UNIVERSIDADE FEDERAL DE MINAS GERAIS

ESCOLA DE CIÊNCIA DA INFORMAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO

PROCESSAMENTO DE LINGUAGEM NATURAL: CARACTERIZAÇÃO DA PRODUÇÃO CIENTÍFICA DOS PESQUISADORES BRASILEIROS

Nome: Ana Paula Ladeira

Tese apresentada ao Curso de Doutorado em Ciência da Informação da Universidade Federal de Minas Gerais, como requisito para obtenção do título de Doutor em Ciência da Informação, sob orientação da Prof.ª Dra. Lídia Alvarenga.

Belo Horizonte, Novembro de 2010.

Ana Paula Ladeira

PROCESSAMENTO DE LINGUAGEM NATURAL:

CARACTERIZAÇÃO DA PRODUÇÃO CIENTÍFICA DOS PESQUISADORES BRASILEIROS

Tese apresentada ao Curso de Doutorado em Ciência da Informação da Universidade Federal de Minas Gerais, como requisito para obtenção do título de Doutor em Ciência da Informação.

Área de concentração: Organização e Uso da informação

Orientadora: Prof.ª Dra. Lídia Alvarenga.

Belo Horizonte, Novembro de 2010

Ladeira, Ana Paula.
L154p Processamento de linguagem natural [manuscrito] : caracterização da produção científica dos pesquisadores brasileiros / Ana Paula Ladeira. – 2010.
259 f. : il., enc.

Orientadora: Lídia Alvarenga.
Apêndices: f. 258-259
Tese (doutorado) – Universidade Federal de Minas Gerais, Escola de Ciência da Informação.
Referências: f. 250-257

1. Ciência da informação – Teses. 2. Processamento da linguagem natural (Computação) – Teses. 3. Recuperação da informação – Teses. I. Título. II. Alvarenga, Lídia. III. Universidade Federal de Minas Gerais, Escola de Ciência da Informação.

CDU: 025.4.03

Ficha catalográfica: Biblioteca Profª Etelvina Lima, Escola de Ciência da Informação da UFMG

UFMG

Universidade Federal de Minas Gerais
Escola de Ciência da Informação
Programa de Pós-Graduação em Ciência da Informação

FOLHA DE APROVAÇÃO

"PROCESSAMENTO DE LINGUAGEM NATURAL: CARACTERIZAÇÃO DA PRODUÇÃO CINTÍFICA DOS PESQUISADORES BRASILEIROS"

Ana Paula Ladeira

Tese submetida à Banca Examinadora, designada pelo Colegiado do Programa de Pós-Graduação em Ciência da Informação da Universidade Federal de Minas Gerais, como parte dos requisitos à obtenção do título de **"Doutora em Ciência da Informação"**, Linha de Pesquisa **"Organização e Uso da Informação (OUI)"**.

Tese aprovada em: 05 de novembro de 2010.

Por:

Profa. Dra. Lídia Alvarenga - ECI/UFMG (Orientadora)

Profa. Dra. Marisa Brascher Basilio Medeiros - UnB

Profa. Dra. Ana Maria Rezende Cabral - Profa. Aposentada ECI/UFMG

Profa. Dra. Beatriz Valadares Cendón - ECI/UFMG

Profa. Dra. Marlene Oliveira Teixeira de Melo - ECI/UFMG

Prof. Dr. Renato Rocha Souza - FGV/RJ

Aprovada pelo Colegiado do PPGCI

Profa. Gercina Ângela B. O. Lima
Coordenadora

Versão final Aprovada por

Profa. Lídia Alvarenga
Orientadora

Av. Antônio Carlos, 6627 – Sala 2003 – Campus Pampulha – Cx. Postal 1606 – CEP: 30161-970 – Belo Horizonte – MG
Telefone: (31)3409-6103 – Fax: (31)3409-5207 – www.eci.ufmg.br/ppgci – e-mail: ppgci@eci.ufmg.br

UFMG

Universidade Federal de Minas Gerais
Escola de Ciência da Informação
Programa de Pós-Graduação em Ciência da Informação

ATA DA DEFESA DE TESE DE **ANA PAULA LADEIRA**, matrícula: 2006203007

Às 14:30 horas do dia 05 de novembro de 2010, reuniu-se na Escola de Ciência da Informação da UFMG a Comissão Examinadora aprovada pelo Colegiado do Programa de Pós-Graduação em Ciência da Informação em 21/10/2010, para julgar, em exame final, o trabalho intitulado ***Processamento de linguagem natural: caracterização da produção científica dos pesquisadores brasileiros***, requisito final para obtenção do Grau de DOUTORA em CIÊNCIA DA INFORMAÇÃO, Área de Concentração: Produção, Organização e Utilização da Informação, Linha de Pesquisa: Organização e Uso da Informação (OUI). Abrindo a sessão, a Presidente da Comissão, Profa. Dra. Lídia Alvarenga, após dar conhecimento aos presentes do teor das Normas Regulamentares do Trabalho Final, passou a palavra à candidata para apresentação de seu trabalho. Seguiu-se a argüição pelos examinadores com a respectiva defesa da candidata. Logo após, a Comissão se reuniu sem a presença da candidata e do público, para julgamento e expedição do resultado final. Foram atribuídas as seguintes indicações:

Profa. Dra. Lídia Alvarenga - Orientadora	APROVADA
Profa. Dra. Marisa Brascher Basilio Medeiros	APROVADA
Profa. Dra. Ana Maria Rezende Cabral	APROVADA
Profa. Dra. Beatriz Valadares Cendón	APROVADA
Profa. Dra. Marlene Oliveira Teixeira de Melo	APROVADA
Prof. Dr. Renato Rocha Souza	APROVADA

Pelas indicações, a candidata foi considerada APROVADA.

O resultado final foi comunicado publicamente à candidata pela Presidente da Comissão. Nada mais havendo a tratar, a Presidente encerrou a sessão, da qual foi lavrada a presente ATA que será assinada por todos os membros participantes da Comissão Examinadora.

Belo Horizonte, 05 de novembro de 2010.

Profa. Dra. Lídia Alvarenga
Orientdora - ECI/UFMG

Profa. Dra. Marisa Brascher Basilio Medeiros
UnB

Profa. Dra. Ana Maria Rezende Cabral
Profa. Aposentada - ECI/UFMG

Profa. Dra. Beatriz Valadares Cendón
ECI/UFMG

Profa. Dra. Marlene Oliveira Teixeira de Melo
ECI/UFMG

Prof. Dr. Renato Rocha Souza
FGV/RJ

Obs: Este documento não terá validade sem a assinatura e carimbo da Coordenadora.

Prof. Geruza Angela B. O. Lima
Coordenadora do Programa de Pós-Graduação
em Ciência da Informação - ECI/UFMG

RESUMO

Sinais evidentes de contribuições de grandes campos disciplinares marcaram e têm influenciado fortemente as pesquisas na área de processamento de linguagem natural (PLN), dentre eles a ciência da computação, a ciência da informação e a linguística. Sendo assim, a presente tese pretendeu utilizar o conhecimento acumulado ao longo dos últimos 40 anos em PLN e revelado no ARIST, como referência para selecionar e analisar a produção científica da comunidade acadêmica nacional da área. As publicações nacionais foram coletadas automaticamente da Plataforma Lattes, e um instrumento de seleção automática foi construído a partir da análise de assunto dos artigos de revisão do ARIST. Este instrumento foi utilizado para selecionar, de maneira automática, as publicações nacionais atinentes para a área de PLN. Dentre as 621 publicações consideradas da área, definiu-se o material empírico, constituído por uma amostra de 68 trabalhos, que foi submetido à análise de conteúdo. Essa análise permitiu elucidar as temáticas discutidas pela comunidade científica nacional. Ao analisar todas as publicações atinentes para a área de PLN, observou-se que a grande maioria da produção científica foi publicada depois do ano 2.000. Além disso, a participação da ciência da informação tem sido muito modesta, sendo que a ciência da computação e a linguística foram responsáveis por quase 85% da produção nacional. Doze pesquisadores foram responsáveis por mais de 20% de toda a produção nacional, sendo que dentre eles, nove são da ciência da computação, dois da linguística, e um é da engenharia elétrica. Além disso, vale destacar que dentre esses doze pesquisadores, sete fazem parte do grupo de pesquisa NILC. Dentre as problemáticas mais discutidas, foi possível observar que: a tradução foi intensamente abordada na década de 90; os estudos com indexação diminuíram a partir da década de 80; e que as pesquisas sobre classificação passaram por um período de dormência na década de 90; e que existe uma tendência clara na área de PLN de desenvolvimento de pesquisas em sumarização automática. Outro aspecto que a pesquisa revelou foi que a ciência da informação tem priorizado as pesquisas em indexação automática, seguido da análise de conteúdo, enquanto que a ciência da computação tem priorizado as pesquisas em tradução e sumarização. A análise de conteúdo realizada nas 68 publicações selecionadas permitiu revelou que a recuperação de informação foi a problemática que teve maior destaque na produção científica nacional. Dos trabalhos analisados sobre sumarização, observou-se que somente dois usaram a abordagem profunda e produziram sumários, e que a maioria das pesquisas em sumarização automática tem privilegiado a abordagem empírica (para gerar extratos). As pesquisas em tradução automática têm utilizados métodos estatísticos e regras de transferências, com resultados muito próximos. Apesar das pesquisas em PLN estarem ocorrendo em campos disciplinares diferentes da ciência da informação, os estudos realizados precisam ser conhecidos, pois esta última pode se beneficiar das ferramentas computacionais desenvolvidas, aplicando-as em processos clássicos inerentes ao campo, tais como catalogação, recuperação e representação de informação.

ABSTRACT

Natural language processing researchs (NLP) has being made by researchers from areas as computer science, information science and linguistics. This thesis aims to use the knowledge accumulated over the past 40 years in NLP and published in ARIST, as a reference to select and to analyze the scientific production of the Brazilian academic community in the area. Brazilian publications about NLP were collected automatically from Lattes database (http://lattes.cnpq.br/). The tool for automatic selection of NLP publications from Brazilian Lattes database was built by analyzing the subject of review articles of ARIST. A total of 621 publications were automatically related to NLP area and were retrieved from Lattes database. A randomly sample of 68 papers from this total was submitted to content analysis. This analysis allowed identifying the main issues about NLP discussed by the Brazilian scientific community. We observed that the majority of Brazilian publications were published after the year 2000. Moreover, the participation of information science has been very modest in NLP publication. However, computer science and linguistics were responsible for almost 85% of Brazilian production. Twelve investigators were responsible for more than 20% of all Brazilian production, and among them, nine were from computer science, two from linguistics, and one from electrical engineering. Besides, it is noteworthy that among the twelve main researchers, seven were part of just one research group that works with computational linguistics, the NILC - Núcleo Interinstitucional de Lingüística Computacional (http://nilc.icmc.sc.usp.br/). Among the most discussed issues, we observed the following: translation was discussed intensively in the 90's, indexing studies decreased after the 80's, studies about classification became inactive during the 90's, and there is a clear trend in the area of NLP to develop automatic summarization. Another aspect revealed by the analysis was that information science has focused mainly on automatic indexing and content analysis, while computer science has focused primarily on automatic translation and summarization. The content analysis performed on 68 sample publications showed that retrieval information was the issue most prominent in Brazilian scientific production. Only two papers that worked with summarization used a deep approach to produce summaries. The most research in automatic summarization emphasized on empirical approach to generate extracts. Researches on automatic translation using statistical methods and transfers rules obtained very similar results. Brazilian studies on NLP involve different disciplines from information science. These studies should to be well known by the researchers from information science whose can benefit from the computational tools developed that can be applied in classical processes such as cataloging, information representation and retrieval.

Agradecimentos

É chegada a hora de sair de cena para agradecer os verdadeiros atores principais desta produção. Agradeço a Deus por ter me dado força, saúde e sabedoria para conduzir o meu Doutorado. Agradeço a Profa. Lídia Alvarenga por ter me adotado na essência da palavra: foi muito mais que minha orientadora. Agradeço ao Prof. Renato Souza por ter sido o primeiro a me receber na ECI e me incentivar a integrar a equipe de pesquisadores da ciência da informação. Agradeço a secretaria do PPGCI, em especial a Gisele por se mostrar sempre disponível e pronta pra nos atender. Agradeço a todos os meus colegas da ECI, responsáveis por momentos únicos e inesquecíveis. Agradeço a Daniela Lucas, que foi um anjo que apareceu na minha vida, quando eu mais precisava de uma amiga. Agradeço ao UNI-BH, em especial a coordenação do curso de ciência da computação, pelo apoio e incentivo dado nos últimos anos. Agradeço a todos os meus colegas do UNI-BH pelo carinho, em especial as professoras Miriam Maia e Magali Barroso, que sempre tiveram por mim, um carinho muito além do que o profissional exigiria. Agradeço a toda a minha família pelo apoio incondicional e pelo incentivo dado por toda a minha vida. Agradeço, em especial, aos meus pais e meus irmãos que torcem por mim e comemoram comigo cada conquista. Agradeço aos meus sobrinhos Gil, Livia e Caio por serem fãs da Tipoia, e a grande razão da minha vida. Agradeço ao Bráulio, por ter cruzado o meu caminho e por participar diretamente e intensamente não apenas desta tese, mas de toda a minha vida. Agradeço pelas orientações estatísticas, pela companhia nas noites mal dormidas, pela compreensão quanto às ausências, por me acompanhar e me apoiar, enfim, por estar ao meu lado.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e siglas

AAAI – *Association for the Advancement of Artificial Intelligence*

ACL –*Association for Computational Linguistics*

ACM – *Association for Computing Machinery*

ARIST - *Annual Review of Information Science and Technology*

ASIST - *American Society for Information Science & Technology*

ATN - *Augmented Transition Network*

CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico

EM - *Expectation-Maximization*

HMM - *Hidden Markov Models*

HPSG – *Head-driven phrase structure grammar*

k-NN - *k-Nearest Neighbor*

NILC - Núcleo Interinstitucional de Linguística Computacional

PLN – Processamento da Linguagem Natural

POS – *part-of-speech* (partes do discurso)

RBC – Raciocínio baseado em casos

REM - Reconhecedor de Entidades Mencionadas

RST - *Rhetorical Structure Theory*

SIGIR – *Special Interest Group on Information Retrieval*

SOM - Rede Neural *SelfOrganizing*

SVM - *Support Vector Machine*

TFIDF – *term frequency x inverse document frequency*

TFISF – *term frequency x inverse sentence frequency*

TLG - Teoria do Léxico Gerativo

VT - Teoria de Veins

WER - *word error rates*

SUMÁRIO

1. Introdução

Nas últimas décadas tem-se observado grande aumento na quantidade de informação armazenada e disponibilizada em documentos, principalmente eletrônicos. Acredita-se que, atualmente, grande parte das informações encontradas está no formato textual, tornando fundamental que os mecanismos de análise e processamento sejam focados nesse tipo de informação (BAEZA-YATES; RIBEIRO-NETO, 1999).

No entanto, o grande acúmulo de conhecimento registrado trouxe problemas no acesso e recuperação de documentos nos sistemas de informações documentais. Observa-se que os mecanismos de busca e localização destas informações não tem sido suficientes para resolver esse problema, fazendo com que o usuário se sinta sobrecarregado e perdido diante desse volume de dados e informações (WIVES, 2004).

Esta explosão informacional foi identificada por Vannevar Bush (em 1945) como sendo um problema crítico e fonte de preocupação de várias pessoas. A solução por ele proposta foi utilizar as tecnologias de informação para tornar mais acessível este acervo crescente de conhecimento (BUSH, 1945). Segundo Saracevic (1996), nos anos 50, inúmeros cientistas e pesquisadores começaram a investir no problema e a desenvolver possíveis soluções apontadas por Vannevar Bush. E foi nesta época que o termo Recuperação de Informação (RI) foi introduzido por Calvin Mooers (em 1951) englobando aspectos intelectuais da descrição das informações e da busca, além das máquinas e técnicas utilizadas neste processo. Saracevic (1996) complementa que “a recuperação de informação tornou-se então uma solução bem sucedida encontrada pela ciência da informação e em desenvolvimento até os dias de hoje” (p. 44).

Observa-se, no entanto, que esse grande volume de informação tornou inviável, nos sistemas de recuperação de informações, os processos manuais de indexação, e consequentemente de classificação de documentos. Além disso, diante do aparente esgotamento das estratégias tradicionais de representação e busca de informação em sistemas de recuperação de informação (SOUZA, 2005), a melhoria da eficácia desses sistemas depende dos resultados em várias linhas de pesquisa sobre processos de organização da informação.

E neste sentido, apesar do enfoque principal desta pesquisa ser o processamento da linguagem natural (PLN), em diversos pontos esta se encontrará com o conceito de linguagem documentária, principalmente no que se refere à utilização destas como linguagens de representação de conhecimento.

As linguagens documentárias têm sido utilizadas por unidades de informação para descrever o conteúdo dos documentos. As linguagens documentárias, sejam sistemas de classificação, cabeçalhos de assunto, palavras-chave, lista de descritores ou tesauros, pertencem à mesma família, têm o mesmo objetivo e apresentam várias características em comum. Guinchat e Menou (1994) complementam que as linguagens documentárias são usadas normalmente no momento de entrada de dados dos sistemas de informação, ou seja, “no tratamento intelectual dos documentos” (análise conceitual e tradução) (p. 133). Ainda segundo os autores, os estudos sobre linguagens documentárias privilegiam seus aspectos linguísticos, o que as aproxima das linguagens naturais.

No que tange a linguagem natural, Souza (2005) considera que existem diversas tentativas de se abordar esses processos de representação e recuperação de conhecimento em textos, mas a sua real integração demanda análises concomitantes em diferentes áreas do conhecimento e campos de pesquisa, como a ciência da informação, a linguística, a ciência da computação, a psicologia cognitiva, a comunicação, a sociologia, a antropologia, dentre outras. Saracevic (1996) complementa que “os problemas básicos de se compreender a informação e a comunicação, suas manifestações, o comportamento informativo humano, (...), incluindo as tentativas de ajustes tecnológicos, não podem ser resolvidos no âmbito de uma única disciplina” (SARACEVIC, 1996, p. 48).

Sabe-se que a recuperação, usando linguagem natural, já vem sendo estudada há muitos anos e tem sido o apoio mais concreto para os recentemente criados motores de busca na web. Tais instrumentos foram construídos a partir de abordagens criadas no âmbito das ciências da computação e da informação, em seus primórdios, quando o trabalho de desenvolvimento de pesquisas envolvia pesquisadores de ambas as áreas.

Sinais evidentes de contribuições de grandes áreas marcaram e têm influenciado fortemente as pesquisas na área de processamento de linguagem natural (PLN): a linguística, a ciência da computação e a ciência da informação. Esta

massa crítica formada representa uma considerável contribuição à investigação científica, não apenas quantitativamente como qualitativamente. Contribuições como a gramática de Chomsky, da década de 60, a teoria matemática da comunicação de Shannon & Weaver e o modelo do espaço vetorial de Gerard Salton foram contribuições fundamentais para o desenvolvimento da área. É inegável também a importância do advento das tecnologias, ao longo da década de 80, que permitiram que grandes experimentos fossem realizados em intervalos de tempo menores (influenciados pelos testes executadas pelo *Cranfield Institute of Technology*, em 1957, reconhecidos até hoje pela utilidade e importância).

Embora muito se tenha avançado nesse campo de pesquisa, é fato que ainda há muito por ser feito. Em um primeiro momento, observa-se que, não existem restrições tecnológicas e que o computador permitiu, ou pelo menos seria capaz de permitir, o acesso sem fronteiras, de quaisquer pontos do planeta, aos acervos, não somente de referências, mas também de textos completos, disponibilizados virtualmente. Sendo assim, o grande desafio que se apresenta para os próximos anos é: apesar dos computadores terem evoluído em sua capacidade de armazenamento e rapidez de processamento, os registros de conhecimento continuam sendo inscritos em uma miríade de línguas, transformando o sonho do acesso livre e universal ao conhecimento numa verdadeira metáfora da Torre de Babel.

Ferneda (2003) destaca que "a internet, particularmente a Web, evidencia a dificuldade inata dos computadores no tratamento adequado da informação, na acepção dada ao termo pela ciência da informação" (p. 123). Além disso, ele complementa que essa inabilidade é reconhecida pelos desenvolvimentos recentes da Web em que buscam a criação de novas linguagens que objetivam "uma maior valoração semântica aos documentos da Web". É interessante observar que no projeto da Web Semântica estão inseridos conceitos e idéias que há muito tempo são utilizados pela ciência da informação no tratamento documental (FERNEDA, 2003, p. 123).

> Recuperar informação implica operar seletivamente um estoque de informação, o que envolve processos cognitivos que dificilmente podem ser formalizados através de um algoritmo. [...] a equiparação dos significados supostamente implícitos pelos significantes depende de uma análise intelectual (FERNEDA, 2003, p. 124).

Apesar de Saracevic (1996) afirmar que a base da relação entre a ciência da informação e a ciência da computação está na aplicação dos recursos computacionais na recuperação da informação, assim como nos produtos e serviços, suspeita-se que exista uma distância teórica entre estas ciências, no que se refere às pesquisas desenvolvidas sobre processamento de linguagem natural. Ferneda (2003) afirma que este distanciamento pode ser justificado, num primeiro momento, pelo fato dessas ciências definirem informação de maneira diferenciada. Em uma análise mais aprofundada verifica-se que a informação, objeto de interesse comum de ambas as ciências, é paradoxalmente o que mais as distancia (FERNEDA, 2003, p. 1).

A história do desenvolvimento de uma teoria da informação começou com os trabalhos de Claude Shannon e Warren Weaver, e com a publicação da teoria matemática da informação, em 1949. Eles propuseram um modelo onde um solicitante seleciona uma mensagem, que é enviada por meio de um canal até o receptor. No entanto, vários problemas e ambiguidades têm sido identificados nesse modelo (CORNELIUS, 2002). Dentre os problemas e limites estão questões relacionadas à ignorância do solicitante e do receptor e à capacidade humana nesse processo de comunicação. Saracevic (1999), por exemplo, considera informação como sendo “um *signal* ou uma mensagem para decisão, envolvendo processo cognitivo resultante da interação da mente com o texto, permitindo assim conectar-se com um contexto social. (...) Informação é usada dentro de um contexto e em relações” (p. 397).

A teoria de Shannon & Weaver foi amplamente questionada por semanticistas que a consideraram uma visão simplificada da comunicação, devido ao caráter hermenêutico e interpretativo de todo o processo de transferência, desde a representação até a recuperação da informação. A informação pode ser vista como sendo “o significado de uma mensagem juntamente com um contexto relevante do receptor, (...) e o conteúdo de informação é uma construção subjetiva do receptor” (CORNELIUS, 2002, p. 412). O significado é obtido durante uma interação tendo como base o contexto dos indivíduos. Este conceito é compartilhado por Le Coadic (1996) quando afirma que “a informação comporta um elemento de sentido. É um significado transmitido a um ser consciente por meio de uma mensagem (...)”.

Ainda segundo Cornelius (2002), num processo de transferência, o emissor e o receptor devem ter alguma forma de relação social, e complementa: “para que uma comunicação seja possível, pressupõe-se que exista um sistema social baseado em uma linguagem compartilhada” (p. 403). Outro ponto importante é o conhecimento do receptor no momento que a informação é recebida. Em outras palavras, dependendo do estado de conhecimento existente no receptor, diferentes inferências podem ser feitas a partir de uma mesma informação. Foi baseado nesse pensamento que Brookes, em 1980, propôs a equação fundamental da ciência da informação, que postula o fato de que a informação afeta o estado de conhecimento do receptor de maneiras distintas, dependendo do estado da sua mente. Um dos problemas identificados na equação de Brookes refere-se ao fato de não sabermos como medir o quanto o estado de conhecimento foi alterado diante do recebimento de uma informação: “É difícil entender a transformação da estrutura de informação dentro da estrutura de conhecimento sem uma medida de mudança para avaliá-la” (CORNELIUS, 2002, p. 408).

Além disso, Ferneda (2003) destaca que a utilização de recursos computacionais no tratamento da informação parte de reduções ou simplificações do conceito de informação que na maioria das vezes mostram-se insuficientes para os objetivos da ciência da informação, mesmo quando restrito ao processo de recuperação de informação (p. 122). E complementa que o ato de interpretar uma informação, de forma individual ou coletiva, é dependente da existência de um sujeito, e que os modelos quantitativos desconsideram a presença de tal sujeito (FERNEDA, 2003, p. 123).

Ferneda (2003) destaca ainda que, ao iniciar o seu trabalho de doutorado, ele se perguntava como a ciência da computação poderia contribuir para o avanço da ciência da informação, já que, para ele, muitos recursos computacionais estavam sendo ignorados. Ao final, ele se questionava como a ciência da informação poderia contribuir para o avanço da ciência da computação (p. 125). Além disso, complementa que durante a elaboração do seu trabalho, foram consultadas diversas dissertações e teses em ciência da computação que versam sobre o tratamento da informação textual. Muitas delas mostraram desconhecer até mesmo a existência da ciência da informação, e apresentam como sendo novos, métodos e técnicas que há muito tempo estavam sendo utilizados por esta ciência. Por outro lado, ele destaca

que quando se trata da utilização de métodos computacionais no tratamento da informação, observa-se na literatura da ciência da informação "reações que vão desde o ceticismo até o otimismo exagerado, mostrando também desconhecimento sobre a ciência da computação" (FERNEDA, 2003, p. 125).

Observa-se que a ciência da informação tem uma preocupação mais hermenêutica com a informação, focando assim nos conceitos de significado, contexto, interpretação e representação. Já a ciência da computação, procura e necessita dar um enfoque automatizado para a informação. Assim, a informação precisa ser representada de tal maneira que possa ser posteriormente manipulada e extraída por processos automatizados, o que exige que a mesma seja convertida em alguma estrutura lógica. Mesmo a linguagem natural, considerada uma alternativa ampla e abrangente para representar um determinado conhecimento[1], precisa ser convertida em alguma estrutura computável a partir da qual seja possível extrair conhecimento, sob pena de ser reduzida ou simplificada, conforme discutida anteriormente. Saracevic (1996) complementa que:

> (...) a ciência da computação trata de algoritmos que transformam informações enquanto a ciência da informação trata da natureza da informação e sua comunicação para uso pelos humanos. Ambos os objetos são inter-relacionados e não competidores, mas complementares (SARACEVIC, 1996, p. 50).

No entanto, recente discussão tem apontado para um possível esvaziamento de pesquisas e de produção científica na ciência da informação, tanto no que se refere a recuperação de informação mas, principalmente o processamento de linguagem natural. Esta suspeita pode ser confirmada, num primeiro momento, observando-se o número de capítulos de revisão publicados no ARIST sobre PLN desde a sua criação: na década de 60 foi um artigo por ano, enquanto que, durante toda a década de 90 até então, foram publicados somente dois capítulos de revisão (um em 1996 e outro em 2003).

O tema *processamento de linguagem natural* é, sem dúvida, pertinente para uma área que busca conhecer-se melhor, como pode ser observado em inúmeras publicações que refletem a mesma preocupação (MUELLER E PECEGUEIRO, 2001; PINHEIRO E LOUREIRO, 1995; MÜELLER, CAMPELLO E

[1] Neste momento, não serão discutidas as demais características da linguagem natural como linguagem de representação de conhecimento, tais como ambiguidade e dependência do contexto.

DIAS, 1996; GONZÁLEZ DE GÓMEZ, 2000; MIRANDA E BARRETO, 2000; MÜELLER, MIRANDA E SUAIDEN, 2000; PINHEIRO, 2000).

