientes em admissão	epitélio de vaca	esclarecimento de questões
médicos de unidade	crianças amamentadas já	equipe de pesquisadores
itens de prescrição	número de alíquotas	reservatório de oxigênio
diferença estatística entre	gravidade de problema	coletor de drenagem
menor nível socioeconômico	crianças de faixa	circuito de ventilador
importância de aleitamento	início de sucção	meio de contraste
marcador de reabsorção	implementação de IHAC	baixa escolaridade materna
viés de observação	deficiência de crescimento	condição de mamífero
presença de estridor	dieta de lactente	sucção de RN
lesões não relativas	condição socioeconômica desfavorável	responsáveis de pacientes
lesões de gravidade	laboratório de microbiologia	acidente vascular isquêmico
gravidade de desconforto	vida de filhos	ocorrência de complicações
desempenho de escore	serviço de puericultura	atividade física regular
avaliação de estridor	cirurgia de cardiopatia	tromboplastina parcial ativada
avaliação de desconforto	importância de fatores	tempo de trombina
apenas lesões leves	superfície de mucosa	vigência de tratamento
análise de desconforto	desnutrição protéico energética	preparo de medicação
Seqüelas de intubação	contagem de arcos	perda de aerossol
diferente entre grupos	atrofia vilosa parcial	pacientes de NEB
aceleração de crescimento	meses de idade	ministração de medicação
importância de sucção	ocorrência de hemorragia	gravidade de crise
ausência de malformações	até sétimo dia	dose em grupo
redução de hipoxemia	procedimento de intubação	chegada a PA
pronto Atendimento Pediátrico	motivo de intubação	condições de saúde
inalações com broncodilatador	percentual de admissões	técnica de PCR
baixo poder aquisitivo	número de leitos	taxas de S
estatura entre percentis	necessidade de leitos	pacientes com Ome
persistência de canal	média norte americana	episódios de OMA
participação em pesquisa	maior população pediátrica	menor em crianças
média de crianças	assistência intensiva pediátrica	consumo de refrigerante
calazar em IMIP	média de permanência	análise de freqüência
reconhecimento de realidade	choro de filho	duração de choro
hipertensão sistólica isolada	moléculas de adesão	falta de tempo
crianças com IMC	tratamento de BVA	criança com anemia
introdução de medicamento	pacientes com sibilância	achados de exame
método de Ballard	identificação etiológica viral	cirurgia de epilepsia
criança com pais	capacidades de criança	fator de ativação
prevalência de amamentação	referência de NCHS	grupo de tratamento
pressão arterial normal	paciente mais jovem	perda de seguimento
internação em unidade	revisão de prontuários	filha de pais
teste WPPSI R	objetivo de estudar	grau de satisfação
cuidados intensivos neonatais	invasão de estruturas	classificação de doença
crianças de escola	carência de vitamina	descida de leite
diminuição de taxa	teste mais utilizado	crianças com otite
descritas por autores	média desvio padrão	casos de otite
tipo de doença	uso de substância	vírus de imunodeficiência
resposta a imunização	escore de gravidade	relato de uso
Pesquisa de UFMG	manutenção de asma	necessidades de criança
último trimestre gestacional	desconforto respiratório neonatal	mãe de bebê
toxoplasmose em gestantes	significativa entre grupos	encontradas em adultos
soropositividade em gestantes	obesidade de pais	natureza de objeto
proporção de gestantes	risco de pneumotórax	mulheres que iniciam
comparação de curvas	critério de gravidade	intercorrências de mama
porcentagem de gordura	vírus Epstein Barr	fala de entrevistadas
crianças sem lesão	serviço de medicina	diminuição de síntese
uso de corticoterapia	cicatrização de lesão	maioria de pais
dia de avaliação	subpopulações de linfócitos	crianças mais graves
pacientes em faixa	processo de envelhecimento	mal de ausência
uso de enzimas	pesquisa de Id	maior em pacientes
randomização de exposição	pacientes de casuística	tempo de consulta
gordura corpórea percentual	ocorrência de alterações	sorologia para CMV
alto valor energético	critérios de infecção	otite média crônica
severidade de doença	antecedente de sepse	nível socioeconômico inferior
presença de anticorpos	tratamento de FC	mutações mais freqüentes
tratamento de epilepsia	região de coluna	infecção de orelha
indução de resposta	parâmetros de mineralização	idade em óbito
grupo em dieta	aquisição de massa	idade em diagnóstico
contribuição de alimentos	Caracterização de adolescentes	fato de revisões
Dietary Reference Intakes	CMO de coluna	encerramento de estudo
meninas de EA	sucção em seio	eletrólitos em suor
dias de entrevista	presença de hábitos	distúrbio ventilatório obstrutivo
diversas síndromes genéticas	mordida aberta anterior	OMA de repetição
doenças de tireóide	maus hábitos orais	clínica mais rápida
ausência de doenças	hábitos orais nocivos	não apenas linguagem
total de positivos	hábitos orais deletérios	indicadores de aleitamento
prevalência de sensibilização	Mecanismo de sucção	utilização de fármacos
predomínio de sensibilização	características de mães	forma não verbal
padrão de sensibilização	avaliação de sintomas	estenose aórtica grave
níveis de IgE	situações de violência	antiinflamatórios não esteróides
lactentes não amamentados	maioria de adolescentes	Uso de antiinflamatório