Vale destacar que toda ciência deve ser cumulativa, derivada e publicada, ou seja, continuar sempre progredindo, utilizando o conhecimento anterior para a produção de novos; partir sempre de algo existente; e finalmente, ser publicada para que o resultado de uma pesquisa possa ser assimilado pela comunidade (ZIMAN, 1979[2] *apud* CAMPOS LEAL, 2005). Para isto, os pesquisadores devem ser exaustivos ao usar referências e citações, uma vez que é a partir delas que se torna possível desenvolver a propriedade cumulativa da ciência (CAMPOS LEAL, 2005).

Diante disso, torna-se fundamental voltar o olhar para os pesquisadores nacionais e contemporâneos e analisar como a temática *processamento de linguagem natural* tem sido abordada, a partir de resultados oriundos de pesquisas realizadas nos últimos anos, em sociedades consideradas mais avançadas e detentoras de frentes de pesquisa nessa área.

Essas contribuições acumuladas historicamente foram identificadas a partir da análise dos capítulos de revisão publicados no *Annual Review of Information Science and Technology (ARIST)*, versando sobre processamento de linguagem natural, ao longo dos últimos 40 anos, ou seja, desde a sua criação em 1966. Para analisar a produção científica nacional e contemporânea, utilizou-se a Plataforma Lattes do CNPq, que permite o acesso aos currículos de todos os pesquisadores associados a entidades de pesquisa, para identificar aqueles que estejam pesquisando sobre o tema foco da presente pesquisa.

Sendo assim, o **objetivo geral** desta pesquisa é utilizar o conhecimento acumulado ao longo dos últimos 40 anos em PLN e revelado no ARIST, como referência para selecionar e analisar as publicações nacionais, identificando assim a produção científica da comunidade acadêmica nacional da área. Em síntese, pretende-se olhar para o passado e revelar “o que" foi desenvolvido ao longo dos últimos 40 anos, e aplicar os resultados dessa observação como parâmetro para analisar a pesquisa nacional.

Dentre os **objetivos específicos**, e considerando como problema principal de pesquisa o processamento de linguagem natural, tem-se:

[2] ZIMAN, John Michael. Conhecimento Público. Ed. Itatiaia, v.8, 164 p., 1979.

1. Construir um instrumento de seleção (critério de atinência[3]) das publicações da área de PLN, a partir dos artigos de revisão do ARIST, publicados nos últimos 40 anos;

2. Selecionar o material empírico a ser analisado tendo como base o instrumento criado anteriormente: produção científica dos pesquisadores brasileiros sobre processamento de linguagem natural;

3. Caracterizar essa produção científica, confirmando a atinência determinada pelo parâmetro criado, e identificando os conceitos inerentes a área de PLN (além dos evidenciados pelos artigos de revisão do ARIST);

Espera-se que esta análise venha a contribuir para pesquisas futuras e, consequentemente, para o desenvolvimento científico da área de processamento de linguagem natural e em recuperação de informação em documentos textuais.

1.1. Organização da tese

No capítulo 2 será apresentado o processo metodológico adotado na presente pesquisa, juntamente com todas as etapas seguidas: na seção 2.1 serão apresentados os critérios usados na construção do instrumento de seleção das publicações da área de PLN; na seção 2.2 serão apresentados os passos realizados para a seleção das publicações nacionais sobre PLN; na seção 2.3 serão apresentadas as categorias de análise usada para examinar a produção científica obtida anteriormente. No capítulo 3, optou-se por apresentar, como referencial teórico, os enunciados extraídos dos capítulos de revisão do ARIST, a partir dos quais o critério de seleção foi construído, visto que essa síntese representa um recorte da literatura analisada neste momento. No capítulo 4, são apresentados os resultados obtidos nesta pesquisa: na seção 4.1 é apresentada a análise horizontal, realizada utilizando-se os títulos das publicações obtidas, enquanto que na seção 4.2 são apresentados os resultados obtidos a partir da análise vertical, obtida adentrando-se no conteúdo das publicações selecionadas. Finalmente, no capítulo 5 é apresentada a conclusão desta pesquisa.

[3] Espera-se que este instrumento seja capaz de indexar as publicações da área de PLN, de acordo com a temática dos documentos (do inglês *aboutness)*. A dificuldade em se definir a atinência de um documento será discutida posteriormente.

2. Metodologia de Pesquisa

O objetivo geral deste trabalho é analisar a produção científica nacional na área de processamento de linguagem natural, a partir do conhecimento revelado ao longo dos últimos 40 anos no ARIST. Diante disso, a seguinte metodologia de pesquisa foi definida: i) construção de um instrumento de seleção das publicações da área de PLN; ii) seleção das publicações nacionais e contemporânea sobre PLN; e iii) análise da produção científica (obtida em ii) e identificação das temáticas da área de PLN reveladas nestas publicações.

A construção do instrumento de seleção se deu através da utilização da análise de assunto, que permitiu extrair conceitos que traduzem a essência dos artigos de revisão do ARIST analisados. Esse instrumento foi utilizado para selecionar, de maneira automática, as publicações nacionais julgadas atinentes para a área de PLN. Este instrumento de seleção tornou-se necessário, visto que na Plataforma Lattes, usada como fonte para obtenção das publicações nacionais, são apresentados somente os títulos das publicações dos pesquisadores, o que impossibilitou, num primeiro momento, que fosse realizada uma análise de conteúdo de todas as publicações recuperadas. Após a obtenção dessa amostragem documental, pôde-se finalmente, por meio de critérios estatísticos e cronológicos, justificar a definição do material empírico destinado à análise final. Essa análise foi realizada utilizando-se técnicas de análise de conteúdo com o objetivo principal de elucidar as temáticas discutidas pela comunidade científica nacional.

Bardin (1977) apresenta a análise de conteúdo como sendo um método empírico, não existindo assim uma metodologia bem formada, mas apenas algumas recomendações. Segundo a autora, a análise de conteúdo apresenta duas funções: função heurística, que enriquece a tentativa exploratória, aumentando a propensão à descoberta "para ver o que dá", e a função da prova, a fim de confirmar uma diretriz pré estabelecida (BARDIN, 1977, p. 30). Neste sentido, a análise de conteúdo pode ser aplicável à presente pesquisa, visto que as duas funções precisaram ser aplicadas de maneira complementar: para confirmar as temáticas relevadas pelo ARIST (função da prova) e para ver o que foi desenvolvido *a posteriori* (função heurística). A FIG. 1 apresenta uma síntese da metodologia adotada nesta pesquisa.

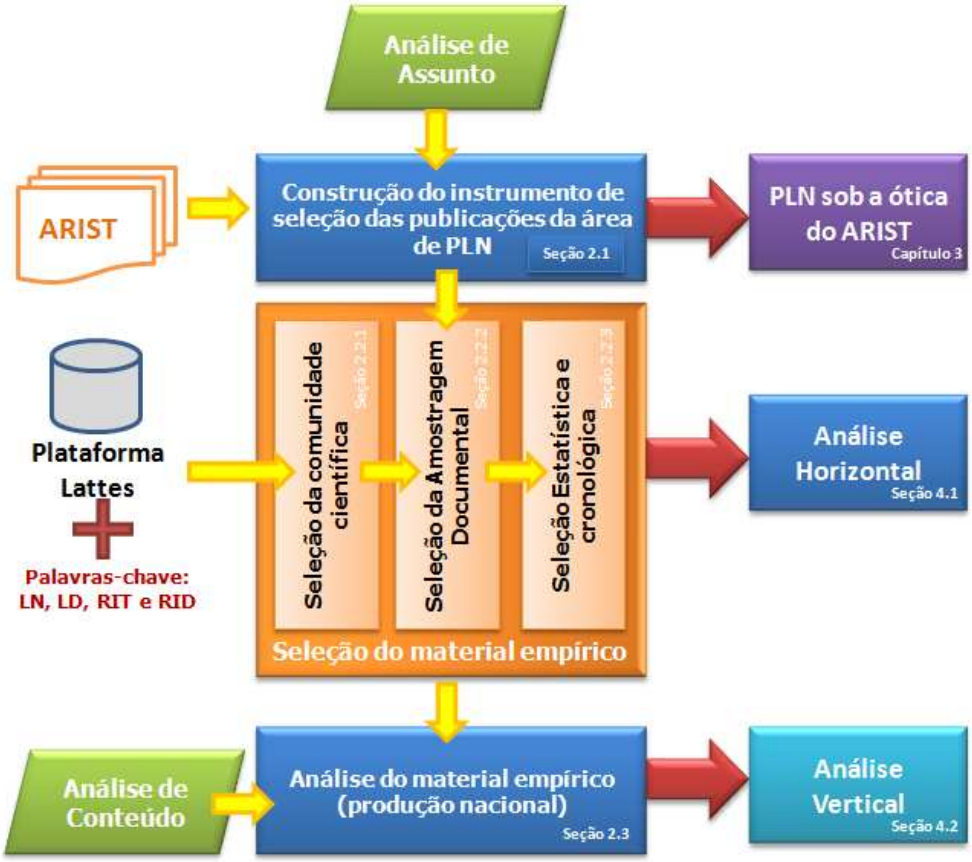

FIGURA 1 – Metodologia adotada na presente pesquisa

Vale destacar que, na FIG. 1 estão ilustradas as etapas do processo metodológico adotado nesta pesquisa, juntamente com as seções deste documento onde as mesmas são discutidas. Nos losangos verdes estão representadas as duas principais técnicas usadas nesta pesquisa: análise de assunto e de conteúdo, que serão detalhadas em momento oportuno. Do lado direito da figura, e obtidos pela seta em vermelho, estão os resultados produzidos: a PLN sob a ótica do ARIST, que apresenta a seleção de enunciados produzida pela análise de assunto (e apresentada no capítulo 3); os resultados da análise horizontal, usando os atributos descritivos das publicações obtidas (apresentados na seção 4.1); e os resultados da análise vertical, obtidos a partir da análise de conteúdo das publicações do material empírico construído (apresentados na seção 4.2).

2.1. Construção do instrumento de seleção

O desafio apresentado na primeira etapa do processo metodológico foi construir um parâmetro conceitual que fosse capaz de indexar publicações da área de PLN utilizando-se apenas os seus títulos. Isso se tornou necessário, visto que a Plataforma Lattes, usada para obter as publicações nacionais, disponibiliza a produção dos pesquisadores cadastrados, na forma de listas de referências bibliográficas. Sendo assim, este parâmetro foi utilizado para verificar, de maneira automática, a atinência de uma determinada publicação à área de PLN baseando-se somente no seu título.

Sabe-se que existem divergências entre os pesquisadores quanto à tradução do termo *aboutness* e que estudos têm tentado formular uma definição conceitual, descrevendo a proximidade com outros termos adotados, tais como assunto, temacidade, tema, tópicos, etc (GUEDES, 2009; FUGITA, 2003). Apesar disso, no contexto desta pesquisa, optou-se por utilizar o termo atinência como sendo a tradução de *aboutness*, uma vez que será empregado para verificar se uma determinada publicação pertence à área de PLN.

O referido instrumento de seleção foi elaborado respaldando-se na garantia literária de uma grande fonte de disseminação científica mundial: o ARIST, periódico escolhido dada a sua importância no panorama da ciência da informação no Brasil e no mundo. O ARIST procura apresentar ao leitor uma revisão geral, analítica, acessível e com autoridade das tendências e desenvolvimentos significativos nas áreas de interesse da ciência da informação. Os tópicos abordados variam de ano para ano, refletindo o dinamismo da disciplina e a diversidade das perspectivas teóricas da ciência e da tecnologia de informação. Apesar de alguns tópicos clássicos continuarem em evidência (bibliometria, recuperação de informação), o ARIST tem ampliado a sua abrangência com o intuito de conectar a ciência da informação a outras comunidades acadêmicas e profissionais. É produzido pela *American Society for Information Science & Technology (ASIST)*, que desde 1937 tem tentado fazer com que profissionais da informação possam pesquisar teorias e técnicas novas, que melhorem a representação e o acesso à informação.

Os capítulos de revisão do ARIST são escritos por especialistas da área, convidados pelos editores, e caracterizam-se por sua capacidade de refletir sobre

um tema em um determinado tempo, adotando uma abordagem horizontal sem aprofundar em especificidade.

Desde sua primeira edição, em 1966, foram publicados 11 (onze) artigos de revisão[1] dedicados ao processamento de linguagem natural, sendo que somente os três mais recentes trazem no título a expressão *"natural language processing"* (os capítulos anteriores eram intitulados *"automated language processing"*). Esses capítulos de revisão foram submetidos à análise de assunto com o intuito de extrair os conceitos que reflitam a essência das pesquisas apresentadas. Os enunciados extraídos desta etapa do processo metodológico da pesquisa são apresentados no capítulo 3.

Observa-se que cada tipo de comunicação cientifíca requer diferentes estilos de construção, que reflitam seus diferentes objetivos e público. O estilo de um artigo científico é formulado, seguindo normalmente um padrão específico, correspondendo às partes clássicas preconizadas para um trabalho científico. Num artigo científico é importante conter informação suficiente para permitir que os leitores possam entendê-lo e eventualmente repetí-lo. Já os artigos de revisão geralmente são mais difusos do que um trabalho inédito de pesquisa e compreendem revisões de trabalhos significativos sobre uma temática e que tiveram impacto na comunidade científica. Acredita-se que o público de um artigo de revisão é maior do que de artigos científicos, e irá abranger do iniciante ao especialista da área. Uma revisão concentra-se estritamente em noticiar os avanços feitos nos últimos anos, sem conter necessariamente o relato de uma pesquisa inédita e novos resultados. Um artigo de revisão é uma fonte secundária porque arrola fontes de outros autores, indicando caminhos e não adentrando em especificidades. Estas características são determinantes num processo de análise de assunto.

Segundo Dias e Naves (2007), a análise de assunto é o processo de ler um documento para extrair conceitos que traduzam a essência de seu conteúdo. Segundo os autores, esta tarefa está sujeita à interferência de diversos fatores ligados à pessoa do profissional que a realiza, como nível de conhecimento prévio do assunto do documento, formação e experiência, subjetividade, além de fatores linguísticos, cognitivos e lógicos (DIAS; NAVES, 2007, p. 9).

Especialistas em recuperação de informação são os primeiros a declarar que a indicação de termos apropriados, capazes de representar o conteúdo de itens

[1] A listagem completa dos artigos de revisão analisados encontra-se no Apêndice A.

de uma coleção são, ao mesmo tempo, a mais importante e a mais difícil de todas as operações normalmente usadas no processamento de informações contidas em documentos (SALTON; McGRILL, 1983[2] *apud* DIAS; NAVES, 2007, p. 30).

A questão da subjetividade presente no processo de indexação é lembrada por vários especialistas da área, pois envolve julgamento, e consequentemente, oscila muito no seu nível de concordância apresentando discrepâncias (DIAS; NAVES, 2007, p. 30).

Diante disso, vale destacar que a análise de assunto dos capítulos de revisão do ARIST foi realizada por uma especialista da computação (doutoranda), priorizando as expressões mais significativas, que pudessem ser usadas para indexar publicações relevantes da área de PLN. Para cada capítulo de revisão, elaborou-se uma estrutura contendo grupos significativos de conceitos com diferentes blocos empíricos baseados principalmente na estrutura do texto adotado por cada autor. Tentou-se, durante esta etapa, realizar uma diagramação mais fiel possível às estruturas dos textos, na tentativa de preservar a classificação de conceitos feita pelos próprios autores. Observou-se que não existe um padrão homogêneo entre os diversos textos, visto que, cada capítulo apresentou uma estrutura de tópicos diferentes, assim com abordagens diferentes, condizentes com o momento focalizado. Dentre os conceitos extraídos, procurou-se identificar objetos de estudo, teorias, processos específicos, produtos e/ou ferramentas, além de outras temáticas apresentadas pelos autores ao longo dos anos. Tais estruturas foram então analisadas com o intuito de organizar todos os conceitos encontrados em grandes categorias, baseando-se na própria classificação feita pelos autores.

A partir dessa análise, quatro grandes categorias foram definidas: a dos conceitos computacionais, a dos conceitos linguísticos, a das aplicações e a das técnicas e métodos. Considerou-se conceitos **computacionais** os termos relacionados a automação, tais como os atributos (automático e computacional), processos (implementação e algoritmos[3]), dentre outros. Dentre os conceitos **linguísticos**, estão os termos gerais relacionados à linguagem e à linguística (linguagem, língua natural, português e inglês), além de elementos a ela relacionados (classes gramaticais – verbo, adjetivo, advérbio), dentre outros. A categoria **técnicas e métodos** foi criada para incorporar os recursos usados pelos

[2] Salton, G. and McGill, M. J. (1983) Introduction to Modern Information Retrieval. McGraw-Hill, New York, NY.
[3] No contexto desta tese, algoritmo foi considerado o processo de construção de uma solução automatizada.

autores na construção das aplicações citadas. Finalmente, nas **aplicações**, procurou-se incluir os produtos resultantes das pesquisas da área de processamento de linguagem natural. Nesta categoria encontram-se os tradutores automáticos, a indexação automática, os respondedores automáticos, a análise de conteúdo e de estilo, além da recuperação de informação em documentos ou textos.

Dois sub-produtos foram gerados, concomitantemente, a partir dessa etapa do processo metodológico: o instrumento de seleção das publicações atinentes à área de PLN, constituído de termos distribuídos nas quatro categorias descritas anteriormente; e os enunciados que emergiram da análise de assunto realizada no ARIST, e apresentados no Capítulo 3. A estrutura de tópicos adotada no referido capítulo permitiu extrair os termos alocados nas quatro categorias definidas no instrumento de seleção. Ao discutir os Aspectos da Linguagem (seção 3.2) procurou-se extrair, dos capítulos de revisão do ARIST, os termos incluídos na categoria de conceitos linguísticos. A partir da discussão apresentada sobre as Teorias Linguísticas (seção 3.3), foi possível identificar as técnicas e métodos relatados pelos autores dos capítulos de revisão do ARIST e compor assim a categoria de mesmo nome. As aplicações da área de PLN foram identificadas e discutidas na seção 3.4 (Aplicações). Os termos incluídos na categoria de conceitos computacionais foram extraídos ao longo de todo o fichamento realizado.

Todas as expressões e termos, identificados a partir da análise de assunto nos artigos de revisão do ARIST, foram utilizados considerando-se também as devidas derivações em gênero e número, quando for o caso, e de idioma, tanto em inglês (idioma usado no ARIST), como em português (idioma predominante nas publicações coletadas na Plataforma Lattes).

Durante a análise de assunto dos capítulos de revisão do ARIST, considerou-se fundamental manter uma relação entre exaustividade e especificidade, visto que o objetivo era indexar publicações atinentes a área de PLN. Assim, procurou-se privilegiar termos indexadores mais genéricos uma vez que assuntos muito específicos tendem a não aparecer nos títulos. Sabe-se que as palavras-chaves definidas com fim da atividade de recuperação, precisam ser determinadas de forma a representar o assunto (WITTEN *et al.*, 1999) e que alcancem os maiores índices de precisão e revocação (BAEZA-YATES; RIBEIRO-NETO, 1999). Sendo assim, alguns conceitos e termos foram descartados, conforme será discutido a seguir.

A expressão "recuperação da informação" foi considerada relevante somente quando ocorrer com algum termo relacionado a informação textual, tais como texto ou documento. Assim, quando aparecer sozinha será descartada. Os termos relacionados à estatística, tais como probabilidade, frequências, dentre outros, também foram descartados, exceto quando ocorresse junto com algum termo relacionado a informação textual. Em outras palavras, procurou-se manter como conceito central os termos relacionados ao processamento de linguagem natural.

Durante a análise de assunto dos capítulos de revisão do ARIST, inúmeras gramáticas foram identificadas, tais como gramática transformacional de Chomsky (encontrada nos artigos dos anos de 1966, 1967, 1968, 1969, 1971, 1973, 1976 e 1987), gramática de casos de Fillmore (1969, 1971 e 1973), gramática de estrutura de frase (*phrase structure grammar*) (1966, 1967, 1968, 1976, 1987), dentre outras. No entanto, utilizou-se como termo indexador apenas o termo gramática (e suas variações de número e idioma). A mesma consideração foi feita para o conceito *parser*, descartando-se os tipos inerentes às gramáticas. Descartou-se também todos os critérios de software, tais como escalabilidade (2003), portabilidade ou transportabilidade (1987, 2003), robustez (1987, 1996), dentre outros encontrados durante a análise desses capítulos, por julgar que os mesmos são muito específicos e como tais tendem a não aparecer no título.

Tendo definido os termos indexadores (listados no Apêndice B), e agrupado-os nas quatro grandes categorias apresentadas anteriormente, o instrumento de seleção das publicações pertinentes à área de PLN foi finalmente construído (FIG. 2) baseando-se nas definições da área apresentadas pelos próprios autores dos capítulos de revisão do ARIST (discutidos na seção 3.1).

Segundo Bobrow *et al.* (1967), a área de processamento automático de linguagem cobre, numa visão ampla, qualquer uso do computador para processar qualquer tipo de linguagem humana. Walker (1973), em seu artigo, engloba todos os estudos teóricos e práticos do uso do computador ou de técnicas computacionais no processamento de linguagem, especialmente a linguagem natural. Becker (1981) define processamento automático de linguagem como sendo a manipulação por computador de dados não-numéricos (normalmente palavras em Inglês). Warner (1987), autor do primeiro capítulo com título “processamento de linguagem natural”, define a área como "uma área de pesquisa e aplicações que exploram como a linguagem natural usada como entrada em sistemas de computadores pode ser

manipulada e armazenada de forma que preserve certos aspectos do original" (WARNER, 1987, p. 79). Segundo Chowdhury (2003), autor do capítulo de revisão mais recente, processamento de linguagem natural é uma área de pesquisa e de aplicação que explora como os computadores podem ser usados para processar e manipular texto ou discurso em linguagem natural para fazer coisas úteis.

Assim, o processamento de linguagem natural pode ser considerado como sendo qualquer utilização do computador para manipular linguagem natural. Desta maneira, para uma publicação ser considerada atinente para a área de PLN, esta deve:

- apresentar no título um conceito computacional juntamente com um conceito linguístico, uma aplicação ou uma técnica ou método, conforme apresentado na FIG. 2, ou;
- ser publicada em evento ou periódico em cujos títulos apresentem termos do conceito computacional, e apresentar no título um conceito linguístico, ou uma aplicação, ou uma técnica ou método.

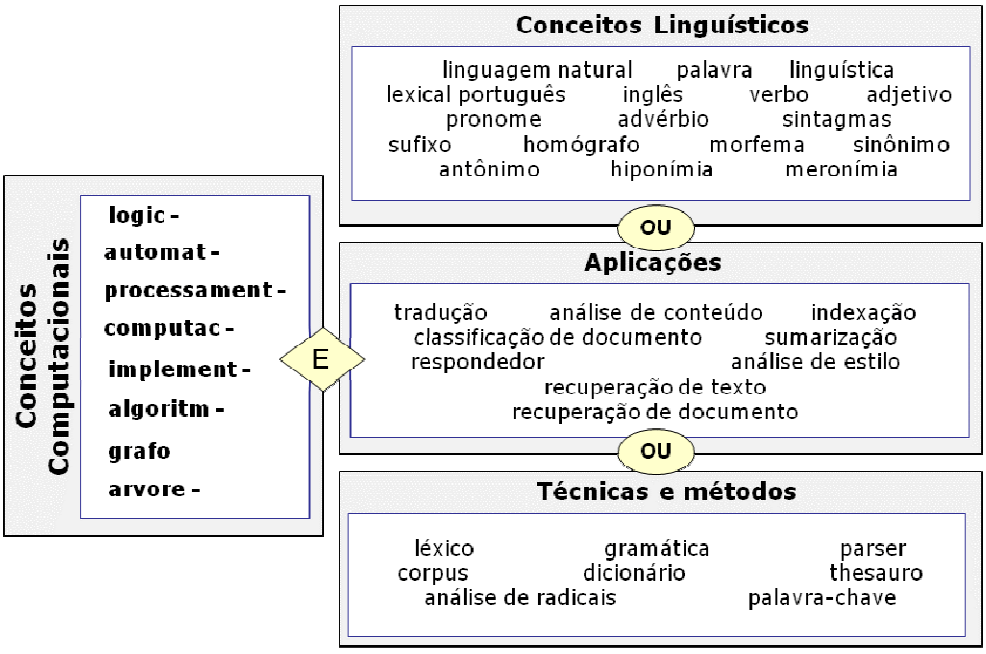

FIGURA 2 – Instrumento de seleção construído a partir da análise de assunto dos capítulos de revisão do ARIST.

Vale destacar que, na FIG. 2, são apresentados somente os radicais dos termos usados para representar a utilização das variações de gênero e número implementadas no escopo desta pesquisa.

2.2. Seleção do material empírico

Tendo definido os termos indexadores que compõem o parâmetro conceitual, a próxima etapa foi obter o material empírico a ser utilizado na última fase desse projeto, ou seja, as publicações nacionais e contemporâneas relacionadas à área de PLN. Para isto, inicialmente, procurou-se identificar a **comunidade científica nacional** (seção 2.2.1), isto é, os pesquisadores cadastrados na Plataforma Lattes, que apresentaram pesquisas na área. Os currículos desses pesquisadores foram analisados com o intuito de selecionar, dentre todas as publicações cadastradas na Plataforma Lattes, aquelas que, de acordo com o parâmetro conceitual construído, fossem consideradas pertinentes à área de PLN. A essa seleção de documentos deu-se o nome de **amostragem documental** (seção 2.2.2). No entanto, tornou-se fundamental avaliar se os critérios de seleção usados foram eficientes o suficiente para recuperar as publicações da área, e ao mesmo tempo, não descartar, erroneamente, publicações atinentes à área de PLN (seção 2.2.2.1). Tendo verificado a validade do critério usado, definiu-se, por meio de **seleção estatística e cronológica** (seção 2.2.3), a amostra de documentos que compôs o material empírico para a análise de conteúdo (seção 2.3), realizada na última etapa desse processo metodológico.

2.2.1. Seleção da comunidade científica

Inicialmente aventou-se a possibilidade de se utilizar o diretório de grupos de pesquisa do CNPq para identificar os pesquisadores na área de processamento de linguagem natural. No entanto, simulações realizadas mostraram que somente pesquisadores das áreas da linguística e da ciência da computação possuem grupos de pesquisa, formalmente cadastrados no CNPq, que abordam processamento de linguagem natural. Em função disto, optou-se por utilizar a Plataforma Lattes do CNPq para identificar os pesquisadores que estejam pesquisando sobre o tema desta pesquisa.

Diante do objetivo desta etapa – identificar os pesquisadores que desenvolveram pesquisas sobre processamento de linguagem natural – utilizou-se como critério de busca, encontrar os pesquisadores doutores que tenham atualizado o seu currículo nos últimos dois anos, tendo como assuntos **linguagem natural** e

linguagem documentária. Esses assuntos foram utilizados, tendo em vista que, conforme mencionado anteriormente, as linguagens documentárias têm sido utilizadas por unidades de informação para descrever o conteúdo dos documentos, e portanto devem ser levadas em consideração num projeto sobre processamento de linguagem natural. Além disso, para verificar a hipótese de que a grande parte das pesquisas sobre linguagem natural está concentrada na ciência da computação, realizou-se também a coleta de currículos a partir da Plataforma Lattes, buscando-se como assunto **recuperação de informação em documentos** e/ou **em textos**. Vale destacar que inúmeras consultas na Plataforma Lattes foram realizadas utilizando-se essas palavras-chaves em português e em inglês, assim como com as devidas variações de número e genêro.

O processo de busca da Plataforma Lattes procura qualquer ocorrência das palavras usadas como critérios de busca no conteúdo do currículo cadastrado pelo pesquisador. Vale ressaltar que todas as informações obtidas na referida plataforma foram fornecidas pelo próprio pesquisador, não sendo foco deste trabalho validá-las.

Diante do volume de dados manipulados durante esta etapa do projeto e dado a formação da doutoranda (ciência da computação), optou-se por desenvolver programas de computador[4] que realizassem as coletas na Plataforma Lattes de maneira automática. Inúmeras coletas foram realizadas durante a realização desta pesquisa, mostrando como a Plataforma Lattes é um ambiente dinâmico no que se refere a atualização dos currículos. Todos os resultados apresentados foram obtidos a partir de dados coletados no dia 10 de novembro de 2009.

Utilizando-se como critério de busca o assunto “linguagem natural” foram retornados 411 currículos[5]; 23 currículos foram retornados utilizando-se como critério “linguagem documentária”[6]; 114 currículos ao utilizar “'recuperação de informação' e documento” como assunto[7]; e 119 currículos utilizando-se “'recuperação de informação' e texto”[8]. Para cada pesquisador, manteve-se um arquivo texto contendo todos os artigos e trabalhos completos publicados pelo pesquisador em toda a sua vida, ou seja, todas as referências que encontram-se cadastradas na Plataforma

[4] Os programas foram desenvolvidos utilizando-se a linguagem de programação Java, que permitiam a coleta e análise (*parsing*) do código HTML de cada currículo retornado.
[5] Esses 411 currículos são listados no arquivo "estatisticasLN.txt" disponível no CD-ROM que acompanha esta tese.
[6] Apresentados no arquivo "estatisticasLD.txt".
[7] Apresentados no arquivo "estatisticasRID.txt".
[8] Apresentados no arquivo "estatisticasRIT.txt".

Lattes[9].

Os resultados dessa seleção totalizaram 541 pesquisadores, sendo que: 95 foram retornados por mais de um assunto: um pesquisador (Nair Yumiko Kobashi da ciência da informação) foi retornado nos quatro assuntos, 29 pesquisadores foram retornados em três assuntos, 65 para dois assuntos e 446 foram recuperados em apenas um assunto[10].

Cada currículo foi avaliado, buscando-se a produção científica de cada pesquisador. Diante do objetivo de identificar as grandes contribuições para a área, descartou-se todas as publicações em eventos de iniciação científica. Assim, para cada currículo recuperado, foram coletados automaticamente a instituição vinculada, a área de atuação e os artigos e trabalhos completos, publicados pelo pesquisador em periódicos e em anais de congressos.

Os 541 pesquisadores coletados foram responsáveis, ao longo de toda a sua vida produtiva, por um total de 27.626 publicações[11]. No entanto, vale destacar que, este montante não é constituído apenas por publicações da área de PLN, e sim por todos os trabalhos publicados pelos pesquisadores que foram recuperados pela Plataforma Lattes. Assim, no sentido de obter as publicações realmente pertinentes a área de PLN, o próximo passo foi descartar as repetições oriundas de co-autorias, para então aplicar o critério de seleção. Neste momento, algumas dificuldades foram encontradas devido à inconsistência nos dados cadastrados na Plataforma Lattes. Essas divergências dificultaram uma análise baseada na referência bibliográfica, visto que para que um procedimento automático de casamento de padrões funcionasse, era necessário que o mesmo trabalho fosse referenciado nos currículos de todos os seus co-autores da mesma maneira, o que não acontece. No entanto, conforme mencionado anteriormente, não cabe a este trabalho validar as entradas e a consistência na normalização das referências apresentadas pelos pesquisadores, e sim considerá-las como sendo uma situação possível de existir.

Essas dificuldades foram antecipadas por Mascarenhas Silva (2007), em sua tese de doutorado, quando advertiu que muitos cientistas não se dão conta que sua produção documental poderia servir, no futuro, como objeto de estudo e que esta produção permitiria estudar "a evolução das políticas de pesquisa e ensino

[9] Estes arquivos criados para cada pesquisador retornado pelas consultas realizadas na Plataforma Lattes estão disponibilizados no diretório "todosCurriculos".
[10] Disponibilizado no arquivo "estatisticasPesqAssunto.txt".
[11] Todas as publicações são listadas no arquivo "todasPublicacoes.txt".

científicos, a evolução desta ou daquela disciplina ou ainda o papel deste ou daquele cientista no desenvolvimento da ciência" (p. 78).

Das 27.626 publicações coletadas originalmente, 337 foram citadas da mesma forma nos currículos dos seus co-autores e portanto puderam ser descartadas por um processo de casamento de padrões[12].

Esta divergência dificultou a análise de co-autoria, de maneira automática, uma vez que um mesmo trabalho pode ser cadastrado na Plataforma Lattes com referências diferentes. Por exemplo, o trabalho “Desafios do Processamento de Línguas Naturais” publicado por Vera Lúcia Strube de Lima, Maria das Graças Volpe Nunes e Renata Vieira no 34° Seminário Integrado de Software e Hardware em 2007 foi cadastrado de maneira diferente nos três currículos (citando os autores em ordem diferente).

Mesmo utilizando-se a forma de citação, cadastrada pelo próprio pesquisador no seu currículo, observou-se que tal forma era adotada somente no próprio currículo, mas as co-autorias eram cadastradas sem seguir o padrão. Assim, todas as publicações de um único autor eram cadastradas seguindo o formato cadastrado, mas para as demais, fez-se necessária a utilização de uma heurística (comparando as iniciais dos nomes). Das 2.338 co-autorias analisadas, 334 (cerca de 14%) foram identificadas por meio desse processo heurístico.

Além disso, 88 publicações tiveram que ser corrigidas manualmente, pois apresentavam o mesmo título e o mesmo ano de publicação mas apresentavam autores diferentes. Para essas inconsistências sinalizadas pelo programa de computador desenvolvido, a doutoranda verificou na internet, no site do evento ou do periódico, qual seria a correta autoria do trabalho, e corrigiu manualmente.

Ao final desse processo de análise de co-autorias, a lista de pesquisadores, que inicialmente era de 541, passou a contar com 1.209 co-autores[13]. Considerando-se somente as publicações potencialmente atinentes, de acordo com o critério de seleção definido, são 1.003 pesquisadores envolvidos[14]. Para estes pesquisadores, que não foram obtidos pela consulta na Plataforma Lattes, e sim pela análise de co-autoria de alguma publicação, atribuiu-se à área e macro-área como sendo 'desconhecida'. Estes pesquisadores não foram recuperados pelas consultas na Plataforma Lattes, por um dos seguintes motivos: ou

[12] As 27.289 publicações restantes estão apresentadas no arquivo "todasPublicacoesSemRep.txt", sem as repetições.
[13] Esta lista de 1.209 co-autores está apresentada no arquivo "saidaPesquisadoresRelevantes.txt".
[14] Disponíveis no arquivo "saidaPesquisadoresSomenteAtinentes.txt".

não é doutor, ou não tem atualizado seu currículo, ou não tem currículo cadastrado na Lattes. Vale destacar que, destes 668 pesquisadores desconhecidos, 448 foram associados a somente uma publicação, 127 a somente duas publicações, 54 a três publicações e 20 a quatro publicações. Vale ressaltar que, dentre os pesquisadores desconhecidos que apresentaram mais de 5 publicações estão José Gabriel Pereira Lopes e Ricardo Baeza-Yates, que publicam muito na área de recuperação de informação em documentos mas que por serem, Português e Chileno, respectivamente, não apresentam currículos cadastrados na Plataforma Lattes.

Conforme destacado anteriormente, a Plataforma Lattes retorna os currículos que apresentaram os assuntos pesquisados (*linguagem natural, linguagem documentária e recuperação de informação em documentos e/ou em textos*), não garantindo que estes currículos, assim como as publicações obtidas, sejam realmente pertencentes à área de PLN.

O parâmetro conceitual construído foi aplicado em todas as publicações para que fosse possível analisar historicamente a produção nacional. Além disso, como os conceitos usados como critérios de seleção foram estabelecidos, tendo como base os capítulos de revisão do ARIST (de 1966 a 2003), aplicá-los somente às publicações recentes poderia gerar um retrato distorcido da realidade.

Sendo assim, o instrumento de seleção definido anteriormente, a partir da análise de assunto do ARIST, foi utilizado para determinar quais destas publicações eram realmente consideradas atinentes à temática processamento de linguagem natural.

2.2.2. Seleção da amostragem documental

O processo de coleta dos currículos na Plataforma Lattes, assim como o de aplicação do parâmetro conceitual – utilizado como instrumento de seleção das publicações – foram realizados usando programas de computadores desenvolvidos pela doutoranda na linguagem de programação Java, especialmente para o escopo deste trabalho. Tanto a coleta das publicações, como o processamento das mesmas, foram feitos de maneira automática com o intuito de minimizar a subjetividade e a interferência do avaliador humano envolvido, visto que a doutoranda tem a ciência da computação como área de formação.

Os títulos de todas as publicações coletadas na Plataforma Lattes foram analisados segundo os termos e expressões contidos no parâmetro conceitual definido anteriormente (e apresentado na FIG. 2). Sabe-se que avaliar a atinência de uma publicação para a área, baseando-se somente no título, pode ser considerada uma estratégia frágil. No entanto, o volume de publicações obtidas (mais de 27 mil), assim como o meio de registro – somente a referência bibliográfica cadastrada na Plataforma Lattes – impossibilitou que outra abordagem fosse adotada. Assim, assumindo-se que o título represente uma condensação dramática de um conteúdo que, pode estar contido ou desenvolvido em centenas de páginas (DIAS e NAVES, 2007, p. 53), acredita-se que ele possa ser usado numa primeira análise, com as devidas restrições.

Conforme mencionado anteriormente, o processo automatizado de coleta dos curriculos dos 541 pesquisadores recuperados na Plataforma Lattes, obteve 27.626 publicações, sendo que 482 foram descartadas: 334 por serem referências duplicadas (em função de co-autorias) e 148 por, apesar de serem referências escritas de maneira diferentes, era o mesmo título, publicado no mesmo evento ou periódico, pelos mesmos autores. Assim, descartando-se as publicações repetidas, obteve-se um universo de 27.144 publicações.

Aplicando-se o critério de seleção automática, definido a partir da análise de assunto do ARIST, 831 foram consideradas potencialmente atinentes, enquanto que 26.313 foram desconsideradas[15]. Diante disso, tornou-se fundamental avaliar a capacidade do critério utilizado em indexar as publicações da área.

2.2.2.1. Avaliação do critério de seleção automática

O próximo passo foi avaliar a qualidade do critério de seleção automática definido a partir da análise de assunto do ARIST. O critério deveria ser eficiente o suficiente para recuperar as publicações da área, e ao mesmo tempo, não descartar publicações atinentes à área de PLN. Sendo assim, a validação do processo de seleção foi realizada por meio de julgamento humano, analisando-se manualmente, tanto títulos de publicações que foram selecionadas, como de outras que foram descartadas.

[15] Estas listagens encontram-se, respectivamente, nos arquivos "saidaPublicacoesRelevantes.txt" e "saidaPublicacoesNaoRelevantes.txt".

Para a análise das 831 publicações selecionadas de forma automática, optou-se por fazer um censo, quando todos os títulos destas publicações foram avaliados manualmente pela doutoranda. Ao analisar manualmente essas 831 publicações, pode-se constatar que 31 eram publicações repetidas, e por apresentarem pequenas diferenças nos titulos, não foram identificadas pelo processo automático. Assim, descartando-se as repetições, e avaliando-se então 800 títulos, 621 foram identificados como sendo realmente de publicações pertinentes à área de PLN, e 179 não eram atinentes. Estes resultados estão sintetizados na TAB. 1.

Para as 26.313 publicações que foram descartadas pelo processo automático de seleção, optou-se por uma análise por amostragem, já que o volume dessas publicações inviabilizou a leitura de todos os títulos envolvidos. No cálculo do tamanho da amostra, considerou-se que, no máximo 5% das publicações teriam sido descartadas erroneamente (p=0,05). Se for usada uma margem de erro de 2% (E=0,02), o cálculo do tamanho da amostra foi feito por (HULLEY *et al.*; 2006):

$$n_1 = \frac{4 \times (z_{\alpha/2})^2 \times p \times (1-p)}{(2 \times E)^2}$$

Onde:

- n_1 = tamanho da amostra considerando populações de tamanho infinito
- $Z_{\alpha/2}$ = 1,96 (para intervalos de 95% de confiança)
- E = margem de erro da estimativa = 0,02
- p = proporção esperada de relevância = 0,05

$$n_1 = \frac{4 \times (1{,}96)^2 \times 0{,}05 \times (1-0{,}05)}{(2 \times 0{,}02)^2} = 456$$

Com base na fórmula anterior, n_1 = 456. Agora, considerando que a população amostrada é de tamanho finito (N = 26.313), o tamanho da amostra deve

ser ajustado pela equação abaixo:

$$n = \frac{N \times n_1}{N + n_1}$$

Onde:

- n = tamanho da amostra para pesquisas em populações finitas
- n_1 = tamanho da amostra considerando populações de tamanho infinito
- N = tamanho da população amostrada

$$n = \frac{26.313 \times 456}{26.313 + 456} = 448$$

Sendo assim, dentre as 26.313 publicação que foram descartadas pelo processo de seleção, definiu-se selecionar aleatoriamente uma amostra de 448 publicações para analisar os títulos manualmente por meio de julgamento humano[16]. Após análise manual, verificou-se que apenas 11 publicações eram pertinentes a área de PLN e como tal deveriam ter sido selecionadas pelo parâmetro conceitual. As demais 437 publicações realmente deveriam ter sido descartadas (TAB. 1).

TABELA 1
Resultados obtidos pela avaliação manual dos títulos das publicações selecionadas pelo critério de seleção criado.

Algoritmo recuperou o documento?	O documento é realmente atinente (julgamento humano)?		Total
	Sim	Não	
Sim	621	179	800
Não	11	437	448
Total	632	616	

Estes resultados apontam para uma taxa de 98% de sensibilidade[17] (621/632) e 71% de especificidade (437/616). Em outras palavras, dado que uma publicação é pertinente, o critério de seleção automática tem 98% de chance de selecioná-la, enquanto que dado que não é pertinente à área, tem-se 71% de

[16] A listagem das publicações descartadas pelo critério de seleção automática e analisadas manualmente está no arquivo "saidaPublicacoesNaoRelevantes.xls".
[17] Sensibilidade, especificidade, predição positiva e predição negativa são métricas estatísticas usadas para avaliar a qualidade de testes diagnósticos.

chance de não selecioná-la. É interessante observar que o método de seleção automática prioriza a sensibilidade, ou seja, se um documento for atinente à área de PLN, ele tem grande chance de ser selecionado (98%), evitando-se assim descartar qualquer documento que seja importante para caracterizar a área.

Uma outra observação interessante refere-se a suposição que foi feita para o cálculo do tamanho da amostra (n) de publicações descartadas pelo critério de seleção automática. Para o cálculo de n, supôs-se que, no máximo, 5% das publicações teriam sido descartados erroneamente. Após a seleção e análise da amostra, verificou-se que somente 2% das publicações (11/448) tinham sido descartadas erroneamente, o que confirma a suposição feita para o cálculo de n.

Além da sensibilidade e especificidade, os dados da TAB. 1 permitem o cálculo dos valores de predição positiva e negativa do critério de seleção automática de artigos. O valor de predição positiva foi de 78% (621/800), isto é, um artigo selecionado de forma automática tem 78% de chance de realmente ser pertinente à área de PLN. Já o valor de predição negativa foi de 98% (437/448), o que reforça a boa triagem feita pelo algoritmo: uma publicação descartada de forma automática tem 98% de chance de realmente não ser da área de PLN.

Diante disso, as 621 publicações realmente atinentes podem ser usadas para representar a produção científica nacional na área de PLN.

Assim, análises estatísticas foram realizadas a partir das referências destas 621 publicações, com o intuito de identificar a distribuição dessa produção por área dos autores, por ano, por temática, dentre outros resultados apresentados na seção 4.1 – Análise Horizontal[18]. Esta análise foi intitulada de horizontal por ter sido realizada baseando-se apenas nas características descritoras das publicações (título, ano de publicação, autores e áreas de vinculação).

Para que fosse factível adentrar no conteúdo das publicações foi necessário identificar, dentre as 621 relevantes, as que seriam submetidas à análise de conteúdo (considerada Análise Vertical). Assim, na próxima seção, serão apresentados os critérios estatísticos e cronológicos usados para definir o material empírico usado na última etapa desta pesquisa, ou seja, as publicações que serão submetidas à análise de conteúdo. Os resultados alcançados são apresentados na seção 4.2.

[18] Optou-se por não incluir nas análises, as 11 (onze) publicações consideradas atinentes, dentre as que foram descartadas, diante dos índices de sensibilidade e especificidade obtidos pelo critério de seleção automático adotado neste trabalho.

2.2.3. Seleção estatística e cronológica

Conforme discutido anteriormente, a coleta automática na Plataforma Lattes recuperou uma amostra de 541 pesquisadores que, juntos, foram responsáveis por 800 publicações potencialmente atinentes, de acordo com o instrumento de seleção criado a partir da análise de assunto dos capítulos de revisão do ARIST. Analisando-se manualmente os títulos dessas 800 publicações, observou-se que 621 eram realmente atinentes à área de PLN, e que portanto poderiam ser usadas para caracterizar a produção nacional da área. Estas publicações deveriam ser analisadas, no entanto, em função do grande número de artigos envolvidos, optou-se por utilizar uma amostragem.

Para o cálculo do tamanho da amostra de publicações a serem avaliadas, considerou-se que dos 621 trabalhos definidos como atinentes à área de PLN, pelo menos 95% são realmente pertinentes e consequentemente têm informações e conteúdos suficientes para a caracterização da produção científica e nacional da área de PLN. Sob esta hipótese e usando uma margem de erro de 5%, obtêm-se o seguinte tamanho de amostra (HULLEY *et al.*, 2006):

$$n_1 = \frac{4 \times \left(z_{\alpha/2}\right)^2 \times p \times (1-p)}{\left(2 \times E\right)^2}$$

Onde:

- n_1 = tamanho da amostra considerando populações de tamanho infinito
- $Z_{\alpha/2}$ = 1,96 (para intervalos de 95% de confiança)
- E = margem de erro da estimativa = 0,05
- p = proporção esperada de conteúdo relevante = 0,95

$$n_1 = \frac{4 \times (1{,}96)^2 \times 0{,}95 \times (1-0{,}95)}{(2 \times 0{,}05)^2} = 73$$

Com base na fórmula anterior, n_1 = 73. Considerando que a população amostrada é de tamanho finito (N = 621), então o tamanho da amostra deve ser

ajustado pela equação abaixo:

$$n = \frac{N \times n_1}{N + n_1}$$

Onde:

- n = tamanho da amostra para pesquisas em populações finitas
- n_1 = tamanho da amostra considerando populações de tamanho infinito
- N = tamanho da população amostrada

$$n = \frac{621 \times 73}{621 + 73} = 65$$

Tem-se então que, aproximadamente 65 publicações devem ser submetidas à análise de conteúdo. Definido o tamanho da amostra, o próximo passo é definir a forma de se obter a amostra, ou seja, o processo de amostragem. Nesta pesquisa, optou-se por obter uma amostra aleatória estratificada, onde "os elementos são divididos em grupos mutuamente exclusivos e dentro dos quais são sorteadas amostras casuais simples" (SILVA, 1998).

Analisando as 621 publicações atinentes ao longo dos anos, foi possível observar que nos últimos onze anos, ou seja de 1.999 até 2.009, foram publicados 75% de todas as publicações consideradas atinentes. Para os anos anteriores a 1.999, tem-se apenas 25% das publicações relevantes. Em função deste resultado, optou-se por uma amostragem estratificada por estes dois períodos, de forma proporcional. Ou seja, 75% da amostra, equivalente a cerca de 50 artigos, seriam selecionados de forma aleatória do período de 1.999 a 2.009. No período anterior a 1.999, seria sorteada 25% da amostra, equivalente a aproximadamente16 artigos.

Como o período de 1.999 a 2.009 equivale a 11 anos, optou-se por sortear 5 artigos de cada um destes anos, o que totaliza **55 trabalhos**. No outro período, anterior a 1.999, existem 21 anos, sendo que apenas 13 apresentaram dois ou mais artigos atinentes. Neste período, optou-se por sortear um artigo em cada ano que apresentasse duas ou mais publicações, totalizando **13 artigos**. Desta forma, a amostra analisada, de tamanho n = (55 + 13) = 68, teria uma maior

representatividade, por possuir publicações sorteadas de cada ano envolvido na pesquisa, numa proporção diretamente relacionada à importância do período em termos de número de publicações atinentes a área de PLN.

Mais uma vez, foi feito um processo automático que dentre uma amostra finita de dados sorteia n elementos (sorteio automático). Assim, para os anos de 1.999 a 2.009 foram sorteados 5 publicações, enquanto que, para os demais anos, apenas uma era sorteada.

Conforme mencionado anteriormente, os resultados desta pesquisa foram divididos em dois momentos. No primeiro, foram apresentados os resultados obtidos analisando-se todas as 621 publicações consideradas atinentes para a área de processamento de linguagem natural. Como esta análise foi realizada considerando-se as informações descritivas das publicações (título, autores e áreas) optou-se por referênciá-la por **análise horizontal ou superficial**. Dentre os resultados apresentados, estão análises estatísticas envolvendo a distribuição dos pesquisadores autores dessas publicações por área de vinculação, por produção científica, por temáticas ao longo dos anos, assim como grupos de pesquisas. A segunda parte dos resultados, contém as discussões que emergiram durante a análise de conteúdo dos artigos selecionados. Tendo em vista que esta análise permitiu adentrar no conteúdo das 68 publicações sorteadas, optou-se por chamá-la de **análise vertical ou profunda.** As categorias de investigação usadas durante a análise de conteúdo das publicações selecionadas são apresentadas na próxima seção (seção 2.3).

2.3. Análise de conteúdo do material empírico

Tendo definido a amostra de publicações que serão analisadas, a próxima etapa desta pesquisa consistiu em submetê-las à análise de conteúdo. Segundo Bardin (1977), para a realização da análise de conteúdo, algumas categorias e subcategorias relacionadas ao objeto de pesquisa podem ser estabelecidas antes do processo propriamente dito, ou a medida que vão sendo observadas (p. 123). Ainda segundo Bardin (1977), é realizar a leitura “pra ver no que vai dar” (p. 20).

Inicialmente, os artigos foram analisados com o intuito de analisar as problemáticas discutidas pelos autores, a metodologia adotada e os resultados

alcançados. No entanto, durante a leitura, e tendo em vista que a presente pesquisa tem como objetivo analisar a produção científica da comunidade acadêmica nacional, observou-se que seria interessante analisar também se o trabalho apresentava experimentos práticos e para qual idioma o mesmo estava voltado. Neste momento, não nos interessa identificar o idioma no qual o artigo estava escrito, e sim o idioma usado durante os experimentos, se este for o caso. Além disso, outro aspecto que emergiu da leitura está relacionado com os métodos de avaliação usados como parte da metodologia do trabalho. Para alguns artigos, observou-se que os autores usaram estratégias automáticas para avaliar o trabalho, enquanto que outros recorreram a julgamento humano para avaliar os resultados alcançados. Vale ressaltar que, métodos quantitativos de avaliação tendem a ser mais aplicáveis a trabalhos que apresentem experimentos, conforme será apresentado posteriormente. Em outras palavras, pode-se afirmar que as categorias utilizadas nesta pesquisa são o resultado de um processo incremental e cíclico, visto que algumas categorias foram emergindo durante a leitura dos artigos. Esta dinâmica acarretou na re-leitura de vários artigos, que já haviam sido analisados, para garantir que todos os selecionados fossem analisados sob as mesmas dimensões. Desta maneira, as categorias de análise definidas, antes e durante a leitura realizada, e utilizadas finalmente nesta pesquisa, são discutidas a seguir e apresentadas na FIG. 3.

FIGURA 3 – Categorias de análise usadas durante a etapa de análise de conteúdo das publicações selecionadas.

Para cada artigo analisado, procurou-se apresentar o artigo, identificando **a problemática abordada**, ou seja, o conjunto de problemas tocantes ao trabalho, destacando os objetivos propostos pelos autores. Com isso, espera-se observar quais as temáticas foram discutidas ao longo dos anos. O objetivo final desta dimensão é identificar os problemas recorrentes pesquisados pela comunidade científica nacional; **(O QUÊ)**

Para cada artigo analisado, procurou-se identificar, **a metodologia adotada**, ou seja, **os métodos e técnicas utilizados** durante a realização do trabalho. Esta categoria tem como objetivo revelar o que os pesquisadores da área tem usado em termos de ferramentas, tanto computacional como linguística, para resolver os problemas apresentados; **(COMO)**

Procurou-se identificar também, para cada artigo analisado, o objeto empírico focalizado, ou seja, o **material empírico utilizado** na realização dos trabalhos. Espera-se com esta análise, identificar, dentre outras coisas, se os trabalhos da área têm apresentado experimentos práticos, ou se são de cunho teórico; se os trabalhos têm priorizado algum idioma; e finalmente, se a comunidade científica nacional foi capaz de criar um *framework* de pesquisa na área de PLN. Entende-se por *framework* como sendo um arcabouço experimental a partir do qual as pesquisas podem ser desenvolvidas. Neste arcabouço devem estar ferramentas desenvolvidas, assim como bases de documentos disponibilizadas (*corpus*) para a comunidade científica. Esta categoria surgiu durante a análise dos artigos, ao observar que alguns artigos, principalmente os mais recentes vem utilizando e re-utilizando ferramentas e corpus desenvolvidos e disponibilizados por outros pesquisadores, em trabalhos anteriores.

E finalmente, procurou-se identificar os **resultados** apresentados pelos autores, extraindo as perspectivas dos autores quanto ao desenvolvimento da área. Além disso, tendo em vista a dificuldade em se avaliar os resultados, procurou-se identificar se foram usados métodos de avaliação automáticos, ou se o mesmo foi avaliado a partir de julgamento humano.

Vale destacar que, dependendo do formato, assim como do conteúdo do artigo, é possível que nem todas as categorias procuradas sejam encontradas. Além disso, conforme discutido anteriormente, as categorias apresentadas são o resultado de um processo incremental e cíclico, o que obrigou que vários artigos fossem analisados várias vezes.

Como resultado da análise de conteúdo realizada, optou-se por apresentar, num primeiro momento (seção 4.2.1), as publicações analisadas organizadas em ordem cronológica dentro de cada categoria utilizada, e num segundo momento (seção 4.2.2), os resultados observados sistematizados no formato de mapas conceituais. Todos os mapas conceituais foram construídos utilizando-se a ferramenta CMap[19]. Os critérios usados na elaboração destes mapas, incluindo *layout*, cores e distribuição das publicações, são discutidos no capítulo 4.

[19] *CMapTools Knowledge kit* - versão 5.04, disponível em http://cmp.ihmc.us

3. PLN sob a ótica do ARIST: uma seleção de enunciados

Embora os onze capítulos de revisão analisados tenham apresentados títulos diferenciados: “processamento automático de linguagem” (nos oito primeiros) e “processamento de linguagem natural” (nos demais), todos tratam a área de PLN como sendo aquela responsável por manipular automaticamente a linguagem não controlada contida normalmente nos documentos textuais. Neste capítulo é apresentada a seleção de enunciados elaborada durante a análise de assunto dos capítulos de revisão do ARIST, tendo como finalidade construir o critério de seleção automática das publicações da área de PLN. As citações são apresentadas em ordem cronológica: do capítulo mais antigo (de 1966) para o mais recente (2003). A estrutura de tópicos adotada neste capítulo (e apresentada na FIG. 4) fomentou as quatro categorias de conceitos usadas no instrumento de seleção (conceitos computacionais, linguísticos, aplicações e, técnicas ou métodos) apresentado na seção 2.1.