Trigramas do *corpus* de Pediatria em ordem de frequência *tf-dcf* — Tabela 12 de 13

uso de aminoglicosídeo	tempo de avaliação	taxa de hospitalização
tronco cerebral normal	rotinas para ajuste	presença de irmãos
tronco cerebral alterados	rotina de assistência	pacientes com cicatriz
otoemissão acústica alterada	publicação com enfoque	lateralidade de RVU
deficiência auditiva neonatal	planilhas como Excel	influência de sexo
atraso em desenvolvimento	perímetro cefálico comprimento	estabelecimento de lesão
resposta de anticorpos	parâmetros não lineares	dependência de oxigenoterapia
tipo de enurese	ganho de perímetro	captação de radionuclídeo
paciente com ENM	dados de crescimento	alta de berçário
disfunção de trato	crianças em acompanhamento	RVU não dilatado
crianças com enurese	comportamento de decréscimo	variedade de condições
complexo esfincteriano uretral	coeficiente de determinação	cuidados em unidade
atividade de complexo	adequação de crescimento	intubação de paciente
incidência de complicações	otites de repetição	instituição de tratamento
dificuldades de sucção	assistência pré natal	uso de opióides
fonte de infecção	tratamento intensivo pediátrico	uso de analgesia
locais mais acometidos	defesa de organismo	Ambulatório de Seguimento
RNs de termo	pequeno para idade	Necrose de pele
freqüência respiratória menor	redução de sedentarismo	alimentação com mamadeira
pacientes com anorexia	diminuição de ingestão	tipo de manifestação
freqüência de depressão	conhecimento de criança	droga de primeira
espectro de autismo	alimentos menos calóricos	tratamento de osteoporose
dor de paciente	gema de ovo	ganho de massa
estudo com crianças	exposição a luz	formação de osso
ausência de amamentação	absorção de cálcio	desenvolvimento de osteoporose
síndrome de ovários	disfunções de cérebro	colágeno tipo I
insulina de jejum	baixa educação materna	medicina pré paga
eclosão de doença	principal manifestação clínica	liberação de leite
diabetes de tipo	funcionamento de paciente	Soma a isto
Estudos em crianças	solução de Ringer	relato de sentimentos
organizações não governamentais	pesquisas clínicas controladas	prescrição de pacientes
precocidade de uso	pacientes com trauma	reflexo de busca
intensidade de dor	osmolalidade de plasma	vida de mulher
brometo de ipratrópio	níveis de cloro	curso de amamentação
qualidade de ingestão	lesão térmica grave	ciclo gravídico puerperal
impacto em mortalidade	insuficiência cardíaca avançada	abandono de amamentação
fase de doença	hipertônica de NaCl	conteúdo de cálcio
doença mais avançada	aumento de osmolalidade	medidas de estatura
confirmação de achados	pacientes com quadro	bolsa auto inflável
sala de aula	maioria de mães	episódio de sepse
transfusão de hemácias	pacientes de UTI	início de alimentação
predominância de sexo	reflexo de sucção	eficácia de lactação
linha axilar média	pressão intra oral	composição de alimentos
número de intubações	movimentos de língua	prognóstico de AVBEH
momento de intubação	oferta hídrica oferecida	principais sinais clínicos
alimentação de lactente	AIG para idade	inclusão citomegálica doença
utilização de medicação	amamentação a seio	etiologia de colestase
médicos com atuação	Estudantes de medicina	ducto biliar comum
minoria de entrevistados	adesão a tratamento	coloração de fezes
grupo de médicos	vacinação contra tuberculose	Haemophilus e Moraxella
dor de neonato	vacina BCG ID	realização de traqueostomia
conhecimento de escalas	uso de vacina	alterações de complacência
atuação em berçário	técnica de aplicação	ausência de dor
introdução de alimentação	tamanho de criança	freqüência de amamentação
atraso de idade	prevalência de tuberculose	exercício de sexualidade
parte de cérebro	positivação de RCHT	Quantidade de alimentos
presença de anemia	leitura de cicatriz	ingestão de calorias
tipo de assistência	exposição a micobactérias	terapêutica anti retroviral
cuidados com filho	aprovação de FDA	população de linfócitos
semana de primeira	administração de BCG	percentuais de linfócitos
incorporação de peso	determinação de morbidade	aplicação de imunobiológico
confluência de curvas	uso de chá	ANOVA de Friedman
inclusão de criança	uso de tocolíticos	observadas em adultos
portador de doença	presença de hemocultura	Mudança de hábitos
motivo de encaminhamento	maior porcentagem de	doenças sexualmente transmissíveis
testes in vivo	escore SNAPPE II	evolução de crianças
soro de paciente	diagnóstico de membrana	tratados com ventilação
número de mamadas	prescrição de medicamentos	subseqüentes comparações múltiplas
adolescentes em Brasil	administração de medicação	nível de hilo
insuficiência cardíaca crônica	ato de brincar	grupos experimentais Dano
tumor de Wilms	instrumentos de investigação	ativação de neutrófilos
regime de condicionamento	hospitais de referência	atenuação de dano
leucemias agudas submetidos	capacidade de síntese	instrumento de avaliação
infusão de células	vítimas de trauma	menores efeitos adversos
doenças mais freqüentes	indicações de intubação	valor de cortisol
TMO em Brasil	entrada de ar	utilização de hidrocortisona
ausência de reflexo	instituição de saúde	trabalho de Annane
gorduras de dieta	possibilidade de diagnóstico	teste de metapirona
início de adolescência	viés de informação	produção de ACTH
óbito durante período	taxas de internação	incidência de insuficiência