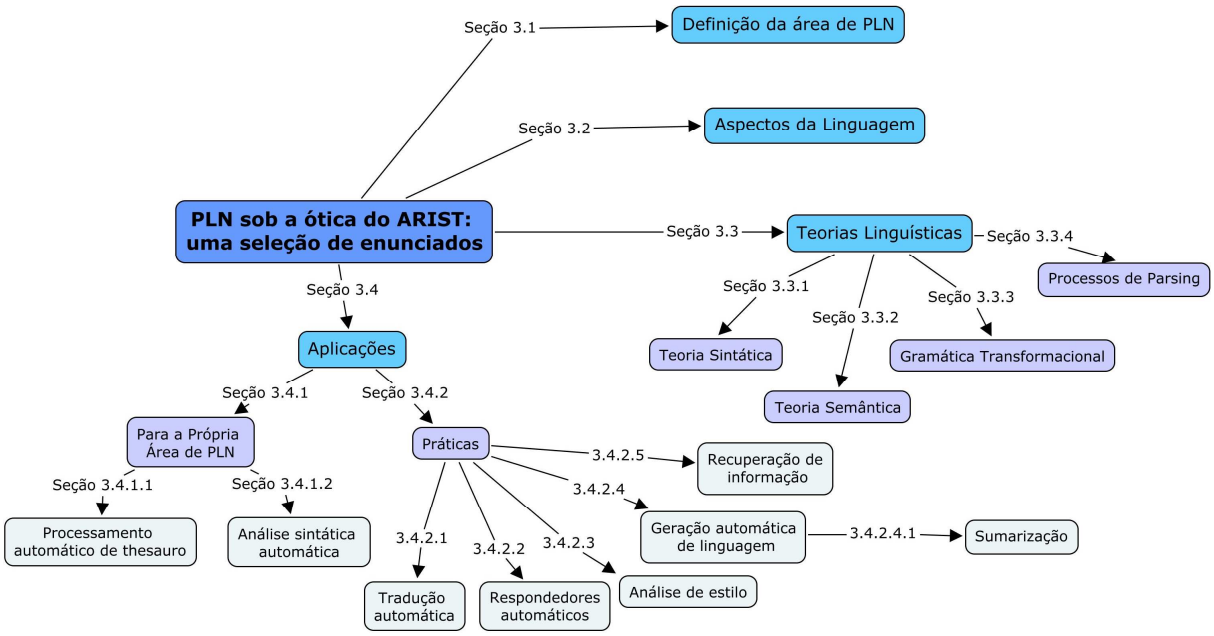

FIGURA 4 - Estrutura de tópicos adotada no Capítulo 3

A partir da seção 3.1 (Definição da área de PLN), foram identificadas as operações lógicas (booleanas) usadas no instrumento: conceitos computacionais E (conceitos linguísticos OU aplicações OU técnicas/métodos). Na seção 3.2 (Aspectos da Linguagem) procurou-se extrair dos capítulos de revisão do ARIST, as

questões relacionadas aos diferentes níveis de análise, assim como as dificuldades e limitações no tratamento automatizado da linguagem, apresentados pelos autores. Os termos identificados nesta seção foram incluídos na categoria de conceitos linguísticos. Na seção 3.3, intitulada Teorias Linguísticas, procurou-se identificar as técnicas e métodos relatados pelos autores dos capítulos de revisão do ARIST para compor a terceira categoria do instrumento de seleção. Finalmente, a partir da seção 3.4 foram identificadas as aplicações da área de PLN. Inúmeras aplicações foram apresentadas e discutidas pelos autores dos capítulos de revisão, desde ferramentas voltadas para a própria área, tais como processadores automáticos de tesauros e de gramática, até ferramentas de aplicabilidade mais abrangente, tais como, sistemas de tradução automática e respondedores automáticos. As aplicações consideradas como sendo voltadas para a própria área foram incluídas na categoria técnicas e métodos, por serem usadas dentro do desenvolvimento de outras aplicações. As aplicações intituladas práticas foram incluídas na categoria de mesmo nome: Aplicações. Os termos incluídos na categoria de conceitos computacionais foram extraídos ao longo de todo o fichamento realizado.

3.1. Definição da área de PLN

Nesta seção é apresentado como os autores dos capítulos de revisão do ARIST definiram a área de processamento de linguagem natural, o que permitiu definir as operações lógicas usadas para combinar as quatro categorias de conceitos definidas no critério de seleção automática.

Segundo Bobrow *et al.* (1967), a área de processamento automático de linguagem cobre, numa visão ampla, qualquer uso do computador para processar qualquer tipo de linguagem. No entanto, os autores destacam que, durante a elaboração do seu capítulo de revisão, o escopo da área foi delimitado: somente foi considerado o processamento de linguagem natural, desconsiderando assim linguagens artificiais como linguagens de programação. Além disso, segundo Bobrow *et al.* (1967) preocupou-se somente com processamento analítico e não estatístico das linguagens naturais, excluindo assim a maioria dos trabalhos em indexação automática, sumarização, análise de conteúdo e de estilo (p. 161). Bobrow *et al.* (1967) afirmam que, de uma maneira resumida, deu-se ênfase a

trabalhos que geram e analisam sentenças em linguagem natural baseando-se em alguma gramática ou base de dados.

Walker (1973) afirma que, a expressão “processamento automático de linguagem” foi usada de maneira mais genérica englobando todos os estudos teóricos e práticos do uso do computador ou de técnicas computacionais no processamento de linguagem, especialmente a linguagem natural. Esta discussão acerca das terminologias e das áreas é apresentada por outros autores envolvendo outras expressões como, por exemplo, linguística computacional e processamento automático de linguagem. Walker (1973) afirma que a linguística computacional é um subconjunto da área de processamento automático de linguagem, apesar da primeira ser o foco principal de toda a área. Ainda segundo Walker (1973), a linguística computacional é um campo interdisciplinar; e a linguística é o principal ponto de partida para as pesquisas tanto em processamento automático de linguagem como em ciência da informação. No entanto, Walker (1973) afirma que, a linguística não aceita o computador ou qualquer computação como um componente metodológico essencial para seu campo, e que, inúmeros linguistas têm questionado não somente a relevância dos resultados das pesquisas em linguística computacional, mas a existência de tais resultados (p. 73). E complementa que técnicas computacionais (computadores) podem ser aplicadas em campos onde alguma formalização e/ou sistematização das descrições têm sido alcançados.

Walker (1973) foi o primeiro a destacar as pesquisas oriundas da ciência da computação, especialmente da Inteligência Artificial (IA), e a apresentar uma nítida preocupação com os recursos computacionais usados nos experimentos realizados na área.

> Procedimentos para dedução e inferência têm se tornado mais sofisticado, e desenvolvimentos em ciência da computação, especialmente em IA, têm resultado em novas técnicas de programação e novas heurísticas (WALKER, 1973, p. 76).

Esta abordagem foi ratificada no capítulo de revisão seguinte, quando Damerau (1976) afirmou que a influência da linguística no processamento de linguagem tem tido uma pequena evidência, e que as principais inspirações advêm de trabalhos em IA e psicologia cognitiva (DAMERAU, 1976, p. 108). Além disso, ele cita Moyne (1975) que compartilha desta opinião e justifica usando o número de simulações linguísticas realizadas pelos cientistas da computação, com sistemas

que requerem um nível específico de detalhes, os quais os linguistas não estão acostumados. Damerau (1976) ressalta que a maioria dos projetos conhecidos tem sofrido mais influência da IA e da psicologia cognitiva do que da linguística, e como consequência, vários, se não a maioria dos trabalhos nesta área, são produtos de departamentos de ciência da computação ou psicologia e não de linguística. Segundo ele, a linguística estuda a linguagem em todas as suas manifestações; a linguística Teórica (*Theoretical Linguistics*) é a parte da linguística voltada a explicar os princípios gerais da organização da linguagem; a linguística Computacional é a parte da linguística onde algoritmos são aplicados a coleções de material de linguagem; enquanto que o processamento de linguagem natural estuda como o computador pode ser usado para processar coleções de dados em linguagem (*language data*) para o propósito de reorganização, extração, etc...

Becker (1981), autor do primeiro capítulo de revisão da década de 80, define processamento automático de linguagem como sendo a manipulação, por computador, de dados não-numéricos (normalmente palavras em Inglês). Segundo Becker (1981), desde o capítulo do ARIST de Damerau (1976), o processamento automático de linguagem (ou ALP do inglês *automated language processing*) tem sofrido mudanças pelos avanços de hardware e software: a popularização dos sistemas para computadores domésticos com capacidade para processar atividades de ALP, tais como manter arquivos bibliográficos e arquivos de dados; o desenvolvimento de dispositivos de *talking* e *listening*; desenvolvimento dos sistemas de OCR; além dos avanços em software, tais como os analisadores (*parsers)* de linguagem, representação de significado por computador, inferência lógica e organização de base de dados.

Warner (1987), autor do primeiro capítulo com título “processamento de linguagem natural”, definiu a área como sendo:

> uma área de pesquisa e aplicações que exploram como a linguagem natural, usada como entrada em sistemas de computadores, pode ser manipulada e armazenada de forma que preserve certos aspectos do original (WARNER, 1987, p. 79)

Segundo Chowdhury (2003), autor do capítulo do ARIST mais recente, processamento de linguagem natural (PLN) é uma área de pesquisa e de aplicação que explora como os computadores podem ser usados para processar e manipular texto ou discurso em linguagem natural para fazer coisas úteis. Segundo ele, as

bases da área de PLN encontram-se em diversas disciplinas, tais como ciência da computação e da informação, linguística, matemática, engenharia elétrica e eletrônica, inteligência artificial e robótica, psicologia, etc.

3.2. Aspectos da Linguagem

Nesta seção procurou-se extrair dos capítulos de revisão do ARIST os termos que irão compor a categoria de conceitos linguísticos relacionados aos diferentes níveis de análise observados no tratamento automatizado da linguagem.

Simmons (1966) afirma que nos 15 anos que antecederam à publicação do seu artigo, o objetivo das pesquisas na área de processamento de linguagem natural era construir sistemas de tradução automática e de recuperação de informação. No entanto, críticas feitas por Bar-Hillel (1963 *apud* Simmons, 1966) e Oettinger (1963 *apud* Simmons, 1966) obtiveram grande influência, ao enfatizarem que “a linguagem é um dos artefatos mais complexos da mente humana, (...) e que nenhum sistema com qualidade poderá ser construído antes de se entender o sistema da linguagem natural” (SIMMONS, 1966, p. 137).

A complexidade da linguagem é ressaltada também por outros autores. Salton (1968), por exemplo, afirma que, enquanto é relativamente fácil isolar palavras individuais em um texto, a interpretação do significado das palavras é bem mais difícil. Além disso, segundo ele, não há um conjunto de regras bem definidas a partir das quais as palavras de uma linguagem podem ser combinadas em grupos ou sentenças com significado (SALTON, 1968, p. 171). E complementa que a correta identificação do significado de um grupo de palavras depende pelo menos em parte do reconhecimento das ambiguidades sintáticas e semânticas, da correta interpretação dos homógrafos, do reconhecimento das equivalências semânticas, da detecção das relações entre palavras, dente outros (SALTON, 1968, p. 172).

Kay e Sparck Jones (1971) afirmaram que, inicialmente, pensava-se que uma linguagem poderia ser estudada independente de outra. No entanto, segundo eles, esta visão mudou radicalmente com o advento da gramática gerativa nos anos 60, tornando-se claro que a doutrina que insistia que as linguagens não são relacionadas com outras, não poderia nunca ser a base de uma verdade científica (p. 141).

Já no final da década de 80, Warner (1987) ressaltou que “a organização dos dados devem incorporar informação em vários níveis: morfológico, gramatical, contextual, assim por diante” (p. 79). Além disso, o autor complementa que, “tanto a entrada como a saída deve ser na forma de sentenças simples, fragmentos de sentenças ou texto conectado”, e que “a linguagem pode ser fornecida e retornada na forma falada ou escrita”. Vale ressaltar que, assim como Warner (1987), a maioria dos capítulos de revisão analisados discute a linguagem na sua forma escrita.

Outro tema comum na literatura é sobre como gerenciar a ambiguidade de maneira computacionalmente eficiente e psicologicamente plausível. Segundo Warner (1987), de uma maneira geral, todo documento pertence a algum assunto, e cada assunto tem sua própria linguagem especializada. Segundo Warner (1987), as principais preocupações na representação de conhecimento são em como identificar o que deve ser representado. Warner (1987) complementa apresentando o ponto de vista de diversos autores na área, começando por Winograd (1981 *apud* Warner, 1987) que considera que:

> (...) sistemas em linguagem natural devem representar fatos sobre estruturas linguísticas, sobre a relação dessas estruturas com o mundo e sobre estruturas cognitivas de pessoas, assim como a necessidade de tratar as ações seguindo a noção de que as pessoas estão inseridas num padrão de interação (...) (WINOGRAD, 1981 *apud* WARNER, 1987, p. 81)

Warner (1987) cita Harris (1984 *apud* Warner, 1987) que ressalta não ser suficiente representar o conhecimento somente com estruturas estáticas, uma vez que a representação de conhecimento requer que uma estrutura interna de significados seja construída. Já Weischedel (1986 *apud* Warner, 1987) apresenta, segundo Warner (1987), uma visão mais ampla, classificando o conhecimento necessário por um sistema em alguns componentes: morfológico e fonético, sintático, semântico e pragmático. Além disso, Weischedel destacou algumas preocupações, tais como “(...) quanto e quais tipos de conhecimento precisam ser explorados para que certas construções sejam entendidas, e em que momento no processo o conhecimento deve ser explorado” (WEISCHEDEL, 1986 *apud* WARNER, 1987, p. 81).

Segundo Warner (1987), os mecanismos formais e as estruturas para representação de conhecimento são cobertos em inúmeros trabalhos. Dentre eles, Harris (1984 *apud* Warner, 1987) com uma visão geral do assunto, Cullingford (1986 *apud* Warner, 1987) que considera a estrutura de conhecimento como sendo uma

amálgama de dependências conceituais, algoritmos de senso-comum e preferências semânticas, Barr e Feigenbaum (1981 *apud* Warner, 1987) que dividem a representação de conhecimento em representação procedural e declarativa, e finalmente Grishman (1986 *apud* Warner, 1987) que apresenta uma visão das principais estruturas e formalismos para representação sintática e semântica de conhecimento (WARNER, 1987, p. 82).

Warner (1987) finaliza suas considerações sobre representação de conhecimento destacando que sistemas sofisticados de linguagem natural necessitam de grandes bases de conhecimento. Além disso, ele cita o trabalho de Hobbs (1984 *apud* Warner, 1987), que apresenta princípios nos quais uma base de conhecimento deve ser construída, e o de Frey *et al.* (1983 *apud* Warner, 1987) que relata técnicas de automação do processo de construção dessas bases (WARNER, 1987, p. 82).

3.3. Teorias linguísticas

Nesta seção procurou-se identificar as técnicas e os métodos relatados pelos autores dos capítulos de revisão do ARIST para compor a terceira categoria do instrumento de seleção: técnicas e métodos. Os autores dos primeiros capítulos de revisão do ARIST apresentaram as discussões acerca da teoria linguística em dois momentos distintos: desenvolvimentos envolvendo teoria sintática e os envolvendo semântica. No entanto, o limiar que determina a fronteira entre as duas teorias se tornou tênue a medida que a área foi se desenvolvendo, como será observado nos enunciados apresentados a seguir. Diante disso, neste capítulo procurou-se manter a mesma estrutura adotada inicialmente (seções distintas para as teorias sintática e semântica), mas no instrumento de seleção criado, optou-se por criar uma categoria genérica com Técnicas e Métodos.

Os primeiros capítulos de revisão analisados (SIMMONS, 1966; BOBROW *et al.*, 1967) deram um enfoque mais teóricos aos seus capítulos e portanto discutiram explicitamente as teorias linguísticas da época. Simmons (1966) apresentou a linguística computacional como sendo subdividida em teoria linguística (onde aspectos de sintaxe foram discutidos), teoria semântica e psicolinguística. Segundo o autor, o progresso nas pesquisas em processamento da linguagem

natural depende dos desenvolvimentos tanto na teoria linguística como na semântica (SIMMONS,1966). Bobrow *et al.* (1967) dividiram o artigo em teoria sintática (com os mais significativos desenvolvimentos teóricos e descritivos em sintaxe), e em teoria semântica (com as principais teorias propostas na literatura em 1966).

Já no artigo de revisão de 1968, Salton, por dar um enfoque mais prático, juntou as duas teorias em uma única seção “Teoria Sintática e Semântica da Linguagem”, onde tentou mostrar meios automatizados para realizar as duas análises. Em Montgomery (1969), a autora iniciou afirmando que as duas principais publicações do ano de 1968 foram em fonologia, ao invés de sintaxe e semântica. Mas mesmo assim, a autora citou dois desenvolvimentos em teoria linguística significantes para processamento de linguagem natural na época. O primeiro foi o reconhecimento da importância da semântica na teoria da gramática, o que, segundo a autora, era o caminho para a automação, com a especificação de uma lógica formal para representação semântica (MONTGOMERY, 1969, p. 152). O outro desenvolvimento linguístico citado foi a elaboração da noção de modelo de desempenho (*performance*). No entanto, a autor destaca que, embora uma explicação completa fosse impossível, era evidente a importância de heurísticas e estratégias de considerável valor em projetos automatizados de compreensão análoga ao humano (MONTGOMERY, 1969, p. 153).

Kay e Sparck Jones (1971) destacaram que a maioria do esforço da linguística computacional estava voltada para sintaxe. Em outras palavras, a análise sintática automática era o tópico mais bem compreendido dentro da linguística. Interesses em problemas semânticos, segundo os autores, estavam crescendo e deveriam se tornar o principal tema de pesquisa em linguística computacional nos anos seguintes (p. 149).

Como principais influências na época, Walker (1973) destaca a gramática gerativa transformacional de Chomsky, a gramática de *string* de Harris, a gramática sistêmica (*systemic*) de Halliday e a gramática estratificacional (*stratificational)* de Lamb, sendo que estas duas últimas não haviam sido citadas nos capítulos de revisão anteriores. Walker (1973) conclui que a maior questão é como a sintaxe e a semântica podem ser combinadas. Segundo o autor, três analisadores têm apresentado forte influência (durante os dois anos revisados por ele): analisador de redes de transição aumentada de Woods (*augmented transition network parser)* (WOODS, 1973), o analisador de Kay *(chart parser)* (KAY, 1967;1973) e o de

Winograd *(program for natural language understanding)* (WINOGRAD, 1971;1972). Já Damerau (1976) iniciou o seu capítulo de revisão afirmando que os autores anteriores enfatizaram os trabalhos em linguística teórica, mas a influência da linguística no processamento de linguagem estava tendo pequena evidência, e que as principais inspirações advinham de trabalhos em IA e a psicologia cognitiva (p. 108).

Becker (1981) não discutiu explicitamente as teorias linguísticas cobertas pelos outros autores, e justifica afirmando que desde o último capítulo do ARIST de Damerau (1976), a área de processamento automático da linguagem ter sofrido mudanças influenciadas pelos avanços de hardware e software. Vale destacar que a teoria transformacional de Chomsky foi intensamente citada e discutida nos primeiros capítulos de revisão, o que não foi observado a partir da década de 80.

Segundo Warner (1987), uma questão muito debatida e ainda não resolvida em processamento de linguagem natural envolve as regras usadas pela sintaxe e pela semântica no processo de análise (*parsing)*. O autor complementa dizendo que alguns pesquisadores têm começado a explorar outras informações como a pragmática (p. 85). Ainda segundo o autor, a exploração de regras de sintaxe e semântica é tema comum na literatura da época que discutem a relação entre memória, significado e sintaxe, e apontam que significado e conhecimento de mundo são cruciais no processo de compreensão da linguagem, e que a sintaxe deve ser utilizada no processo de análise, mas não exclusivamente. Por outro lado, o ator alega que analisadores de linguagem natural principalmente semânticos (“*semantic mainly*”) ou somente semânticos (“*semantic only*”) não são adequados para cobrir uma grande gama de línguas usadas pelas pessoas. Warner (1987) complementa que “o processo de compreender linguagem humana consiste em determinar os significados das especificidades localizadas (*‘utterances’*), mas que as estruturas sintáticas parecem ser uma parada necessária nesse caminho” (p. 85). O próprio autor destacou que, existe uma tendência em a análise baseada em sintaxe incorporar mais a semântica, apesar da sintaxe ainda manter a sua primazia (e cita CHARNIAK, 1983 e MELLISH, 1983).

3.3.1. Teoria Sintática

Apesar de ter afirmado que o progresso nas pesquisas em processamento da linguagem natural dependia dos desenvolvimentos tanto na teoria sintática como na semântica, Simmons (1966), autor do primeiro capítulo de revisão, destacou que grande parte dos esforços em processamento de linguagem estava fortemente embasada em teorias formais da estrutura sintática (p. 139). Simmons (1966) citou Garvin (1965) que listou 12 abordagens para gramática: estado finito (*finite state*), estrutura de frase (*phrase structure*), análise de dependência (*dependency analysis*), formacional (*formational*), transformacional (*transformational*), estratificacional *(stratificational),* dentre outras[1], mas afirmou que certamente a área da linguística mais fértil é a teoria transformacional, escola amplamente desenvolvida por Chomsky e Katz e Postal (na década de 60).

Bobrow *et al.* (1967) consideraram como objetivo dos linguistas ao escrever uma gramática, representar os fatos que os falantes nativos da linguagem conhecem. Este conhecimento tem sido chamado de competência do falante. Assim, uma gramática deve ser pensada como um modelo para a competência ideal. Segundo Bobrow *et al.* (1967), este conhecimento nem sempre é óbvio, depende do contexto e da pronúncia (ou seja, das regras fonológicas). Com isto, várias regularidades da linguagem podem ser capturadas somente através de representações abstratas dos fatos superficiais (p. 162). Ainda segundo Bobrow *et al.* (1967), um objetivo mais distante dos linguistas é encontrar características de todas as linguagens, e então determinar uma especificação mais simples possível destes fatos universais, a parte das características específicas de uma linguagem em particular. Esta abordagem assume que todas as linguagens apresentam similaridade. No entanto, geralmente assume-se que não existe um procedimento para determinar a análise linguística mais criteriosa para uma parte de uma dada linguagem em termos de uma determinada teoria linguística (p. 162).

Bobrow *et al.* (1967, p. 163 a 165) citam como contribuições na área de sintaxe os trabalhos de Lakoff (1965), Fillmore (1966), Chapin *et al.* (1965) que apresentam as regras gramaticais desenvolvidas para o procedimento de análise para gramática transformacional MITRE.

---

[1] Além de outras gramáticas, tais como *word-paradigm, item-and-process, item-and-arrangement, immediate constituent, tagmemic and glossematic.*

Salton (1968) afirmou que, apesar de procedimentos de análise sintática não poderem ser usados para resolver por completo o problema de identificação de conteúdo, o conhecimento de propriedades sintáticas das palavras é importante para reconhecer certas relações que existem entre palavras dentro das sentenças, por exemplo, combinações de sintagmas nominais, preposicionais, adverbiais, e agrupamentos simples de sujeito-verbo-objeto (p. 172). Ainda segundo Salton (1968), a maioria dos sistemas de análise sintática automática é baseada em regras de construção, ou em gramáticas, conhecidas como gramáticas de estrutura de frase, na qual uma sentença em linguagem natural é considerada como sendo constituída de um conjunto de frases justapostas e aninhadas. Segundo o autor, uma gramática de estrutura de frases é normalmente definida por um conjunto de regras de reescrita. A derivação de uma dada sentença produzida por uma gramática de estrutura de frase é especificada pela citação das regras de reescrita, usadas na sua geração, assim como a ordem na qual as regras foram aplicadas. A derivação de uma sentença numa dada gramática pode se representada por uma árvore chamada de marcador de frase ou descrição estrutural (p. 172).

No entanto, Salton (1968) alertou que as gramáticas de estrutura de frases sofrem de várias desvantagens já conhecidas, que diminuem o potencial do seu uso em sistemas de análise automática de conteúdo: não existem métodos que permitem escolher a regra correta quando existem várias derivações possíveis a serem aplicadas a uma dada sentença; e alguns resultados da análise, embora gramaticalmente corretos, podem ser semanticamente inaceitáveis (p. 173). Outro inconveniente da gramática de estrutura de frases é o fato de refletir somente a estrutura de superfície de cada sentença, usada na representação fonética, mas não necessariamente para a interpretação semântica (p. 173.).

Segundo Kay e Sparck Jones (1971), a maioria dos linguistas acredita que uma gramática deve não somente prover um significado para distinguir sentenças de não-sentenças (p. 143). Uma gramática adequada para uma linguagem deve mostrar as partes de uma sentença e deve classificar os tipos de relações que podem ser realizadas entre estas partes, assim como a influência do significado como um todo (propósito principal da gramática) (p. 144). Os autores destacaram que Chomsky considerava que as regras transformacionais, que fazem a mediação entre estruturas profundas e de superfície, não deveriam ter efeito de significado. Assim, os componentes semânticos da gramática precisam ser aplicados somente à

estrutura profunda (p. 144). Ainda segundo os autores, apesar de vários argumentos favoráveis e contrários à natureza da preservação de significado das regras transformacionais, evidências claras para um julgamento final ainda não estão disponíveis (p. 145). Outro problema lógico que tem sido tema de discussão são os chamados quantificadores sintáticos: negação, conjunções, etc... (p. 145). Os autores afirmam que a relação entre linguagem natural e lógica é de fundamental importância sempre que qualquer material textual precisar ser tratado mecanicamente (p. 145).

Kay e Sparck Jones (1971) destacam que alguns semanticistas generatistas advertem que uma gramática transformacional não deve conter um componente semântico separado do componente sintático, mas as regras transformacionais devem mediar diretamente a relação entre as representações semânticas e as estruturas sintáticas de superfície. E complementam que a noção de estrutura profunda, distinta da estrutura semântica, não tem justificativa satisfatória. Um segundo grupo de gramáticos transformacionais, liderados por Fillmore (1968), acredita que grande parte da nossa fala, principalmente os verbos, comporta-se como funções, que recebem um conjunto de argumentos de tipos específicos e podem ser nulos (p. 146). E nesta direção, segundo os autores, a gramática de casos de Fillmore tenta elucidar que a correlação entre marcadores gramaticais em estruturas de superfície e os casos ou tipos de argumentos da estrutura profunda são algumas vezes complexas (p. 146). O apelo da gramática de casos para linguistas computacionais não é difícil de entender, porque funções com argumentos podem ser facilmente modeladas com o cálculo de predicados. Em outras palavras, a gramática de casos provê um conjunto de formas canônicas para sentenças que são facilmente acomodadas num formalismo bem conhecido.

3.3.2. Teoria Semântica

Ao apresentar a teoria semântica, Simmons (1966) cita: a teoria de Katz-Fodor (1963), complementada por Katz e Postal (1964); a teoria de classificação e análise semântica desenvolvida pelo *Cambridge Language Research Unit (CLRU)* (SPARCK JONES, 1964) e a teoria de memória semântica de Quillian (1966) (p. 141). Segundo Simmons (1966), a teoria KF de Katz-Fodor assume que um componente semântico é parte integral da descrição linguística; e que compreende

várias partes: um dicionário, o qual provê o significado de cada palavra ou entrada léxica da linguagem; um conjunto de regras de projeção, que provê meios de interpretar os significados de cada ocorrência de entrada lexical da linguagem; e outro conjunto de regras de projeção para prover meios de interpretar cada *string* produzida pela gramática (p. 141). Segundo Simmons (1966), grande parte da teoria concentra-se no desenvolvimento de uma forma padronizada para o conteúdo das entradas lexicais. Esta forma inclui a palavra de entrada e um marcador sintático, tais como nome, verbo, etc., seguido de um marcador semântico, como animal, humano, macho, etc., seguido ainda por um diferenciador opcional para definição do sentido, e finalmente uma restrição de seleção (p. 142). Apesar da teoria semântica KF ter clareado alguns aspectos da estrutura transformacional, alguns autores acham prematura esta tentativa de formalização diante de uma super simplificação (p. 142).