Trigramas do *corpus* de Pediatria em ordem de frequência *tf-dcf* — Tabela 13 de 13

estimulação de eixo	técnico previamente treinado	Estudo por imagem
dose de ACTH	só a paciente	classe social alta
cortisol sérico dosado	início de febre	criança em idade
aumento de cortisol	intensidade de febre	Hospital Guilherme Álvaro
adrenal com ACTH	exame de urina	transtornos de espectro
necessidade de assistência	estado infeccioso grave	tratamento de SDR
vida de indivíduo	dose de antitérmico	redução de função
incidência de diabetes	disposição de criança	papel de terapia
consumo de bebidas	bacterioscópico de urina	ocorrência de obesidade
Estudos com adolescentes	Redução de apetite	metabolismo de repouso
eficácia de terapêutica	apresentação para uso	indicação de surfactante
contato com paciente	presença de infecção	

C. Etiquetas Semânticas Atribuídas pelo PALAVRAS

Esse anexo apresenta as 174 etiquetas semânticas utilizadas pelo *parser* PALAVRAS. Essas etiquetas são agrupadas em um nível hierárquico, dito, semântico conforme a Figura C.1.

As Tabelas C.1 e C.2 apresentam as etiquetas atribuídas pelo *parser* dentro dos ramos "Concreto" e "Abstrato", respectivamente. Em ambas tabelas, indica-se a codificação utilizada pelo *parser* (*cod.*), o seu significado em Português (*sig.*) e a qual classe a etiqueta pertence (*classe*). Para cada uma das classes está indicado em negrito a etiqueta que representa genericamente todos os termos que pertencem à classe. Essa etiqueta genérica é atribuída pelo *parser* somente quando nenhuma das outras etiquetas específicas dentro da classe pode ser associada.