Simmons (1966) destaca ainda que, Sparck Jones (1965) em sua monografia "*Synonomy and Semantic Classification*" desenvolveu um método para definir o uso sinonímico (sinônimos) das palavras e classificá-las em grupos no thesaurus. Este procedimento é essencial para selecionar uma sentença (normalmente um exemplo de uso contido no dicionário de definições) e para substituir palavras sem alteração do significado (p. 143). Segundo Simmons (1966), a partir do momento em que as palavras de uma lista puderem ser trocadas entre si, significa que elas devem conter um elemento comum de significado. Assim, uma linha é uma lista de palavras que compartilham um uso comum, e são obtidas por um método de julgamento humano (p. 143). Estas linhas resultantes podem ser agrupadas para formar um thesaurus, baseando-se nas palavras que as linhas possuem em comum. Uma abordagem possível é a estatística – usada na teoria CLRU de agrupamento (NEEDHAM, 1965). Segundo Simmons (1966), o principal problema ainda não resolvido das pesquisas de classificação é utilizar a abordagem estatística pra classificar não apenas 500 e sim 50 mil ou 150 mil linhas (p. 143). Após obter a classificação, a teoria CLRU propõe separar as palavras em contextos apropriados usando, ou a medida de distância semântica entre elas na sentença, ou usando padrões de mensagem (*message forms)*, que mostram as combinações permitidas de classes semânticas (thesaurus). A distância semântica é obtida encontrando-se o caminho de uma palavra na sentença até outra palavra, examinando linhas associadas com cada palavra para encontrar palavras em

comum. Duas palavras de uma sentença que são encontradas em uma mesma linha exemplificam a menor distância em comum (por exemplo, 0). Se não for o caso, mas existir uma terceira linha que contenha uma palavra em comum com cada uma das linhas anteriores, a distância é a mais longa possível (por exemplo, 1). Desta maneira, um caminho de distância geralmente pode ser encontrado em qualquer par de palavras (p. 143). Segundo Simmons (1966), os padrões de mensagem são estruturas de significado essenciais que são construídas a partir de classes semânticas. Por exemplo, a sentença “O livro é vermelho” daria um padrão de mensagem “objeto é cor”. Este padrão seria obtido para selecionar o senso de objeto e cor das palavras livro e vermelho. (p. 144).

Segundo Simmons (1966), uma abordagem empírica semelhante para análise semântica é a memória semântica de Quillian (1966). Nesta abordagem, definições do dicionário são codificadas manualmente (*hand-coded*) em um formato adequado para computador, mantendo informação de classe sintática e as relações de dependência de palavras em cada definição. Cada uso de uma palavra é associada a uma definição particular para manter sua identidade (p. 144). Segundo o autor, este modelo de memória semântica tem um componente sintático e procedimentos bem definidos para seus conteúdos definicionais (p. 145).

No capítulo de revisão seguinte, Bobrow *et al.* (1967) destacam que Simmons sugeriu que existem vários fragmentos principais de uma teoria semântica compreensiva, e que estes fragmentos (Katz, Fodor e Postal do CLRU – *Cambridge Language Research Unit*) devem se unir para formar uma teoria semântica harmoniosa. Segundo os autores, o fato é que no ano de 66, inúmeras críticas foram feitas especialmente ao grupo de Katz. Bobrow *et al.* (1967) destacam que outra abordagem que apareceu nesta mesma época foi a de Weinreich e Quillian, ambos mostrando o por quê deles acharem que Katz estava errado. Katz em (1966*)* replicou as críticas, e a discussão entre eles foi apresentada por Bobrow *et al.* (1967) nas páginas 166 e 167, ao relatar que eles acusavam Katz e seus co-autores de simplificarem muito a sua teoria semântica. O grupo de Katz (KFP) assume que a informação semântica armazenada no léxico é mantida na forma de árvores simples com arestas *(links)* sem rótulos e os nodos com dois tipos de rótulos (marcador ou diferenciador). Além disso, o grupo KFP afirmou que o resultado final do componente semântico de uma linguagem é uma informação curta e simples, como por exemplo, um único grupo de propriedades desordenadas para cada sentença (BOBROW *et*

al., 1967, p.166). Segundo Bobrow *et al.* (1967), Quillian propõe que a informação semântica seja representada por estruturas recursivas (grafos) e por configurações complexas, construídas a partir de diferentes tipos de *links*, cuja vantagem seria a possibilidade de se usar o computador (p. 168).

Bobrow *et al.* (1967) afirmam que existem duas técnicas principais para a representação das estruturas de informação semântica associada a linguagem: a primeira é ver esta informação como sendo a gerada por um conjunto de regras recursivas, tais como uma gramática ou um programa de computador; e a segunda é ver esta informação como um tipo de rede com *links* associados (p. 169).

Em seu capítulo de revisão, Salton (1968) alerta para o fato de não existirem métodos de análise sintática disponíveis na época para gerar uma única interpretação semântica para cada sentença bem-formada em inglês. Segundo ele, certos agrupamentos de palavras (frases) podem ser isolados com as fronteiras de sentenças individuais, mas as relações entre componentes de frase e entre frases individuais permanecem amplamente desconhecidas. Salton (1968) considera que, para propósitos práticos, é necessário contar com a gramática de estrutura de frase para análise de entradas em linguagem natural, preferivelmente com a mais refinada gramática transformacional, apesar da primeira (gramática de estrutura de frase) ainda ser indefinida em partes, e ser custosa para se implementar (p. 176). E finaliza:

> (...) mesmo se for assumido que análises absurdas podem, de alguma maneira, ser descartadas, e que gramáticas (transformacionais) implementadas atribuiriam as mesmas análises a várias estruturas derivadas, ainda não existiria solução para o problema de escolher quais os significados apropriados para as palavras que normalmente podem ter diferentes (até nos mesmos contextos) (SALTON, 1968, p. 176).

Segundo Salton (1968), vários modelos de análise semântica têm sido propostos, baseados em dicionários de entradas com marcadores semânticos apropriados, e em uma gramática transformacional para o reconhecimento da estrutura sintática (p. 176).

3.3.3. Gramática Transformacional

Dentre todos os tipos de gramática, a que sem dúvida influenciou os trabalhos da área de processamento de linguagem natural foi a gramática gerativa de Chomsky. Conforme mencionado anteriormente, esta teoria foi enfaticamente

discutida e apresentada nos dois primeiros capítulos de revisão, e justamente por esta razão, Simmons (1966) e Bobrow *et al.* (1967) serão usados como as principais referências nesta seção. Os demais artigos apresentaram uma discussão sobre a teoria transformacional sem defini-la com detalhes.

O primeiro capítulo de revisão a descrever detalhadamente a estrutura da gramática transformacional (ou gerativa) de Chomsky foi o de 1967, enquanto que Salton (1968) apresentou uma descrição simplificada, Montgomery (1969) voltou a descrevê-la detalhadamente, e os demais capítulos de revisão não privilegiaram tal teoria.

Segundo Simmons (1966), Chomsky, em seu livro *"Aspects of the Theory of Syntax*", revisou vários aspectos da teoria de formalização, onde afirmou que uma gramática estruturada em frases (*phrase structure grammar*) é suficiente somente para gerar as sequências (*strings)* essenciais considerando-se a complexidade das sentenças em inglês. Segundo o autor, usando operadores para transformar e combinar estes núcleos (*kernels)* seria possível produzir uma variedade de estruturas de sentenças complexas (p. 140). Segundo Simmons (1966), esta teoria é voltada principalmente para a capacidade gerativa da gramática. Ou seja, dado uma gramática e um ponto de partida, a teoria permite, num primeiro momento, a geração de estruturas profundas de sentenças, a partir de um sistema de geração de estruturas de frases. E num segundo momento, permite a combinação destas em estruturas mais complicadas (chamadas de estruturas de superfície) e em outras sentenças pelo uso de transformações opcionais, tais como aquelas que produzem forma ativa e passiva da base, uma negação, combinação, adição ou remoção de termos. Simmons (1966) destacou que estas transformações são sempre aplicadas a estruturas sintáticas completas (ou seja, a árvores) e não em nós terminais (p. 140). Simmons (1966) destacou que revisões da teoria transformacional resultaram em formalizações mais concisas e mais genéricas das estruturas sintáticas (p. 140).

Bobrow *et al.* (1967) consideraram como base *(frameworks)* linguística dos trabalhos em processamento de linguagem, a teoria de Chomsky (e outros) introduzida em *"Syntactic Structures"* e elaborada em *"Aspects of the Theory of Syntaxe"* sendo que este último foi amplamente citado nos capítulos de revisão do ARIST.

Segundo Salton (1968), a gramática transformacional leva em consideração não somente a estrutura superficial como também a estrutura profunda

de uma sentença, visto que as estruturas profundas tentam considerar aspectos semânticos da interpretação da sentença (p. 174).

Segundo os autores dos primeiros capítulos de revisão, a gramática transformacional consiste de três componentes principais: o sintático, o semântico e o fonológico, detalhados a seguir. O componente sintático é formado por um sub-componente base, um sub-componente transformacional e um léxico. Segundo Bobrow *et al.* (1967), o componente base, por meio de substituições léxicas, produz estruturas para representar todos os relacionamentos semânticos possíveis. Salton (1968) detalha afirmando que componente base gera para cada sentença um conjunto de marcadores de frases generalizados, que consiste de uma *string* terminal com uma descrição do tipo de estrutura. Em outras palavras, o componente base consiste de um componente categorial, ou seja, um sistema de regras de escrita que gera um conjunto de *strings* básicas com suas descrições estruturais associadas ou “marcadores de frase” (MONTGOMERY, 1969, p. 147).

Segundo Montgomery (1969), os componentes semânticos e fonológicos são puramente interpretativos. Seus objetivos são correlacionar representações semânticas e fonológicas com estruturas geradas pelo componente sintático (p. 147). Assim, dificuldades com inserções lexicais, assim como as representações de estruturas profundas têm levado alguns discípulos de Chomsky a mudar a noção da natureza gerativa do componente sintático, sustentando o componente semântico como elemento criativo e o sintático como interpretativo (p. 148). Ainda segundo Montgomery (1969), apesar da discussão acerca do caráter gerativo ou interpretativo dos componentes sintáticos e semânticos, a questão essencial é se postulados universais como os propostos por Chomsky (1966), Katz e Fodor (1969) e Katz e Postal (1969) são de fato suficientemente universais para lidar com as características em comum de todas as linguagens naturais. Montgomery (1969) conclui que um sentimento geral, especialmente considerando a noção de estrutura profunda, é que eles não são (p. 148).

3.3.4. Processo de análise *(parsing)*

Nos primeiros capítulos de revisão do ARIST observou-se uma predominância de analisadores *(parsers)* com componentes sintáticos. Salton (1968) afirmou que enquanto procedimentos de análise sintática não podiam ser usados

para resolver por completo o problema de identificação de conteúdo, o conhecimento de propriedades sintáticas das palavras era importante para reconhecer certas relações que existem entre palavras dentro das sentenças, por exemplo, combinações de sintagmas nominais, preposicionais, adverbiais e agrupamentos simples de sujeito-verbo-objeto (p. 172).

No entanto, no início da década de 70, Walker (1973) concluiu que a maior questão era como a sintaxe e a semântica poderiam ser combinadas no desenvolvimento de analisadores. O mesmo autor citou Sparck Jones e Wilks (1983) que ressaltava a tendência em processamento de linguagem natural na direção de gramáticas de estrutura de frase e analisadores determinísticos, além de uma maior integração de sintaxe e semântica.

Damerau (1976), em seu capítulo de revisão, considerou o analisador *(parser)* como sendo um dos principais componentes de um sistema de processamento automático de linguagem e acredita que este problema foi o primeiro a ser tratado pela linguística computacional, mas que continuará sendo um tema substancial de pesquisas.

Warner (1987) considerou o analisador *(parser)* como sendo o componente central do sistema para processamento de linguagem natural, e apresentou a seguinte definição de Sparck Jones “(...) processo computacional que obtém sentenças individuais ou textos conectados e converte-os para alguma estrutura de representação útil para posterior processamento” (p. 83).

O primeiro analisador citado foi o MITRE, apresentado em Simmons (1966) e em Bobrow *et al.* (1967), quando os autores afirmaram que para analisar uma entrada em língua inglesa por computador, eram necessários programas de análise sintática, assim como sistema MITRE.

Bobrow *et.al.* (1967) apresentaram o trabalho de Kuno (1966): um algoritmo que recebe uma gramática livre de contexto e converte-a para uma gramática em formato padrão de um analisador preditivo (*predictive*). Esta forma padrão é argumentada de tal maneira que as árvores produzidas pelo analisador provê informações sobre derivações que seriam encontradas usando a gramática livre de contexto original. Este analisador argumentado e preditivo (*argumented predictive*) foi comparado com dois outros algoritmos de análise *(parsing)*: o algoritmo seletivo *top-botton* semelhante ao algoritmo de correção de erros *(error-correcting parse algorithm)* de Irons (1963*)* e o algoritmo de Sakai-Cocke (1961).

Esta comparação foi feita baseando-se nos critérios de eficiência, complexidade do programa e tempo de processamento. A conclusão da comparação foi que o analisador argumentado e preditivo é comparável, se não superior, mas que a escolha por um ou por outro depende muito mais da aplicação em questão (p. 174). Outro analisador citado por Bobrow *et al.* (1967) foi o trabalho de Sager, derivado de Harris (1966), no qual a saída da análise é uma *string* que representa o esqueleto da sentença.

Walker (1973) e Damerau (1976) escreveram, sem dúvida, os dois capítulos que mais privilegiaram os analisadores *(parser)*. Walker inclusive afirmou que todas as citações apresentadas foram organizadas no formato de um catálogo, tanto para ilustrar a variedade, quanto pelo fato que a sua complexidade não permitiria que qualquer agrupamento fosse feito. Em seu capítulo de revisão, Walker cita vários analisadores, mas conclui que três deles tiveram grande influência durante os dois anos em que escreveu: o analisador de Woods, o analisador de Kay e de Winograd.

Walker (1973) citou também o sistema CUE de Loveman *et al.* (1973), que fazia uso de análise sintática baseada na teoria linguística de Harris para processar conteúdo de texto científico. Além disso, Walker (1973) citou Sager e Grishman por descreverem um analisador que não decompõe sentenças em *strings*. Sager mostra como gramáticas para linguagens técnicas podem ser desenvolvidas e aplicadas juntamente com analisadores para analisar textos científicos.

Dentre as referências feitas por Damerau (1976), está a ATN de Woods (1970), considerada por ele um exemplo de sucesso. Damerau (1976) apresentou alguns tipos de analisadores, destacando as dificuldades encontradas em cada abordagem. Segundo o autor, o tipo mais simples de análise deve meramente examinar quais palavras ocorrem na sentença. No entanto, cada palavra é associada a uma entrada lexical complexa que identifica outras palavras ou conceitos que podem ser esperadas, dado que uma palavra particular ocorreu (p. 119).

Outro tipo conceitualmente simples de análise é o casamento de padrões (*pattern matching)*, onde uma palavra é comparada com *strings* de entrada e o casamento acontece se tanto a palavra, como o padrão, foram encontrados na entrada (p. 119.). Outro sistema ainda considerado simples por Damerau (1976) é identificar quando certas palavras são distintas por terem significado estrutural, ou

seja, usados principalmente para relacionar uma palavra ou conceito a outro (p. 119). Damerau (1976) apresentou ainda alguns problemas que aumentam a complexidade destes casamentos, além de afirmar que sistemas destes tipos devem ser tolerantes a erros cometidos pelo usuário ao definir entradas não previstas pelo sistema (p. 120).

Damerau (1976) também citou Harris com sua proposta de análise de *string*. Este *parser* tem uma parte livre de contexto e outra parte com restrições, sendo esta última a que provê um tipo de testes de compatibilidade para remoção de ambiguidade, o que é difícil ou impossível em gramáticas livre de contexto (p. 120).

Damerau (1976) cita outro analisador (descrito por Medema) composto por uma sequência de conversores que transformam uma árvore em outra, chamada de árvore de decisão flexível *(flexible decision tree*) que aceita uma decisão até um determinado ponto, onde alguma outra mudança (transformação) remove a ambiguidade. A noção de *wait-and-see* descrita por Marcus é semelhante, onde as regras são invocadas pelo casamento de padrões, mas também por prioridade. Marcus assume que a estrutura da linguagem natural tem sempre informação suficiente para decidir qual o próximo passo a ser seguido pelo analisador (p. 121).

Outra abordagem citada por Damerau (1976) foi Heidorn (1975b) que discute o analisador para gramática de estrutura de frase argumentada (APSG), o qual as regras de estrutura de frase são argumentadas por condições arbitrárias e ações de construção de estrutura, algo como um compilador orientado à sintaxe (p. 121). Damerau (1976) concluiu que pesquisas em análise (*parsing*) mostram que o campo de processamento automático de linguagem natural está migrando da arte para a tecnologia: “as técnicas computacionais disponíveis na época nos permitem gastar mais tempo em problemas da linguagem e menos em problemas da ciência da computação de *parsing*” (p. 123).

Segundo Warner (1987), a questão do determinismo foi proposta por Marcus (1980) quem reivindica que o seu analisador PARSIFAL faz linguisticamente generalizações significantes e é psicologicamente exato. A característica chave do PARSIFAL é que não é necessário que nenhuma análise paralela ou *backtracking* seja realizada na sentença e que ela falha somente nos casos de complexidade psicológica óbvia. Warner (1987) ressalta que embora a realidade psicológica do *parser* determinístico de Marcus tenha sido questionada, determinismo é ainda uma importante questão dentro do processamento de linguagem natural (p. 86). Becker

(1981) complementa afirmando que o *parser* de Marcus nunca faz *backtracking* por nunca revisar sua estratégia. No entanto, Becker afirma que esta história não está no fim por que várias sentenças são semanticamente ambíguas.

Warner (1987), em seu capítulo de revisão, cita Winograd (1983) que descreve algoritmos de análise sintática e enfatiza suas propriedades incluindo pontos relacionados com: completude da análise: parcial ou completa; maneira de atribuir as estruturas profundas ou de superfície; manipulação de entradas ambíguas: *parser* paralelo ou *backtracking*, determinístico; cobertura sintática: incluindo fenômenos mais difíceis tais como os complexos sintagmas nominais e conjunções; domínio específico; dentre outros.

Segundo Warner (1987), o livro organizado por Dowty *et al.* (1985) aponta dois problemas teóricos dos *parsers*: a realidade psicológica de vários mecanismos de *parsing*, verificados por meio de cuidadosos experimentos controlados, e, as propriedades formais que os *parsers* devem ter, incluindo quanto de poder é necessário para descrever adequadamente linguagens humanas e a habilidade dos *parsers* de fazer generalizações linguísticas significantes. Warner (1987) continua afirmando que outros mecanismos de *parsing* com forte componente semântico são discutidos na literatura. Segundo ele, *Case frames* são utilizados para *parser* em Hayes *et al.* (1985), que discutem o poder e as fraquezas dessa abordagem no seu *parser* chamado de Plume. Segundo Warner (1987), *case frames* também são usados por Shimazu *et al.* (1983), que descrevem sua implementação em um analisador semântico da língua japonesa.

Warner (1987) citou outros pesquisadores que exploraram a utilidade do determinismo em várias construções: Milne (1986) mostrou que o *parser* determinístico pode facilmente resolver certas classes de ambiguidade lexical, principalmente no que se refere à categoria lexical (*part of speech*). Kosy (1986) descreveu como um *parser* determinístico pode processar conjunções eficientemente; e Berwick (1983) descreveu uma extensão que aumenta a cobertura sintática do formalismo de Marcus. Charniak (1983) introduziu outro parser baseado no formalismo de Marcus, o qual é semigramatical sob o ponto de vista que aceita sentenças que não seguem a gramática. Carter e Freiling (1984) descreve uma pequena implementação de *parser* chamado PARSER (*Deterministic PARSER*), que pretende reduzir a complexidade das gramáticas determinísticas.

Warner (1987) ainda dedica uma seção para discutir as dificuldades de se

construir *parsers* para manipular tipos específicos de construção, tidos como sendo altamente ambíguos, tais como sintagmas nominais, construções temporais e construções envolvendo conjunções, quantificação e anáfora. Dentre os autores que discutem os problemas de interpretação de sintagmas nominais, Warner (1987) cita Sparck Jones (1985), que argumenta que manipular sintagmas nominais complexos envolve ampla inferência de outros tipos de informação, incluindo informação semântica e conhecimento do mundo. Warner (1987) cita outros problemas para os *parsers*, tais como identificação de anáforas, quantificação e informações temporais (p. 88).

3.4. Aplicações

Analisando os capítulos de revisão focalizados, observou-se que os autores, além de apresentarem inúmeras aplicações, discutiam também algumas propriedades relativas ao desenvolvimento de sistemas que manipulavam linguagem natural, dentre elas a portabilidade, a aplicabilidade, a robustez, a importância da componentização, assim como da reutilização destes componentes.

Inúmeras aplicações foram apresentadas e discutidas pelos autores dos capítulos de revisão, desde ferramentas voltadas para a própria área, como processadores automáticos de tesauros e de gramática, até ferramentas de aplicabilidade mais ampla, como, por exemplo, sistemas de tradução automática e respondedores automáticos. Esta discussão foi apresentada também em Bobrow *et al.* (1967), onde os sistemas computacionais foram classificados em duas vertentes: sistemas que manipulam a linguagem provendo ferramentas linguísticas e sistemas que aceitam questões em linguagem natural e usam algum banco de dados para respondê-la.

Mais uma vez, não existe um limiar nítido entre estas duas supostas categorias. Acredita-se que ambos os tipos de aplicação são importantes para o próprio desenvolvimento da área. Bobrow *et al.* (1967) afirmaram que:

> (...) alguns destes sistemas de perguntas e respostas têm como objetivo desenvolver uma teoria da linguagem eficiente, ao mesmo tempo que as ferramentas teóricas objetivam permitir que sistemas se comuniquem em linguagem natural (p. 172).

Vale destacar que ao longo dos 40 anos pesquisados, é possível observar uma mudança no enfoque das aplicações: inicialmente, era dada maior ênfase às ferramentas linguísticas de processamento sintático e semântico, e nos últimos anos, uma nítida exploração das aplicações práticas. Possivelmente esta mudança tenha sido impulsionada pelos avanços de hardware e software assistido nas últimas décadas, além do interesse crescente de pesquisadores da ciência da computação pela área de processamento de linguagem natural.

Simmons (1966) em seu artigo de revisão discutiu que inúmeros esforços estavam sendo feitos não apenas para tradução automática, mas também outras aplicações úteis de processamento de linguagem natural como análise de conteúdo, indexação automática, classificação e sumarização, respondedor automático (*question answering*), análise de estilo (*stylistic analysis*), dentre outras. Além disso, o autor cita algumas iniciativas de sistemas de análise sintática automática, evidenciando a importância do desenvolvimento de ferramentas voltadas para a própria área.

Bobrow *et al.* (1967) enfatizam as aplicações de cunho mais teórico e apresentam inúmeros projetos que utilizam programas de computadores como suporte para processamento da linguagem. Dentre os projetos citados pelos autores encontram-se programas que auxiliam as tarefas gramatical e lexicográfica (projeto da IBM), que testam a capacidade gerativa da gramática transformacional (p. 172), que realizam análise sintática (sistema MITRE), assim como algoritmos que recebem uma gramática livre de contexto e converte-a para uma gramática em formato padrão para um analisador preditivo.

Salton (1968), apesar de ter apresentado pesquisas em processamento automático de texto, incluindo sintaxe, semântica e métodos de análise estatística de linguagem, enfatizou aspectos práticos das aplicações nas áreas de tradução automática, recuperação da informação e respondedor automático. Além disso, Salton (1968) discutiu sobre os componentes que compõem um processador de linguagem natural. Esta discussão também é apresentada por Damerau (1976), ao afirmar que sistemas de compreensão de linguagem tendem a ser construídos com componentes similares, e que em alguns casos, um componente pode substituir outro similar em um outro sistema, com algumas pequenas alterações (p. 117). Damerau concluiu que padronizações deste tipo reduzem naturalmente o esforço da construção de sistemas, a partir da integração destes sistemas.

Dentre os componentes apresentados em Salton (1968) estão: um analisador sintático para identificar as relações estruturais; um analisador semântico para transformar a saída sintática em entidades não ambíguas em alguma linguagem formal; uma estrutura lógica (cognitiva) dos objetos e relações que representam os significados das entidades da forma como são percebidas pelos humanos (normalmente especificadas por um dicionário semântico); um procedimento de inferência para o reconhecimento de estruturas sintáticas distintas com os significados equivalentes; e um sistema de geração sintático-semântico para produzir declarações em inglês a partir de dadas estruturas formais. Dentre os componentes citados por Damerau (1976) estão o dicionário, um componente de análise morfológica e um analisador *(parser)*.

A aplicabilidade dos sistemas PLN é discutida por inúmeros autores ao considerarem o processamento de linguagem natural como sendo uma etapa presente em qualquer sistema de informação. Montgomery (1969) justifica afirmando que “a linguagem, por ser o principal veículo para comunicar informação na sociedade humana (...), o processamento de dados em linguagem natural (...) é a função básica de qualquer sistema de informação” (p. 153). A autora complementa que um sistema de informação pode ser definido de uma maneira bem simples em termos de elementos chaves e funções básicas, sendo que todos envolvem processamento de linguagem (p. 153).

Warner (1987), em seu capítulo de revisão discute acerca da portabilidade e da robustez dos sistemas PLN. Segundo o autor, uma vez que os sistemas possam operar somente em domínios restritos, um dos grandes problemas é como aplicar as melhores técnicas de um domínio restrito num novo domínio (p. 90). O autor destaca inúmeras experiências com portabilidade de diversos sistemas (linguagem) e aponta a preocupação de vários pesquisadores diante da necessidade de um projeto modular alcançar a portabilidade. Quanto à robustez, o autor afirma que um sistema robusto deve processar qualquer entrada: parcial e/ou mal formada (*ill-formedness*), incluindo metáforas e o contexto das sentenças nos textos, e permitir um diálogo cooperativo entre os participantes. Ele questiona se é possível construir um sistema de linguagem natural completamente robusto, e em que medida este eventual sucesso depende de boas práticas de engenharia, e do conhecimento de processos cognitivos humanos.

O único autor que chamou a atenção para os vários níveis em que a

linguagem pode ser analisada foi Chowdhury (2003), ao sugerir que para compreender linguagens naturais é importante distinguir entre os seguintes sete níveis interdependentes, e adiantam que um sistema de processamento de linguagem natural pode envolver todos ou alguns destes níveis de análise: nível fonético ou fonológico, que trata da pronúncia; nível morfológico, que trata das menores partes das palavras, que carregam um significado, sufixos e prefixos; nível lexical, que trata do significado lexical das palavras e das partes de análises do discurso; nível sintático, que trata da gramática e da estrutura das sentenças; nível semântico, que trata do significado das palavras e das sentenças; nível do discurso, que trata da estrutura de tipos diferentes do texto usando estruturas do original e; nível pragmático, que trata do conhecimento que vem do mundo exterior, isto é, fora do conteúdo do documento.