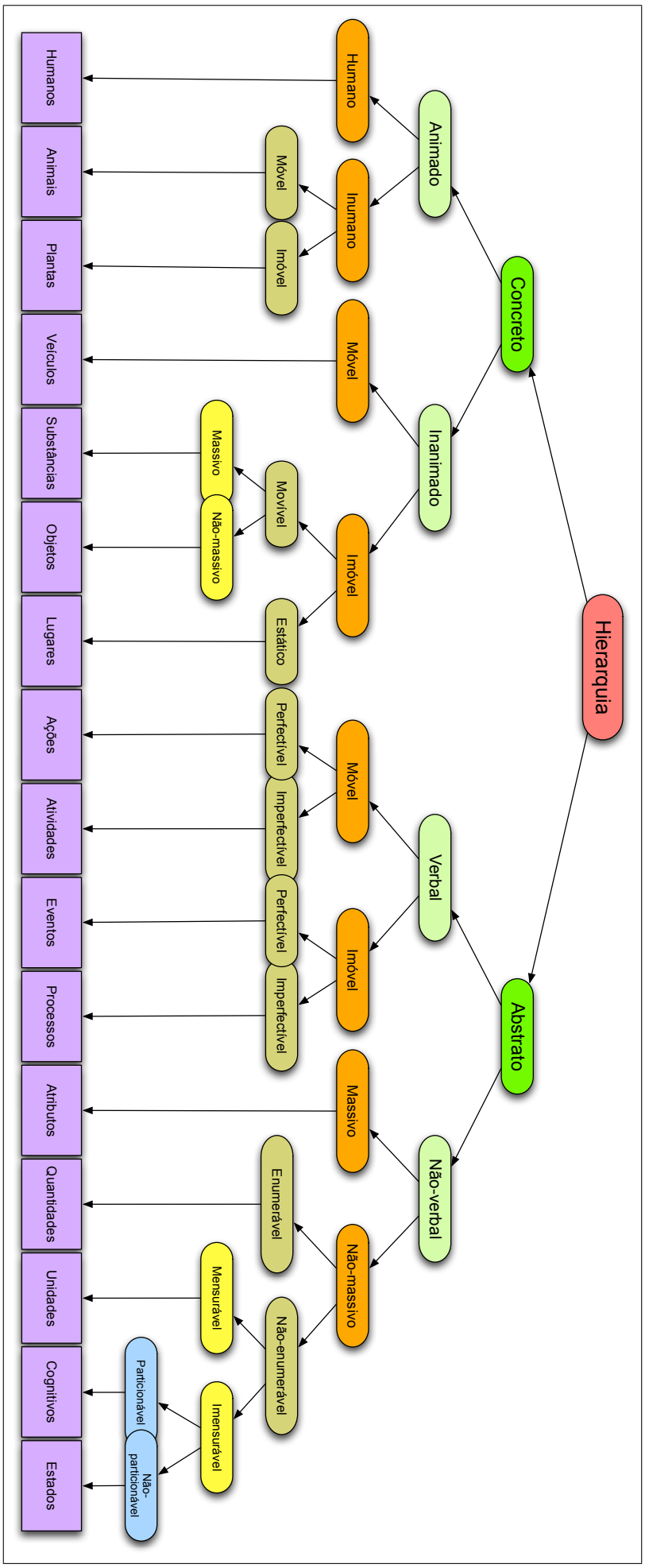


Figura C.1: Hierarquia de classes de etiquetas semânticas encontradas no *parser*.

Tabela C.1: Etiquetas semânticas do ramo Concreto do *parser* PALAVRAS.

cod.	*sig.*	*classe*	*cod.*	*sig.*	*classe*
H	**humanos**	humanos	part-build	partes de construcoes ou veiculos	objetos
HH	grupo de humanos	humanos	coll-cc	coletivos de coisas	objetos
Hattr	humano caracterizado	humanos	cc-h	artefatos	objetos
Hbio	biologicamente humano	humanos	cc-beauty	objetos ornamentais	objetos
Hideo	humano ideologico	humanos	cc-board	objetos planos	objetos
Hmyth	humano mitologico	humanos	cc-fire	objetos de fogo	objetos
Hnat	nacionalidade humana	humanos	cc-handle	objetos manuseaveis	objetos
Hprof	profissao humana	humanos	cc-light	objetos de iluminacao	objetos
Hsick	doente humano	humanos	cc-particle	particulas	objetos
Htit	titulo humano	humanos	cc-r	objetos lisiveis	objetos
A	**animais**	animais	cc-rag	objetos de tecido	objetos
AA	grupo de animais	animais	cc-stone	pedras	objetos
Adom	animal domestico	animais	cc-stick	objetos compridos	objetos
AAdom	grupo de animais domesticos	animais	furn	pele animal	objetos
Aich	animal marinho	animais	mach	maquinas	objetos
Amyth	animal mitologico	animais	con	recipientes	objetos
Azo	animal terrestre	animais	tube	tubos	objetos
Aorn	ave	animais	tool	ferramentas	objetos
Aent	inseto	animais	coll-tool	coletivo de ferramentas	objetos
Acell	animal microscopico	animais	tool-cut	armas brancas	objetos
B	**plantas**	plantas	tool-gun	armas de fogo	objetos
BB	grupo de plantas	plantas	tool-mus	instrumentos musicais	objetos
Btree	arvore	plantas	tool-sail	instrumentos de navegacao	objetos
Bflo	flor	plantas	food	alimentos	objetos
Bbush	arbusto	plantas	food-c	alimentos enumeraveis	objetos
coll-B	coletivo de planta	plantas	food-h	comidas	objetos
an	anotomico	animais	food-c-h	comidas enumeraveis	objetos
amov	anotomico movel	animais	fruit	frutas	objetos
anorg	orgao anatomico	animais	drink	bebidas	objetos
anost	osso	animais	clo	roupas	objetos
anzo	anatomico de animais	animais	cloA	aparatos de animais	objetos
anorn	anatomico de aves	animais	cloH	roupas humanas	objetos
anich	anatomico de animais marinhos	animais	cloH-beauty	aderecos	objetos
anent	anatomico de insetos	animais	cloH-hat	chapeus	objetos
anbo	anatomico de plantas	plantas	cloH-shoe	calcados	objetos
V	**veiculos**	veiculos	L	**lugares**	lugares
VV	grupo de veiculos	veiculos	Labs	lugares abstratos	lugares
Vwater	veiculo aquatico	veiculos	Lciv	aglomeracoes humanas	lugares
Vair	veiculo aereo	veiculos	Lcover	coberturas	lugares
cm	**substancias**	subtancias	Lh	lugares funcionais	lugares
cm-h	artefato	subtancias	Lopening	entradas	lugares
cm-chem	substancia quimica	subtancias	Lpath	caminhos	lugares
cm-gas	substancia gazosa	subtancias	Lstar	astros	lugares
cm-liq	substancia liquida	subtancias	Lsurf	superficies	lugares
cm-rem	remedio	subtancias	Ltip	limites	lugares
mat	substancia material	subtancias	Ltop	lugares geograficos	lugares
mat-cloth	material de vestuario	subtancias	Ltrap	armadilhas	lugares
cc	**objetos**	objetos	Lwater	lugares aquaticos	lugares

Tabela C.2: Etiquetas semânticas do ramo Abstrato do *parser* PALAVRAS.

cod.	*sig.*	*classe*	*cod.*	*sig.*	*classe*
act	**acoes**	acoes	conv	convencoes sociais	processos
act-beat	agressoes	acoes	cord	cordas	processos
act-c	acoes enumeraveis	acoes	cur	moedas	processos
act-d	execucoes	acoes	dir	direcoes	processos
act-s	discursos	acoes	domain	dominio	processos
act-trick	trapacas	acoes	inst	instituicoes	processos
activity	**atividades**	atividades	pos-soc	posicoes sociais	processos
fight	disputas	atividades	sem	produtos semanticos	processos
dance	dancas	atividades	sem-c	produtos cognitivos	processos
sport	esportes	atividades	sem-l	obras musicais	processos
talk	falas	atividades	sem-nons	bobagens	processos
therapy	terapias	atividades	sem-r	obras literarias	processos
dur	duracoes	atividades	sem-s	discursos	processos
event	**eventos**	eventos	sem-w	obras audiovisuais	processos
month	meses	eventos	pict	desenhos	processos
occ	ocasioes	eventos	f	**atributos**	atributos
per	periodo de tempo	eventos	am	massas abstratas	atributos
process	**processos**	processos	f-an	atributos anatomicos	atributos
temp	momentos	processos	f-c	atributos enumeraveis	atributos
percep	perceptiveis	processos	f-h	atributos humanos	atributos
percep-f	sensacoes	processos	f-psych	atributos psicologicos	atributos
percep-l	ruidos	processos	f-q	atributos quantificados	atributos
percep-o	cheiros	processos	f-right	atributos legais	atributos
percep-t	gostos	processos	col	cores	atributos
percep-w	visoes	processos	pos-an	posicoes anatomicas	atributos
wea	climaticos	processos	ac	**enumeraveis**	quantidades
wea-c	climaticos enumeraveis	processos	ac-cat	categorias	quantidades
wea-rain	precipitacoes	processos	ac-sign	simbolos	quantidades
wea-wind	ventos	processos	medida	medida	quantidades
sick	doencas	processos	quantity	quantidade	quantidades
sick-c	sinais	processos	unit	**unidades**	unidades
game	jogos	processos	coll	coletivos	unidades
genre	genero	processos	coll-sem	coletivos semanticos	unidades
geom	geometricos	processos	piece	pedacos	unidades
geom-line	linhas	processos	part	partes	unidades
ism	ideologias	processos	ax	**conceitos abstratos**	cognitivos
ling	linguas	processos	sit	situacoes	cognitivos
meta	meta substantivos	processos	state	**estados**	estados
mon	monetarios	processos	state-h	estados humanos	estados