Conforme mencionado anteriormente, dentre as inúmeras aplicações apresentadas pelos autores dos capítulos de revisão, algumas estão voltadas para o desenvolvimento da própria área de processamento de linguagem natural, enquanto que outras procuram atender um público mais amplo. Sendo assim, nas próximas seções serão apresentadas as aplicações voltadas para a própria área, seguida das aplicações de cunho mais prático. Sabe-se que estas duas categorias não são mutuamente exclusivas, no entanto, ariscou-se rascunhar essa separação tanto na estrutura de tópicos adotadas neste capítulo, como no instrumento de seleção construído. As aplicações voltadas para a própria área foram incluídas na categoria de técnicas e métodos, enquanto que as aplicações práticas foram incluídas na categoria Aplicações.

3.4.1. Aplicações para a própria área de PLN

Dentre as aplicações apresentadas pelos autores dos capítulos de revisão voltadas para a própria área, ou seja, que têm como objetivo o desenvolvimento da própria área, estão os analisadores sintáticos e semânticos, os processadores automáticos de dicionários e gramáticas, dentre outras.

Salton (1968) afirma que provavelmente a maioria dos desenvolvimentos em processamento de linguagem natural noticiados na sua época apresentava uma crescente tendência ao uso *online* e de técnicas com suporte de máquinas ao invés de métodos completamente automáticos para análise linguística (p. 180). Grande

parte dos trabalhos tem então sido dedicada à geração e manipulação de dicionários e tesauros mecanizados/mecânicos *(mechanized dictionaries),* e métodos de análise de linguagem usando gramáticas *online* (p. 180).

3.4.1.1. Processamento automático de tesauro

Já na década de 60 almejava-se utilizar recursos computacionais de hardware e de software tanto na construção automática de dicionários, tendo em vista o tempo gasto no processo manual de construção, como no desenvolvimento de sistemas baseados em dicionários.

Salton (1968), em seu capítulo de revisão, cita inúmeros trabalhos que fazem uso automatizado, ou pelo menos semi-automatizado de dicionários. Dentre eles, Galli e Yamada apresentam um dicionário inglês-inglês usado para classificar palavras em grupos morfológicos, como parte de um programa automático de controle de vocabulário (p. 182); Olney *et al.* também se basearam em dicionários mecanizados voltados para análise morfológica e semântica completa da linguagem (p. 182). Bachrach e Masterman usaram dicionários automáticos multilingues como componentes de um sistema de tradução semi-automático. Segundo Masterman (1967) o uso de dicionários automáticos, mesmo quando usados semi-automaticamente, tendem a procurar palavra por palavra e não fazer uma tradução global (p. 182). Salton (1968) finalizou afirmando que, embora a construção de dicionários completamente automática seja esperada em muitas aplicações, o que se tem observado é uma variedade de trabalhos que incluem dicionários preparados manualmente (p. 182).

Damerau (1976) apresenta inúmeros projetos significantes e afirma que todos os projetos de compreensão de linguagem devem contar com algum tipo de dicionário gerado mecanicamente ou não (p. 117).

3.4.1.2. Análise sintática

Dentre as aplicações voltadas para a própria área, alguns autores apontaram para a utilização do processamento de linguagem natural em análise sintática de modo interativo ou, em outras palavras, processamento interativo (ou *online*) de gramática.

Salton (1968) afirma que vários trabalhos incluíram análise sintática em modo interativo, onde as regras são aplicadas uma por uma, e sua aplicação na derivação de estruturas profundas ou de superfície para várias sentenças de entrada é demonstrada. O usuário tem então a opção de aceitar ou rejeitar certas regras, e então refinar a gramática (p. 183). Segundo Salton (1968), desde que a gramática seja suficientemente testada, ela pode ser aplicada a algumas tarefas de processamento de linguagem – geração de estruturas profundas, e ser usada posteriormente em respondedores automáticos ou em sistemas de recuperação de dados (p. 183). Ainda segundo Salton (1968), no estado atual das pesquisas (pequenas abordagens do uso de gramáticas *online* com propósitos restritos), é difícil dizer se este modo *online* realmente produz ambientes que resultem em uma simplificação substancial do processo de análise da linguagem (p. 183).

3.4.2. Aplicações Práticas

Nesta seção serão apresentadas as aplicações mais enfatizadas ao longo dos 40 anos analisados nos capítulos de revisão do ARIST, dentre elas, tradução automática por máquina, respondedores automáticos, geração automática de linguagem, incluindo geração, sumarização e compreensão, recuperação de informação, dentre outras citadas pelos autores tais como, indexação, análise de estilo, análise de conteúdo, e outras.

3.4.2.1. Tradução automática

Simmons (1966) inicia o seu artigo de revisão afirmando que tradução automática de linguagem era uma aplicação atrativa para computadores, e complementa que a maioria dos projetos por ele apresentados redireciona seus esforços para estudos básicos da estrutura da linguagem, e considera tradução automática um objetivo distante de ser alcançado. Apesar disto, Simmons (1966) destaca inúmeros projetos que conseguiram atingir o objetivo de produzir traduções em sentenças.

Salton (1968) inicia afirmando que trabalhos práticos em tradução completamente automática pareciam estar mais ou menos paralisados. Apesar disto, Salton (1968) destaca o desenvolvimento de alguns estudos teóricos, além de

trabalhos de tradução assistida por computador, nos quais sistemas exaustivos de análise são trocados pela presença de um humano, e da avaliação da exatidão (*accuracy*) e da eficiência de vários sistemas de tradução (p. 190).

Kay e Sparck-Jones (1971) afirmaram que quase toda a pesquisa em tradução automática se baseia na hipótese de que qualquer solução de sucesso deve se basear em teoria linguística e que um programa de computador refletirá a estrutura da teoria em sua arquitetura (p. 153).

Ao apresentar sistemas de tradução automática, Damerau (1976) considerou dois tipos de sistemas: aqueles que precisam de um retorno *(feedback)* humano para produzir saídas corretas, e aqueles que são inseridos em sistemas de produção. Damerau (1976) complementa que ao mesmo tempo em que pesquisas em tradução por máquina trazem vantagens (existe um teste com validade natural), trazem também desvantagens inerentes a necessidade de se usar duas linguagens simultaneamente (p. 133).

Segundo Chowdhury (2003), autor do último capítulo de revisão do ARIST sobre processamento de linguagem natural, com a proliferação da web e das bibliotecas digitais, a recuperação de informação multilíngue transformou-se num dos principais desafios, onde existem dois conjuntos de questões: (1) reconhecimento, manipulação e exibição de múltiplas linguagens, permitindo que os usuários acessem a informação em qualquer linguagem que esteja armazenada; e (2) busca e recuperação da informação em linguagem cruzada, permitindo que os usuários especifiquem suas necessidades de informação em sua linguagem preferida. A tradução do texto, segundo Chowdhury (2003), pode ocorrer em dois níveis: (1) tradução de um texto completo de uma linguagem para outra, com a finalidade de buscar e recuperar; e (2) tradução das perguntas de uma linguagem para uma ou mais linguagens diferentes. Além disso, Chowdhury (2003) afirmou que o projeto de sistemas de tradução automática é um trabalho duro, pois requer seleção cuidadosa de modelos e algoritmos, e que tal problema está longe de ser resolvido, visto que “a linguagem humana é uma área rica e fascinante cujos tesouros somente começaram a ser explorados” (p. 23).

Assim, pode-se observar que tradução automática foi uma aplicação discutida praticamente ao longo dos 40 anos, sendo que nos últimos anos observou-se uma preocupação na tradução multilíngue, principalmente voltada para recuperação de informação.

3.4.2.2. Respondedores automáticos

Simmons (1966) inicia alertando que, embora tenha afirmado que um progresso significativo tem sido feito em pesquisas em respondedores automáticos, deve-se enfatizar que sistemas de perguntas e respostas em linguagem natural completamente automáticos era um objetivo muito distante, além de depender da realização de toda a área de processamento de linguagem. Apesar disto, Simmons (1966) cita um outro trabalho de sua autoria (SIMMONS, 1965) contendo 15 sistemas de perguntas e respostas em linguagem natural.

Segundo Salton (1968), existem alguns modelos para respondedores automáticos: Woods, Bobrow e Fraser, todos de 1968. Algumas páginas do seu artigo de revisão são dedicadas à abordagem de Woods para sistemas respondedores automáticos. Segundo ele, tal sistema é composto por três partes: incluindo um analisador sintático, o qual gera a estrutura profunda de uma consulta de entrada; um interpretador semântico, usado para obter o significado da consulta em termos de certas entidades formais especificadas por uma base de dados; e finalmente, um recuperador, que casa (combina) as estruturas semânticas obtidas do interpretador semântico da consulta com as estruturas reconhecidas na base de dados e constrói uma resposta apropriada (p. 177). Segundo Salton (1968), esta base de dados armazenada no modelo de Woods consiste de: objetos, assim como funções que mapeiam um conjunto de objetos em outros; relações, que substituem alguns verbos e modificadores preposicionais em linguagem natural; e preposições, que são instâncias de relações específicas entre objetos. A interpretação semântica é então, efetivamente, uma tradução da estrutura sintática de uma sentença em uma expressão em uma linguagem de consulta formal, representando o significado da sentença em termo de predicados, funções, e comandos entendidos pelo sistema. O significado de cada predicado, função ou comando é definido por uma subrotina programada que gera um valor verdade para um dado predicado, ou computa um valor funcional para os argumentos que podem ser apresentados em uma dada expressão (p. 177). Segundo Salton (1968), no sistema de Woods, a interpretação semântica é especificada por um conjunto de regras de substituição, semelhante às regras de reescritas usadas na gramática de estrutura de frase.

Segundo Salton (1968), o respondedor de Woods era o mais avançado trabalho na área e demonstrava que a complexidade das regras e o tamanho da

gramática realmente precisam lidar com consultas em um ambiente restrito. No entanto, Salton (1968) finaliza questionando se tais sistemas podem ser adaptados para outras bases de dados e para diferentes ambientes (p. 179). Salton (1968) destaca ainda que inúmeros pesquisadores, que trabalham com sistemas de recuperação de documentos e não com respondedores automáticos, acreditam que uma interpretação semântica completa de cada entrada de texto e requisições de busca é necessária se for para o sistema operar eficazmente. Outros, entretanto, acreditam que uma única interpretação de texto não é desejável em recuperação de documentos (p. 180).

Becker (1981) afirma que as capacidades dos atuais sistemas de recuperação bibliográfica e de processamento de palavras são muito diferentes das capacidades dos respondedores automáticos. Para grande parte, os sistemas de processamento de palavras e recuperação de informação não usam representação de significado por computador, não objetivam o significado, não checam gramáticas, não identificam outros erros além daqueles de *spelling* inválido.

3.4.2.3. Análise de estilo

Simmons (1966) cita Sedelow (1966) que define o subcampo do estilismo computacional (*stylistics computational*), o qual inclui indexação, *frequency counting* e concordância. O subcampo também inclui técnicas para identificar palavras por critérios semânticos e sintáticos. Montgomery (1969) descreve projetos de processamento de linguagem baseados em heurística como o citado por Garvin (1968). Complementa que na área de humanas, a maioria das aplicações é voltada para análise de estilo, ou seja, distinguir autores e falantes a partir da identificação de padrões. Kay e Sparck Jones (1971) destacam a área de análise de estilo e apontam para a tendência de escolha de palavras normalmente feitas pelo autor de um documento (palavras que fazem parte do seu vocabulário)(p. 157). Haas (1996) finaliza afirmando que abordagem estatística também tem sido usada para produzir evidências de estilo para determinar relacionamentos entre autores. Segundo Chowdhury (2003), os métodos estatísticos são usados em PLN para inúmeras finalidades, por exemplo, para remoção de ambiguidade de palavra (*word disambiguation*), para gerar gramáticas e *parsers*, para determinar evidências de estilos dos autores e falantes, assim por diante.

3.4.2.4. Geração automática de linguagem

Ao analisar os capítulos de revisão do ARIST utilizados, foi possível observar que o tópico sobre geração automática de linguagem tem sido discutido sob vários focos, além de serem relacionados a outras aplicações. Salton (1968) destacou que geração automática de frases, com o propósito de identificação de conteúdo, usa principalmente critérios estatísticos para selecionar as unidades de conteúdo, incluindo estatísticas de co-ocorrência para vários componentes da frase, assim como frequência individual relativa e absoluta dos componentes. O modelo estatístico é então complementado por informações gramaticais mínimas consistindo de coeficientes de probabilidade para as classes sintáticas de cada componente da frase (p. 187). A técnica de escolha destas unidades é interessante e leva a buscas com alta precisão. Entretanto, somente 25% a 45% dos pares de palavras válidos são realmente selecionados, levando a uma baixa recuperação (p. 187).

Segundo Haas (1996), a geração automática de linguagem incorpora várias questões importantes do PLN. Primeiro, o sistema deve determinar a resposta apropriada para a pergunta (considerando a idade da audiência, o nível de conhecimento e o propósito do texto), e como esta resposta está representada. As aplicações de geração automática de linguagem podem ser divididas em: geração de texto (*text generation*), sumarização (*text sumarization*) e compreensão de texto (*text understanding*).

Segundo Warner (1987), geração de texto é uma outra área de pesquisa que trata as sentenças como unidades linguísticas conectadas, ao invés de elementos discretos. Warner (1987) destaca as áreas nas quais a geração de linguagem é importante, tais como respondedores automáticos, comunicação com sistemas especialistas, e sumarização de texto, considerada uma aplicação não-interativa de geração de texto (p. 94).

3.4.2.4.1. Sumarização

Segundo Haas (1996), geração automática de resumos de artigos ou de outros tipos de documentos compartilha características com a geração de texto. Haas (1996) destaca que na área de sumarização é possível distinguir extratos, que podem ser criados por sentenças selecionadas do documento original, de sumário,

que envolve a criação de novas sentenças baseadas na informação contida no documento original. Estes métodos de seleção utilizam frequência de termos, palavras-chave do título, a localização das sentenças no documento, as frases que indicam pontos importantes além de outros tipos de evidências. Haas (1996) finaliza afirmando que avaliar um sumário é tão difícil quanto outro tipo de geração de texto.

Segundo Chowdhury (2003), abstração e sumarização automática de texto têm sido usadas de maneira indistinta, com o objetivo de gerar resumos (*abstract*) ou sumários (*summaries*) dos textos. Esta área de pesquisa de PLN está se tornando mais comum na web e em ambientes de bibliotecas digitais. Em sistemas simples de abstração ou de sumarização, as partes do texto - sentenças ou parágrafos - são selecionadas automaticamente baseando-se em critérios linguísticos e/ou estatísticos para produzir o resumo ou o sumário. Interesses recentes em sumarização ou abstração automática de texto são refletidos no crescente número de publicações que aparecem em inúmeras conferências internacionais e workshops incluindo a ACL, a ACM, a AAAI, o SIGIR, e vários capítulos nacionais e regionais das associações. Diversas técnicas são usadas para a sumarização ou abstração automática de texto (p. 9). A maioria dos sistemas de sumarização automática de texto trabalha de maneira satisfatória dentro de uma coleção pequena de texto ou dentro de um domínio restrito. Construir sistemas robustos e independentes de domínio é uma tarefa complexa e que requer muito recurso computacional. Entretanto, experimentos recentes quanto à utilidade da extração automática de palavras-chaves em textos completos, no processo real de sumarização mostraram uma variação considerável entre assuntos, quando comparados à sumarização humana, e que somente 37% dos assuntos encontrou as palavras-chaves e frases úteis para escrever seus sumários.

3.4.2.5. Recuperação de Informação

Para finalizar as discussões inerentes às aplicações apresentadas pelos autores dos capítulos de revisão, procurou-se destacar a utilização de processamento de linguagem natural na recuperação de informação, que tem características diferentes daquele voltado para tradução automática ou para sistemas respondedores automáticos, uma vez que o nível de corretude exigido é diferenciado.

Salton (1968) afirma ser possível fazer uso de uma análise mais simplificada pra extrair o conhecimento de documentos textuais rapidamente e com baixo custo. Assim, não se preocupa em fazer uma completa desambiguação, ao invés disso, a tendência é fazer o que é facilmente feito a mão, e verificar como isto pode ser realmente aplicado a técnicas de processamento de texto (p. 183). Salton (1968) complementa que métodos simplificados de análise sintática baseados principalmente na presença ou na ausência de um tipo de palavra têm sido usados com diferentes propósitos. Segundo ele, estes são atrativos porque necessitam de um aparato relativamente pequeno. No entanto, para o propósito de recuperação de informação, tais métodos são considerados deficientes, uma vez que representação de documentos usando frases sintáticas completas, ao invés de identificadores simples tais como termos únicos, tende a especificar melhor o conteúdo do documento. Além disso, a geração de falsas frases pode também diminuir a precisão por ajudar a retornar itens que não são de fato desejados (p. 185). Salton (1968) cita inúmeros projetos que parecem confirmar os resultados que indicam que frases sintáticas derivadas de sentenças dos documentos representam uma ferramenta de indexação que é normalmente muito específica para propósitos de recuperação de documentos (p. 185). Ainda segundo o autor, um número de estudos havia sido iniciado com o propósito de melhorar o desempenho da recuperação alcançada com métodos de casamento de frases sintáticas, usando procedimentos mais refinados de análise sintática, ao invés de substituir a sintaxe por processos estatísticos alternativos (p. 186). Outro esforço citado por Salton (1968) é o trabalho de Sager (1966): um sistema de recuperação baseado em casamento de frases contidas nos documentos e nas requisições de consultas. Durante a análise, também é usado um dicionário semântico contendo sinônimos, uma variedade de relações de termos, definições, etc. Salton (1968) lamenta que, como o sistema não foi implementado, não era possível avaliar a sua eficiência. Complementa que a precisão parece ser alta porque poucas frases erradas são normalmente geradas. Por outro lado, a revocação (*recall*) deve ser baixa comparando-se com os obtidos por outros métodos menos poderosos (p. 187).

Kay e Sparck Jones (1971) afirmam que, além dos sistemas respondedores automáticos fazerem parte do campo de recuperação de informação, não está clara a divisão entre sistemas de recuperação de documentos e de fatos. Segundo os autores, recuperação de informação é um processo no qual citações

são retornadas em resposta a perguntas, ou seja, no qual perguntas são respondidas. Assim, a diferença principal está no tipo de perguntas e no tipo de resposta. Os autores citam como exemplo a necessidade de saber qual é a população do Peru, que provavelmente seria mais facilmente atendida por um sistema de recuperação de fatos ou respondedor automático, enquanto que sistemas de recuperação de documentos trariam citações de livros que tratam tal assunto.

Segundo Warner (1987), outro tópico relevante em PLN é como processar a linguagem natural dentro de um contexto, ou seja, sentenças inseridas em um diálogo. Neste caso, a ênfase deve ser em desenvolver sistemas flexíveis e cooperativos que não somente apresentem uma resposta, mas que apresentem advertências, antecipem necessidades, etc. Segundo Warner (1987), alguns sistemas de recuperação de fatos em banco de dados resolvem ambiguidades e entradas mal-formadas. Dentre as pesquisas citadas pelo autor estão técnicas para reconhecer mudança no foco em um diálogo, sistemas que corrigem incoerências (*misconceptions*) a partir do diálogo com o usuário, que é orientado a reformular a sua pergunta, visto que é raro o usuário ser claro, direto e preciso nas suas colocações (p. 93).

Haas (1996) afirma que na sua época, recuperação de informação inclui pelo menos quatro diferentes aplicações: recuperação de documentos, recuperação de parágrafos/passagens, classificação de documentos e extração de informação. Segundo Haas (1996), as técnicas de PLN podem ser utilizadas em vários pontos em processos de RI, mas talvez sejam mais comumente usadas na criação e no casamento de representações do documento e da consulta. Haas (1996) cita Lewis e Sparck Jones por sugerirem que as técnicas de PLN são especialmente importantes para indexar e ajudar os usuários a formularem consultas mais efetivas.

Finalmente, pôde-se observar que, apesar dos autores dos capítulos de revisão apresentarem inúmeras aplicações, e somente nos últimos anos descreverem a importância do processamento de linguagem natural na área de recuperação de informação, é evidente que as técnicas aqui discutidas podem e devem ser aplicadas no desenvolvimento da maioria dos sistemas de informações requeridos atualmente.

4. Resultados

Neste capítulo pretende-se apresentar os resultados obtidos a partir da análise horizontal (usando os atributos descritivos) de todas as 621 publicações atinentes à área de processamento de linguagem natural (seção 4.1), e da análise vertical das 68 publicações submetidas à análise de conteúdo (seção 4.2).

4.1. Análise Horizontal das publicações

Nesta seção são apresentados os resultados obtidos analisando-se as características descritivas das publicações consideradas atinentes para a área de processamento de linguagem natural, de acordo com o critério de seleção automático, e confirmado pela validação humana. Dentre os resultados apresentados, estão análises estatísticas envolvendo a distribuição dos pesquisadores autores dessas publicações por área de vinculação, por produção científica e por temáticas ao longo dos anos.

Na TAB. 2 é apresentada a distribuição das publicações ao longo dos anos, de 1973 a 2009. Pode-se observar que 70% das publicações na área de PLN foram publicadas após o ano 2.000. O número, relativamente baixo, de publicações em 2.009 pode ser atribuído ao fato dos dados terem sido coletados em novembro do ano, e depender da atualização dos próprios pesquisadores, o que normalmente não ocorre de imediato. Em função disso, na FIG. 5 optou-se por não apresentar os dados de 2.009 para não causar a falsa impressão de que a produção científica da área diminuiu. Na FIG. 6 é apresentada a evolução acumulativa do número de publicações na área de PLN, destacando que entre os anos de 2.009 e 1.999 foram publicadas quase 75% de todo o volume de trabalhos na área.

TABELA 2
Distribuição das publicações por ano: 1973-2009.

Ano da Publicação	Número de publicações por ano	Percentual das publicações	Percentual acumulado
2009	23	3,7%	3,7%
2008	56	9,0%	12,7%
2007	50	8,1%	20,8%
2006	54	8,7%	29,5%
2005	47	7,6%	37,0%
2004	50	8,1%	45,1%
2003	60	9,7%	54,8%
2002	42	6,8%	61,5%
2001	24	3,9%	65,4%
2000	26	4,2%	69,6%
1999	30	4,8%	74,4%
1998	29	4,7%	79,1%
1997	29	4,7%	83,7%
1996	18	2,9%	86,6%
1995	13	2,1%	88,7%
1994	16	2,6%	91,3%
1993	9	1,4%	92,8%
1992	5	0,8%	93,6%
1991	2	0,3%	93,9%
1990	2	0,3%	94,2%
1989	6	1,0%	95,2%
1988	6	1,0%	96,1%
1987	7	1,1%	97,3%
1986	6	1,0%	98,2%
1985	4	0,6%	98,9%
1984	1	0,2%	99,0%
1983	1	0,2%	99,2%
1982	1	0,2%	99,4%
1981	1	0,2%	99,5%
1980	1	0,2%	99,7%
1975	1	0,2%	99,8%
1973	1	0,2%	100,0%
Total	**621**	**100,0%**	**----**

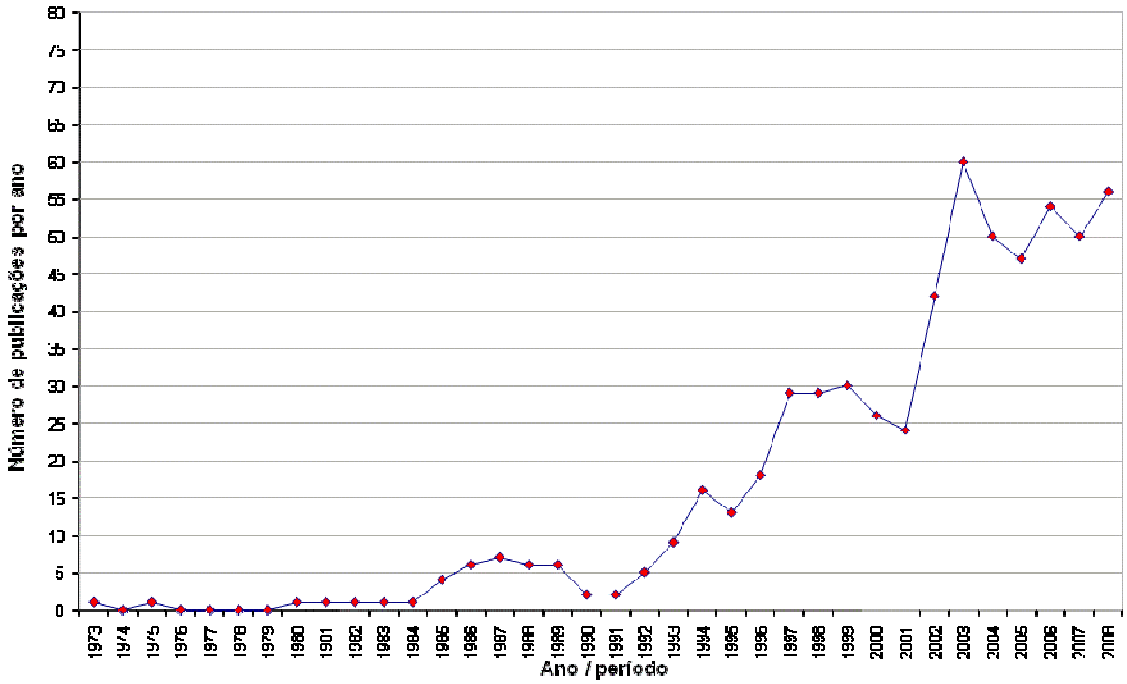

FIGURA 5 – Evolução anual das publicações: 1973-2008.

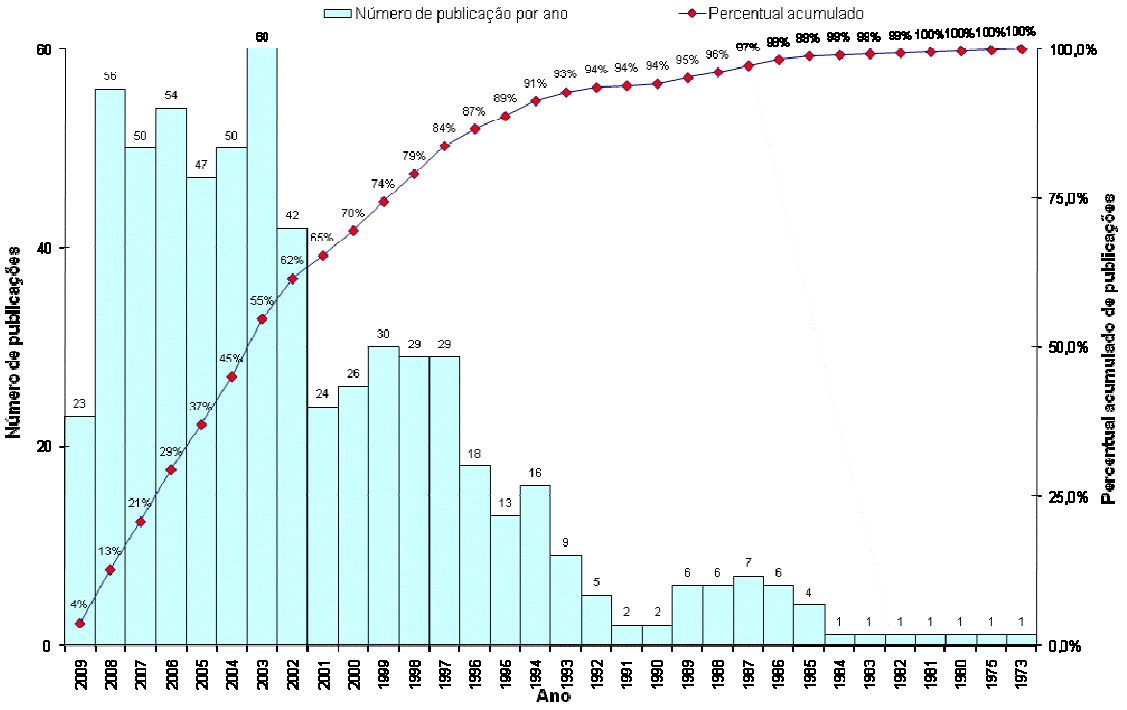

FIGURA 6 – Distribuição acumulativa das publicações: 2009-1973

Com o intuito de avaliar as áreas disciplinares que mais contribuem para o desenvolvimento da pesquisa científica na área de PLN, optou-se por analisar as publicações de acordo com a sua área de vinculação. Considerou-se que a área da

publicação é determinada pela área do seu primeiro autor. Na FIG. 7 é possível observar que 38% (236) das publicações foram publicadas por pesquisadores da área da ciência da computação, enquanto que 34% (213) das publicações tiveram como primeiro autor pesquisadores que não possuíam currículo cadastrado na Plataforma Lattes (pelos motivos discutidos anteriormente). Em seguida, tem-se a linguística com 15% (95) das publicações. Vale relembrar que a área 'desconhecida' foi atribuída aos pesquisadores que não apresentaram currículo cadastrado na Plataforma Lattes, enquanto que 'não informada' foi atribuída aos pesquisadores que, apesar de possuírem currículo cadastrado, não informou, em campo apropriado, a sua área de vinculação. Além disso, 70% dos autores "desconhecidos" foram co-autores de somente uma publicação, o que sugere que a análise pode ser feita, desconsiderando-se esse perfil.

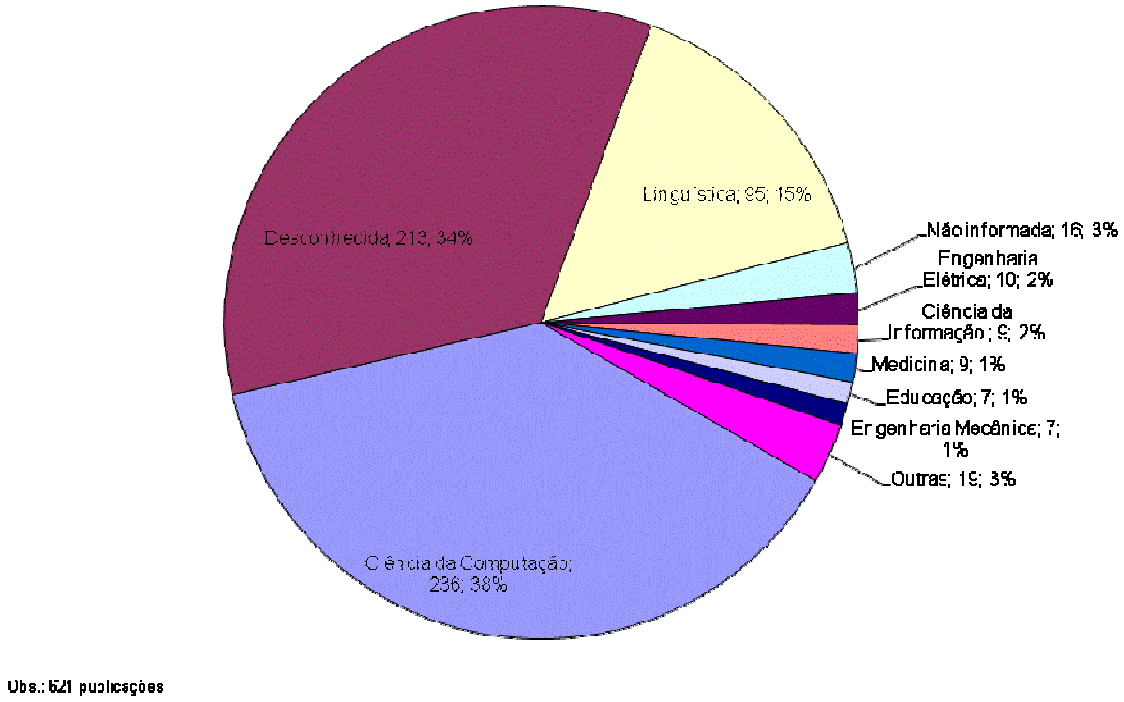

FIGURA 7 – Área das publicações conforme o primeiro autor: 1973-2009

Considerando-se agora apenas as 392 publicações que apresentaram definição de área[1], ou seja, que o primeiro autor possuía área de vinculação definida, é possível observar que a ciência da computação e a linguística juntas são responsáveis por 84% dessas publicações: com 60% e 24%, respectivamente (FIG. 8).

[1] Excluindo-se as publicações cujo primeiro autor era desconhecido (213) ou não informado (16).

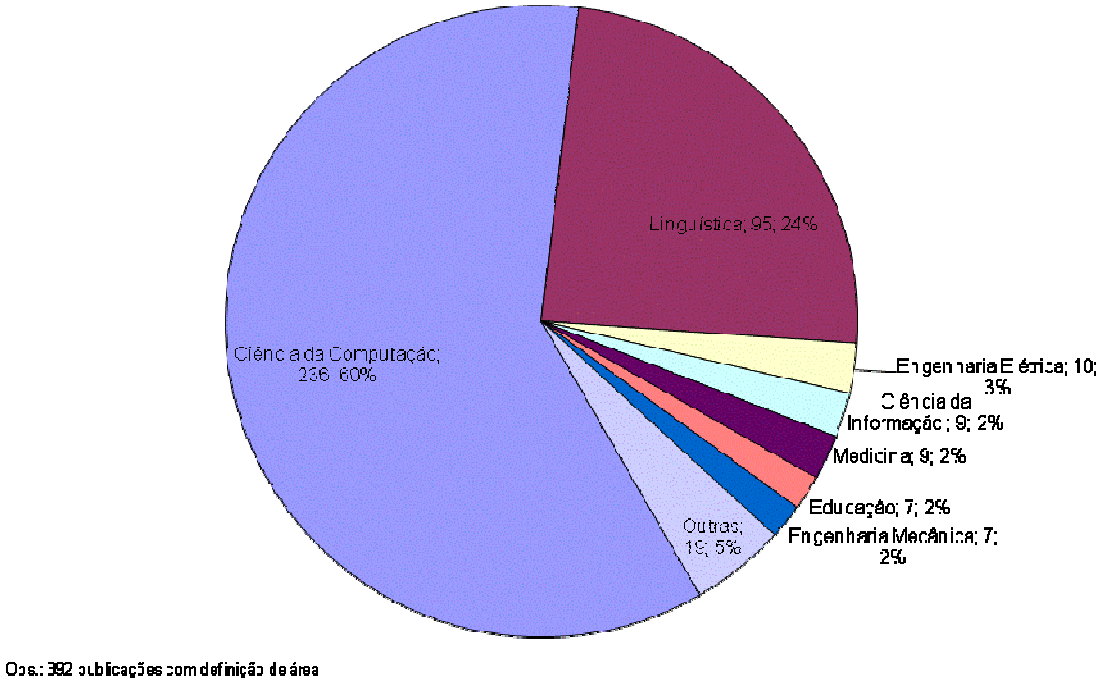

FIGURA 8 – Área das publicações conforme o primeiro autor (1973-2009): análise excluindo as publicações sem definição de área

Na FIG. 9 é apresentada a evolução das áreas da ciência da computação, linguística e ciência da informação, ao longo dos anos. É possível observar que quantitativamente todas as áreas apresentaram um aumento na produção científica após o ano 2.000.

Analisando-se por década (FIG. 10), é possível observar que, proporcionalmente, na década de oitenta, a ciência da computação foi a área mais produtiva, enquanto que a década de noventa, foi a década mais produtiva para a área da linguística. Além disso, os dados apresentados sugerem que a ciência da informação, na década de noventa, recuou a sua contribuição, tentando recuperar nos anos 2.000 (diminuindo de 4% para 1%, e voltando para 3%).

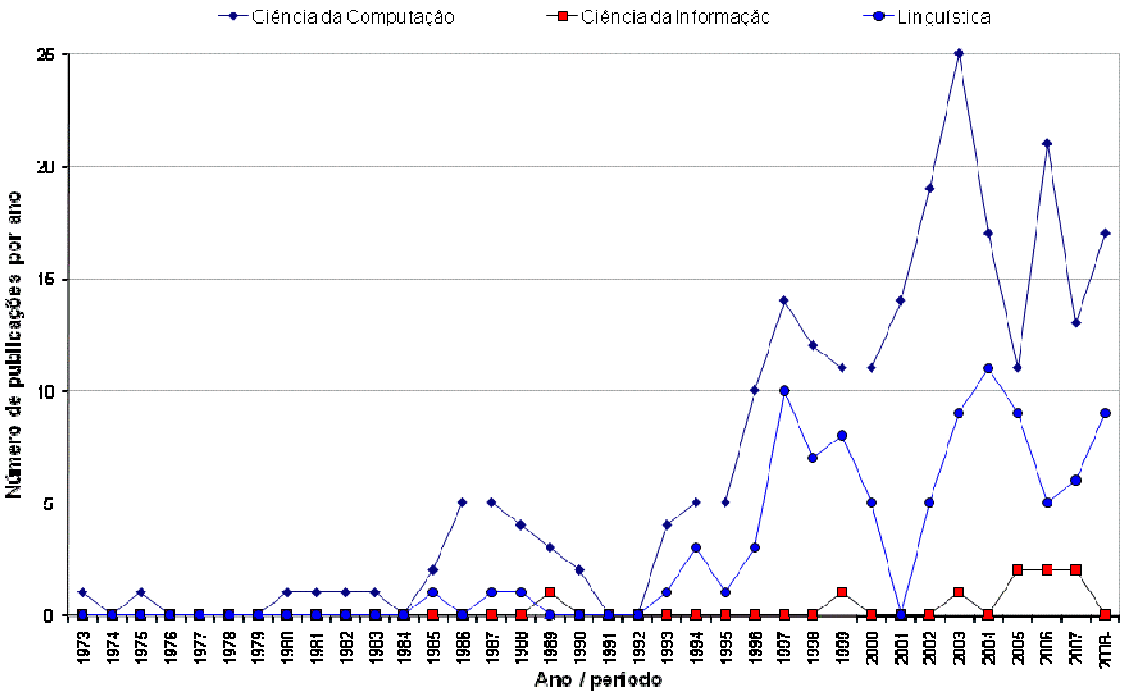

FIGURA 9 – Evolução anual das áreas das publicações definidas pelo primeiro autor: 1973-2009

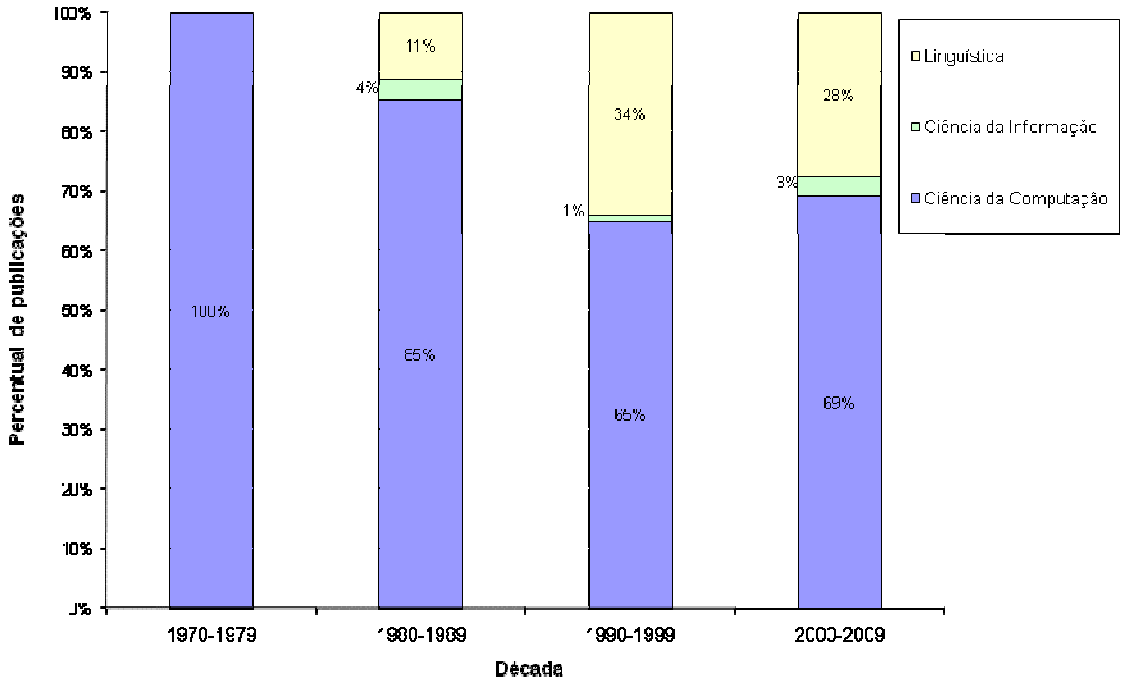

FIGURA 10 – Evolução das áreas das publicações definidas pelo primeiro autor: análise por década (1973-2009)

Na TAB. 3 são apresentados os 12 pesquisadores que mais produziram na área de PLN, e que juntos foram responsáveis por mais de 20% de toda a produção nacional. Vale destacar que dentre eles, nove são da área de ciência da computação, dois são da linguística, e um é da engenharia elétrica. Além disso, dentre eles, sete fazem parte do Núcleo Interinstitucional de Linguística Computacional (NILC), sendo a profa. Dra. Maria Das Graças Volpe Nunes, coordenadora do projeto. O NILC foi criado em 1993 para dar suporte a pesquisa e desenvolvimento de projetos em linguística computacional e processamento de linguagem natural. Originalmente foi concebido por cientistas da computação da Universidade de São Paulo (USP) em São Carlos, mas inclui cientistas da computação e linguistas da Universidade Federal de São Carlos (UFSCar) e da Universidade Estadual Paulista (UNESP) de Araraquara.

TABELA 3
Distribuição das publicações por pesquisador

Nome do Pesquisador	Área de Vinculação	# Publicações	% Publicações	% Acumulada
Maria Das Graças Volpe Nunes	Ciência da Computação	78	4,6%	4,6%
Vera Lucia Strube De Lima	Ciência da Computação	47	2,8%	7,4%
Thiago Alexandre Salgueiro Pardo	Ciência da Computação	41	2,4%	9,8%
Lucia Helena Machado Rino	Ciência da Computação	30	1,8%	11,6%
Renata Vieira	Ciência da Computação	24	1,4%	13,0%
Helena De Medeiros Caseli	Ciência da Computação	22	1,3%	14,3%
Sandra Maria Aluisio	Ciência da Computação	21	1,2%	15,5%
Bento Carlos Dias Da Silva	Linguística	21	1,2%	16,8%
João Luis Garcia Rosa	Linguística	18	1,1%	17,8%
Fernando Gil Vianna Resende Junior	Engenharia Elétrica	16	0,9%	18,8%
Ariadne Maria Brito Rizzoni Carvalho	Ciência da Computação	15	0,9%	19,7%
Aline Villavicencio	Ciência da Computação	14	0,8%	20,5%
Outros pesquisadores (991)	Diversas	1.347	80%	100%

Outro aspecto analisado nas 621 publicações relevantes foi o fato de que grande parte das pesquisas desenvolvidas na área de PLN, aproximadamente 64%, envolve pesquisadores de várias áreas. Além disso, é possível observar na TAB. 4 que essa multidisciplinaridade dobrou entre a década de 80 e os anos de 2.000.

Analisando-se as temáticas apresentadas nos títulos de todas as 621 publicações relevantes, observou-se que alguns termos não apareceram em nenhuma delas: morfema, sinônimo, antônimo, hiponímia e metonímia, da categoria[2] de conceitos linguísticos; e análise de estilo e análise de discurso, da categoria Aplicações.

TABELA 4
Distribuição anual das publicações envolvendo multidisciplinaridade

Década	Publicações por ano	Publicações envolvendo multidisciplinaridade	% publicações envolvendo multidisciplinaridade
1970-1979	2	0	0%
1980-1989	34	12	35%
1990-1999	153	85	56%
2000-2009	432	312	72%
Total	**621**	**397**	**64%**

Observou-se que, dentre os conceitos computacionais, os termos mais frequentes foram ‘automático’, que apareceu em 24% das publicações, e ‘computacional’, que apareceu em 21% das publicações. Dentre os conceitos linguísticos, os termos mais frequentes foram 'português', em 46% das publicações e ‘linguagem natural’, em 36% delas. O interessante foi observar que dentre as aplicações que mais apareceram nos títulos analisados estão 'tradução', que ocorreu em 27% das publicações, 'sumarização', também em 27%, 'indexação', 'classificação' e 'recuperação', em 11% das publicações cada um. Dentre as técnicas abordadas nos trabalhos e evidenciadas nos títulos, observou-se que o termo mais frequente foi 'léxico', que apareceu em 23% das publicações, 'gramática', em 22% das publicações e 'parser', também em 22% delas.

Para cada categoria, procurou-se analisar como essas temáticas foram investigadas ao longo dos anos. As FIG. 11 a 14 apresentam a evolução dos termos por década, sem considerar o período de 1970-1979 em função do pequeno número de artigos (apenas dois).

A partir da FIG. 11 é possível observar que da década de 90 para os anos 2.000, o termo ‘computacional’ passou a ser mais utilizado. Na FIG. 12 é possível observar que da década de 90 para os anos 2.000 houve uma inversão de utilização dos termos indexadores: português passou a ser utilizado com mais frequência, enquanto que a expressão ‘linguagem natural’ deixou de ser usada.

[2] Categorias do critério de seleção automática criado a partir da análise de assunto dos capítulos do ARIST.

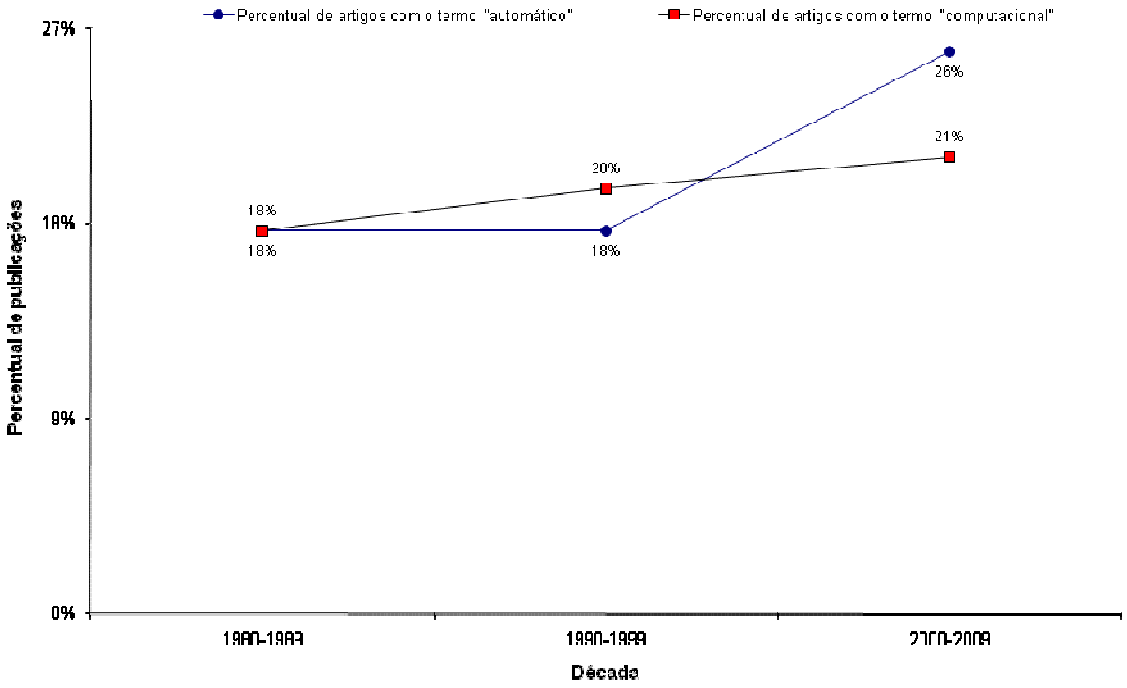

FIGURA 11 – Evolução dos principais termos dentre os conceitos computacionais: análise por década (1980-2009)

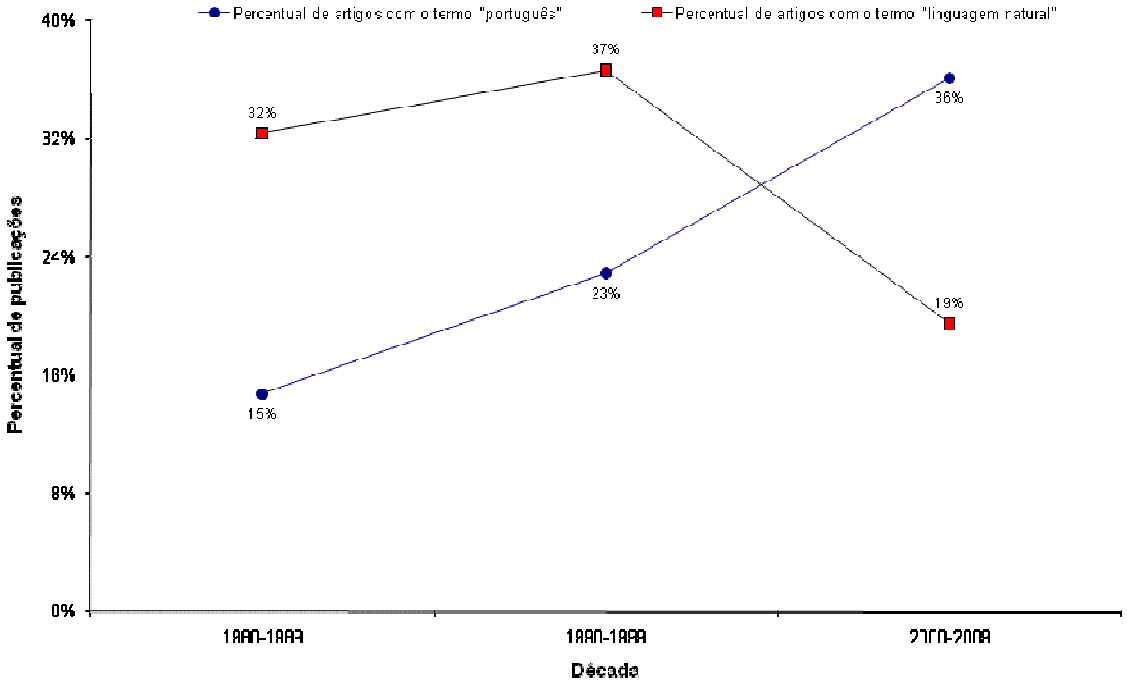

FIGURA 12 – Evolução dos principais termos dentre os conceitos linguísticos: análise por década (1980-2009)

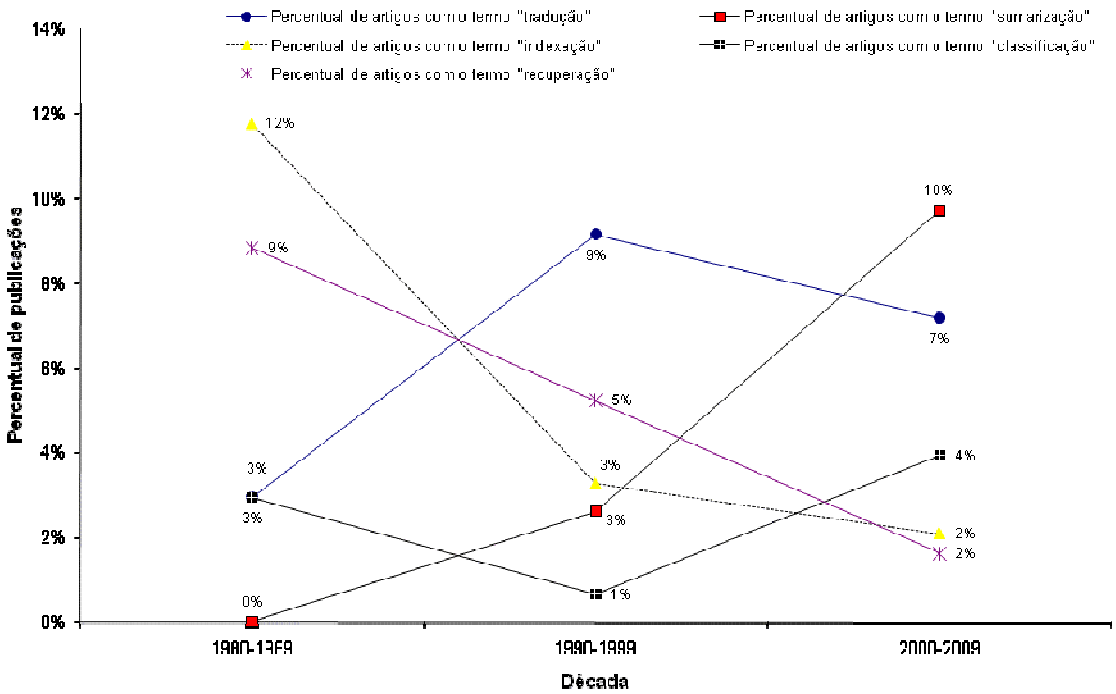

FIGURA 13 – Evolução dos principais termos dentre as aplicações: análise por década (1980-2009)

Na FIG. 13 é possível observar que: a tradução foi intensamente abordada na década de 90; os estudos com indexação diminuíram consideravelmente a partir da década de 80; que as pesquisas sobre classificação passaram por um período de dormência na década de 90; e que existe uma tendência clara na área de PLN de desenvolvimento de pesquisas em sumarização automática.

A partir da FIG. 14 é possível observar que, das três técnicas mais frequentes nos títulos das publicações analisadas, duas (*parser* e gramática) deixaram de ser priorizadas a partir da década de 90.

As próximas análises têm como objetivo avaliar as temáticas que cada área têm pesquisado. Na FIG. 15 é possível observar que a ciência da informação usa intensamente os termos 'sistema' e 'automático', a linguística prioriza os termos 'processamento' e 'computacional', enquanto que a ciência da computação usado todos.

Já na FIG. 16 é possível verificar que nenhum dos termos linguísticos mais frequentes foi usado pelas publicações da área da ciência da informação. Vale lembrar que a área da publicação foi determinada em função da área do seu primeiro autor. Além disso, os termos 'linguagem natural' e 'português' foram os mais

usados pela área da linguística e da ciência da computação.

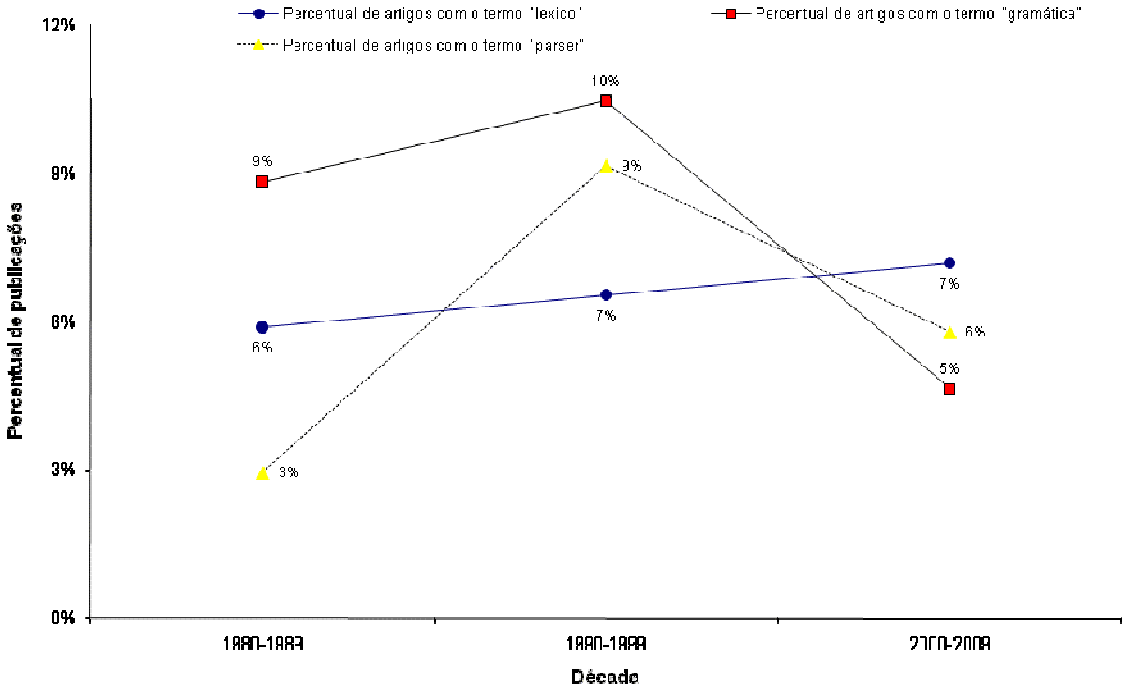

FIGURA 14 – Evolução dos principais termos dentre as técnicas: análise por década (1980-2009)

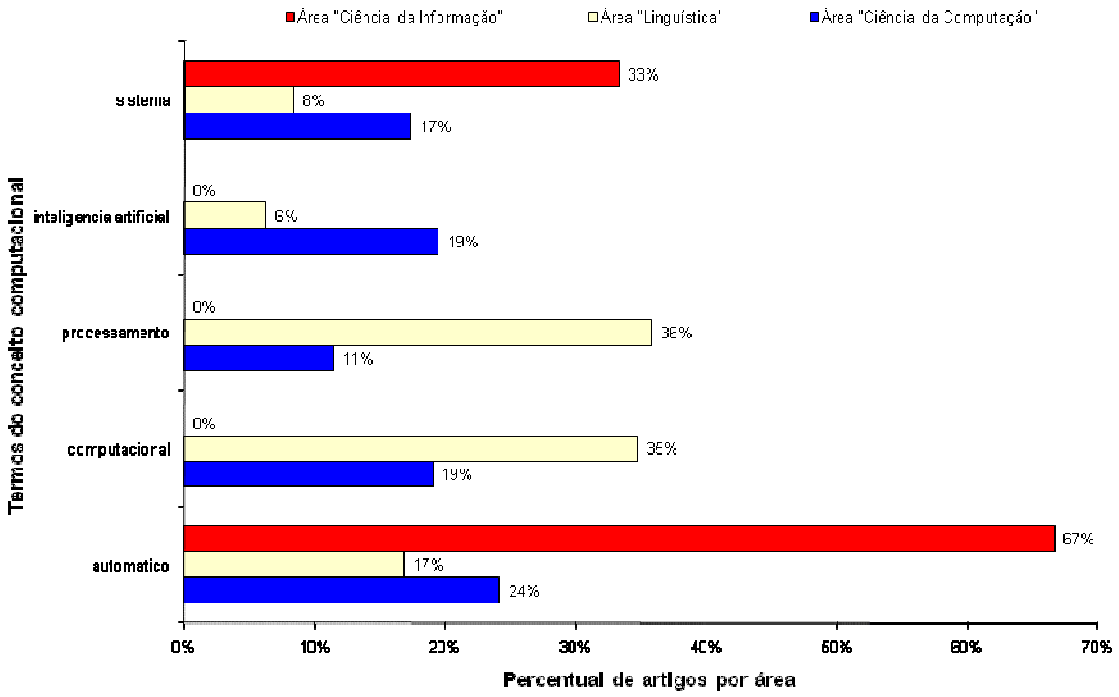

FIGURA 15 – Percentual de artigos de cada área com os principais termos dos conceitos computacionais

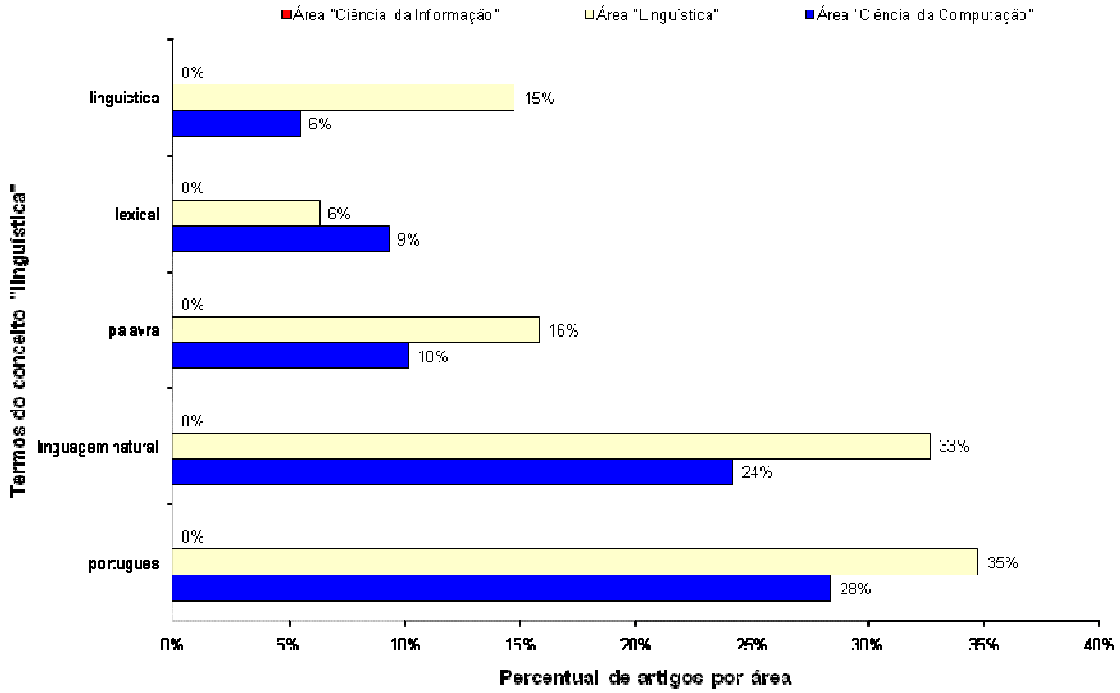

FIGURA 16 – Percentual de artigos de cada área com os principais termos dos conceitos linguísticos

A FIG. 17 revela como a área da ciência da informação tem priorizado as pesquisas em indexação automática, enquanto que a ciência da computação tem priorizado as pesquisas em tradução e sumarização, e a linguística em praticamente nenhuma (um pouco de tradução).

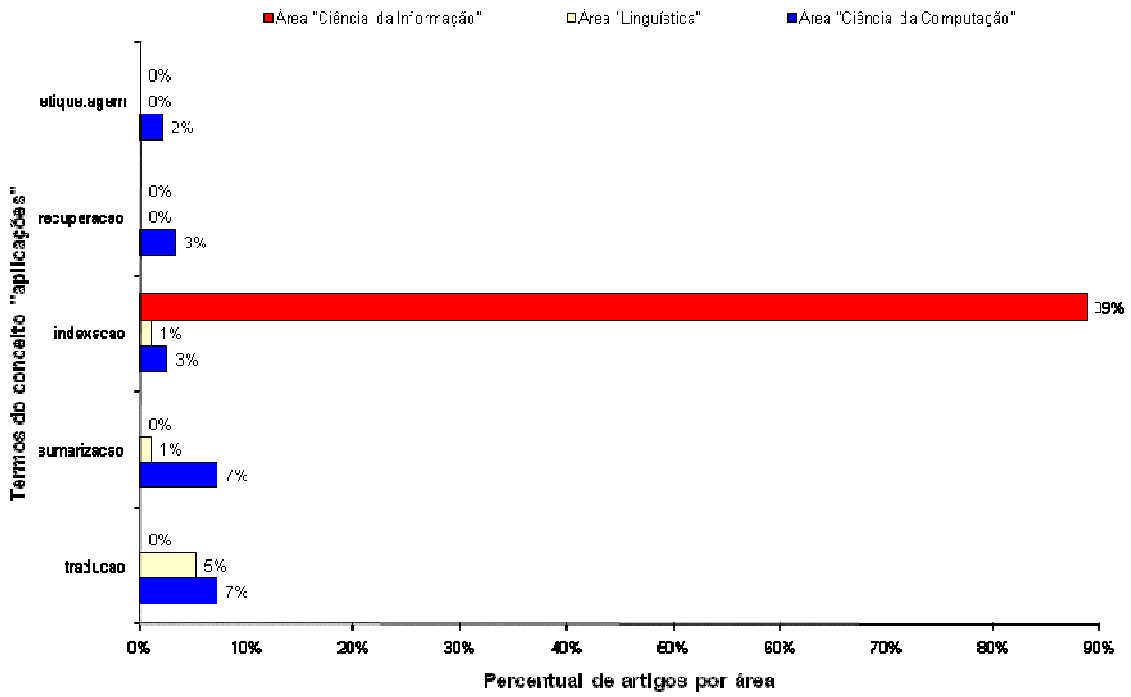

FIGURA 17 – Percentual de artigos de cada área com os principais termos dentre as aplicações

Já a FIG. 18 mostra que, das principais técnicas investigadas, a ciência da informação não priorizou nenhum delas, a área da linguística priorizou o estudo do léxico, enquanto que a computação priorizou as pesquisas em gramática e *parser*.

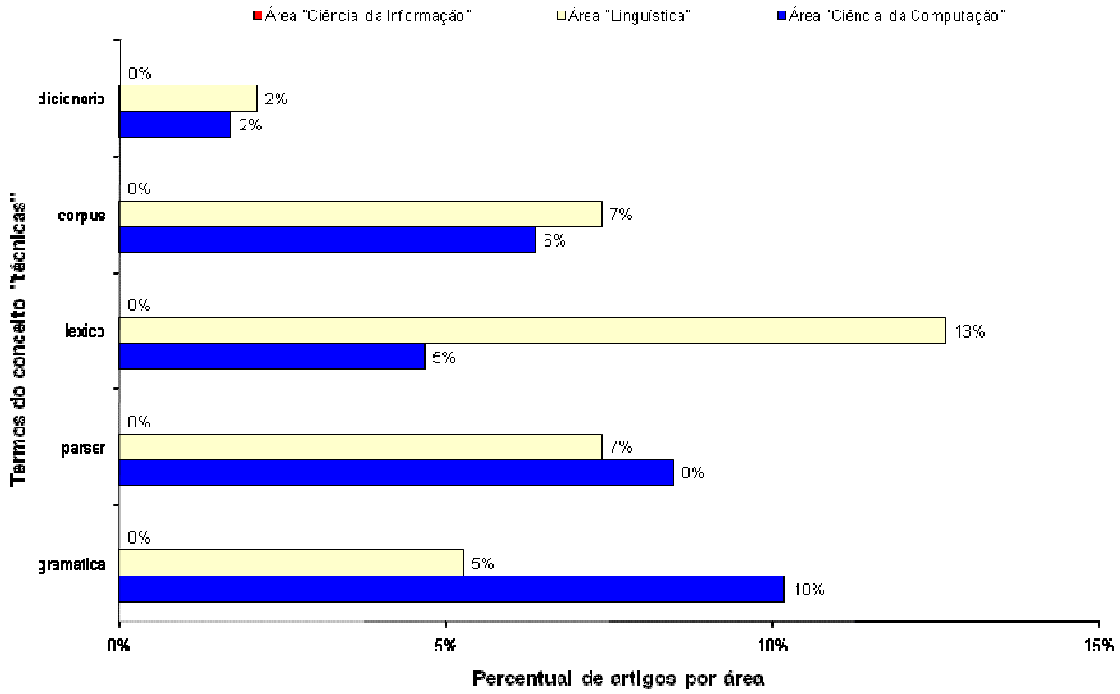

FIGURA 18 – Percentual de artigos de cada área com os principais termos dentre as técnicas

Finalmente, procurou-se identificar o que cada área tem priorizado, dentro dos conceitos definidos e utilizados para selecionar as publicações relevantes.

Na TAB. 5 é possível constatar que, dentre os conceitos computacionais, o segundo mais utilizado pela ciência da computação foi a inteligência artificial, o que sugere que a área tem encontrado espaço na IA para publicar seus trabalhos de processamento de linguagem natural.

TABELA 5
Principais termos dos conceitos computacionais em cada área

	Resultado de cada área do artigo		
Principais termos dos conceitos computacionais	**Ciência da Computação**	**Linguística**	**Ciência da Informação**
1.o	automático (24%)	processamento (36%)	automático (67%)
2.o	inteligência artificial (20%)	computacional (35%)	sistema (33%)
3.o	computacional (19%)	automático (17%)	tecnologia (11%)

Na TAB. 6, duas questões chamam a atenção: o fato da ciência da computação e linguística priorizarem os mesmos conceitos linguísticos, enquanto que a ciência da informação não fez uso de nenhum deles.

TABELA 6
Principais termos dos conceitos linguísticos em cada área

	Resultado de cada área do artigo		
Principais termos dos conceitos linguísticos	**Ciência da Computação**	**Linguística**	**Ciência da Informação**
1.o	português (28%)	português (35%)	-
2.o	linguagem natural (24%)	linguagem natural (33%)	-
3.o	palavra (10%)	palavra (16%)	-

A TAB. 7 mostra como a ciência da informação tem priorizado o estudo da indexação seguido da análise de conteúdo, não havendo uma terceira aplicação. Já a ciência da computação e a linguística apresentaram as mesmas temáticas.

TABELA 7
Principais termos dentre as aplicações em cada área

	Resultado de cada área do artigo		
Principais termos dentre as aplicações	**Ciência da Computação**	**Linguística**	**Ciência da Informação**
1.o	tradução (7%)	tradução (5%)	indexação (89%)
2.o	sumarização (7%)	sumarização (1%)	analise de conteúdo (11%)
3.o	recuperação (3%)	indexação (1%)	-

Na TAB. 8 é possível observar que dentre as técnicas presentes nos títulos dos artigos analisados, a ciência da informação apresentou apenas o thesauro, enquanto que a ciência da computação e a linguística apresentaram ‘parser’ e ‘corpus’ como sendo as técnicas mais pesquisadas, perdendo apenas para ‘gramática’, no caso da ciência da computação, e ‘léxico’ para a linguística.

TABELA 8
Principais termos dentre as técnicas em cada área

	Resultado de cada área do artigo		
Principais termos dentre as técnicas	**Ciência da Computação**	**Linguística**	**Ciência da Informação**
1.o	gramática (10%)	léxico (13%)	thesauro (11%)
2.o	parser (8%)	parser (7%)	-
3.o	corpus (6%)	corpus (7%)	-

Na próxima seção serão apresentados os resultados obtidos a partir da análise de conteúdo realizada em uma amostra estratificada de 68 artigos.

4.2. Análise Vertical ou Profunda das publicações analisadas

Os resultados apresentados nesta seção foram obtidos a partir da análise de conteúdo realizada nas 68 publicações sorteadas dentre as 621 realmente atinentes a área de PLN. Inicialmente, na seção 4.2.1 será apresentada a seleção de enunciados elaborada, seguida da sistematização semântica dessa análise, apresentada na seção 4.2.2.

4.2.1. Análise de Conteúdo das publicações analisadas

As discussões aqui apresentadas não caracterizam uma revisão de literatura, e sim uma visão multidimensionada, determinada pelas categorias de análise usadas, e linear, ou seja, em ordem cronológica, das pesquisas na área de processamento de linguagem natural. Os 68 artigos analisados foram publicados por pesquisadores nacionais e compreende o período de 1986 a 2009. Na TAB. 9 são apresentadas todas publicações analisadas, juntamente com os seguintes atributos identificadores: ano de publicação, autores, título e evento ou periódico de publicação. Em negrito, são apresentados os termos do instrumento de seleção utilizados para recuperar a publicação.

TABELA 9
Publicações submetidas à análise de conteúdo: 1986-2009

Ano de Publicação	Autores	Título	Evento/Periódico
1986	SEMEGHINI-SIQUEIRA, Idmea. ; COSTA, A. ; COHN, P. G. .	Uma **Gramática Conexionista**: Propriedades e Aplicações.	III Simpósio Brasileiro de **Inteligência Artificial**
1987	ZIVIANI, N. ; ALBUQUERQUE, L. C. A.	Um novo método eficiente para **recuperação em textos**	VII Congresso da Sociedade Brasileira de **Computação**
1988	RIPOLL, L. M. B. ; MENDES, S. B. T.	Um modelo **conexionista** para tratamento da ambiguidade **verbal** de um sub-conjunto do **português**.	XV SEMISH - Seminário Integrado de Software e Hardware
1989	FUSARO, P. S. ; ZIVIANI, N.	Uma linguagem de consulta para um **sistema** de **recuperação** de informação em **texto** completo.	IX Congresso da Sociedade Brasileira de **Computação**

1990	STRUBE DE LIMA, V. L.	Tratamento **automatizado** da **língua natural**: rumo a correção automática?.	Letras de hoje
1991	LEFFA, V. J.	O uso do **dicionário** eletrônico na compreensão do texto em **língua estrangeira**	XI Congresso da Sociedade Brasileira de **Computação**,
1992	ROCHA, A. F. ; GUILHERME, I. R. ; THEOTO, M. ; MIYADAHIRA, A. M. K. ; KOIZUMI, M. S.	A **neural net** for extracting knowledge from **natural language** data bases.	IEEE Transactions on **Neural Networks**
1993	ROCHA, R. A. ; ROCHA, B. H. S. C. ; HUFF, S. M.	**Automated Translation** between Medical **Vocabularies** using a Frame-based Interlingua	Seventeenth Symposium on **Computer** Applications in Medical Care
1994	ROBIN, J. P. L.	**Automatic** Generation and Revision of Natural Language Report **Summaries** Providing Historical Background	XI Simpósio Brasileiro de **Inteligência Artificial**
1995	JULIA, R. M. S. ; SEABRA, J. R. ; SEMEGHINI-SIQUEIRA, I.	An Intelligent **Parser** that **Automatically** Generates Semantic Rules during Syntatic and Semantic Analysis	IEEE International Conference on **Systems**, Man and Cybernetics
1996	BARROS, F. A.	Semi-automatic Anaphora Resolution in Portable **Natural Language** Interfaces	XIII Brazilian Symposium on **Artificial Intelligence**
1997	ROSA, J. L. G.	Thematic **Connectionist** Approach to **Portuguese Language Processing**	Iasted International Conference on **Artificial Intelligence** and Soft Computing
1998	OLIVEIRA, ITAMAR LEITE DE ; WAZLAWICK, R. S.	Modular **Connectionist Parser** for Resolution of **Pronominal** Anaphoric References in Multiple Sentences	International Joint Conference on **Neural Network** - IEEE World Conference on Computational Intelligence
1999	CARVALHO, A. M. B. R. ; STRUBE DE LIMA, V. L.	**Processamento** de **Língua Natural**: duas experiências com sistemas multi-agentes.	IX Intercâmbio de Pesquisas em **Linguística** Aplicada – INPLA
1999	KINOSHITA, J.	An Example based Machine **Translation System** working on trigrams.	32nd Annual Meeting of Societas **Linguistica** Eropae
1999	BARCIA, R. M. ; HOESCHL, HUGO ; MATTOS, EDUARDO DA SILVA ; BUENO, TANIA CRISTINA D'AGOSTINI ; GRESSE VON WANGENHEIM, C.	Uso da Teoria Jurídica para **Recuperação** em amplas bases de **textos** jurídicos	Encontro Nacional de **Inteligência Artificial** - XIX Congresso Nacional da Sociedade Brasileira de **Computação**
1999	BERBER SARDINHA, TONY	Estudo baseado em **Corpus** da Padronização **Lexical** no **Português Brasileiro**: colocações e perfis semânticos	PROPOR'99 - IV Encontro para o **Processamento Computacional** da **Língua Portuguesa** Escrita e Falada

1999	VILLAVICENCIO, ALINE	Representing a **System** of **Lexical** Types using Default Unification	Meeting of the European Chapter of the Association for **Computational Linguistics**
2000	JOSE NETO, J. ; MENEZES, C. E. D.	Um Método para a Construção de Etiquetadores Morfológicos Aplicado a **Língua Portuguesa**, baseado em Autômatos Adaptativos	PROPOR 2000 – V Encontro para o **Processamento Computacional** da **Língua Portuguesa**
2000	BERBER SARDINHA, TONY .	Prosódia Semântica na **Tradução** do **Português** e Inglês: um estudo baseado em **corpus**	PROPOR 2000 – V Encontro para o **Processamento Computacional** da **Língua Portuguesa** Falada e Escrita
2000	PADILHA, E. G. ; VICCARI, R. M.	Morfologia da **Língua Portuguesa** com Máquinas de Estados Finitos	PROPOR 2000 – V Encontro para o **Processamento Computacional** da **Língua Portuguesa** Falada e Escrita
2000	LAROCCA NETO, J. ; SANTOS, A. D. ; KAESTNER, C. A. A. ; FREITAS, A. A. ; NIEVOLA, J. C.	A Trainable **Algorithm** for **Summarizing** News Stories	PKDD 2000 - Workshop on Machine Learning and Textual Information Access
2000	DIAS-DA-SILVA, Bento Carlos ; MORAES, Helio Roberto de; OLIVEIRA, Mirna Fernanda de; HASEGAWA, Ricardo; AMORIM, Daniela Angelucci de; PACHOALINO, Christie	Construção de um **Thesaurus** Eletrônico para o **Português** do Brasil	PROPOR 2000 – V Encontro para o **Processamento Computacional** da **Língua Portuguesa** Falada e Escrita
2001	ROSSI, D. ; PINHEIRO, Clarissa ; FEIER, Nara Bressane ; VIEIRA, Renata	Resolução **Automática** de **Correferência** em textos da **Língua Portuguesa**	Revista Eletrônica de Iniciação Científica – REIC
2001	GAMALLO, Pablo ; AGUSTINI, Alexandre ; LOPES, Jose Gabriel Pereira	Selection Restrictions Acquisition from **Corpora**.	10th Portuguese Conference on **Artificial Intelligence** - EPIA. Lecture Notes in Artificial Inteligence – LNAI
2001	GONZALEZ, M. A. I. ; STRUBE DE LIMA, V. L.	**Recuperação** de Informação e Expansão **Automática** de consulta com **thesaurus**: uma avaliação	XXVII Conferência Latinoamericana de **Informática** – CLEI 2001
2001	ORENGO, VIVIANE MOREIRA ; HUYCK, CHRISTIAN	A **stemming algorithm** for the **portuguese** language	8th international symposium on string processing and information retrieval
2001	SOUZA, C. F. R. ; PEREIRA, M. B. ; NUNES, M. G. V.	**Algoritmos** de **sumarização** extrativa de textos em **português**	Workshop da Sociedade Brasileira da Computação
2002	JOSE NETO, J. ; MORAES, M.	Formalismo Adaptativo Aplicado ao Reconhecimento de **Linguagem Natural**	Conferência Iberoamericana em **Sistemas**, Cibernética e **Informática** - CISCI 2002

2002	BIDARRA, J.	Notas para a Especificação de um **léxico computacional**, baseadas em dados de Parafasia Semântica	Congresso Brasileiro de **Computação**, II Workshop de Informática na Saúde
2002	PARDO, Thiago Alexandre Salgueiro ; RINO, L. H. M.	DMSumm: Um Gerador **Automático** de **sumários**	I Workshop de Teses e Dissertações em **Inteligência Artificial**
2002	SCHULZ, S. ; NOHAMA, P. ; BORSATO, E. P. ; MATIAS, L. J. D.	**Indexação** e **Recuperação Automática** de **textos** médicos	VIII Congresso Brasileiro de **Informática** em Saúde – CBIS 2002
2002	BONFANTE, A. G. ; NUNES, M. G. V.	**Parsing** Probabilístico para o **Português** do Brasil.	I Workshop de Teses e Dissertações em **Inteligência Artificial**
2003	ZAVAGLIA, C.	Homonímia no **Português**: tratamento semântico segundo a estrutura Qualia de Pustejovsky com vistas a **implementações computacionais**	Revista Alfa
2003	MARTINS, Claudia A ; MONARD, M. C. ; MATSUBARA, E. T.	Reducing the Dimensionality of Bag-of-**words Text** Representation used by Learning **Algorithms**	**Artificial Intelligence** and Applications
2003	PARDO, THIAGO A. S. ; RINO, LUCIA H MACHADO ; NUNES, M. G. V.	NeuralSumm: Uma Abordagem **Conexionista** para a **Sumarização Automática** de Textos	Encontro Nacional de **Inteligência Artificial** – ENIA'2003
2003	GASPERIN, Caroline Varaschin ; STRUBE DE LIMA, V. L.	Evaluating **Automatically** Computed **Word** Similarity	**Computational Processing** of the **Portuguese** Language – PROPOR 2003
2003	OLIVEIRA, C. M. G. M. ; GARRAO, M. U. ; AMARAL, L. A. M.	Complex **Prepositions** Prep+N+Prep as Negative Patterns in Automatic Term Extraction from Texts	7th Conference on **Computational** Lexicography and Text Research
2004	ALVES, Isa Mara da Rosa ; CHISHMAN, R. L. O.	Ambiguidade e a **Tradução Automática**: uma análise do desempenho	III Colóquio Anual de Lusofonia
2004	SPECIA, L. ; NUNES, Maria das Graças Volpe	Um modelo para a Desambiguação **lexical** de sentido na **Tradução Automática**	II Workshop de teses e dissertações em **Inteligência Artificial**
2004	RINO, L. H. M. ; PARDO, Thiago Alexandre Salgueiro ; SILLA JR, Carlos Nascimento ; KAESTNER, Celso Antonio Alves ; POMBO, M.	A Comparison of **Automatic Summarization Systems** for Brazilian **Portuguese** Texts	XVII Brazilian Symposium on **Artificial Intelligence** – SBIA 2004
2004	ALUISIO, S. M. ; PINHEIRO, Gisele Montilha ; MANFRIN, Aline P M ; OLIVEIRA, Leandro H M de ; GENOVES JR, Luiz C ; TAGNIN, Stella E O	The Lacio-Web: **Corpora** and Tools to Advance Brazilian **Portuguese** Language Investigations and **Computational Linguistic** Tools	4th International Conference on Language Resources and Evaluation – LREC 2004

2004	MATSUBARA, E. T. ; MONARD, M. C. ; BATISTA, G. E. A. P. A. .	Aprendizado Semi-Supervisionado Multi-Visão para a **Classificação** de Bases de **Texto**	Workshop in **Artificial Intelligence** - Jornadas Chilenas de Computacion
2005	PARDO, Thiago A S ; MARCU, Daniel ; NUNES, M. G. V.	Um Modelo Estatístico Gerativo para o Aprendizado Não Supervisionado da Estrutura Argumental dos **Verbos**	III Workshop em **Tecnologia** da Informação e da **Linguagem Humana** – TIL 2005
2005	CASELI, H. M. ; NUNES, M. G. V. ; FORCADA, M. L.	LIHLA: A **Lexical** Aligner Based on Language-Independent Heuristics	V Encontro Nacional de **Inteligência Artificial** – ENIA 2005
2005	SPECIA, L. ; NUNES, Maria das Gracas Volpe ; STEVENSON, Mark	Mining rules for **Word** Sense Disambiguation	III Workshop em **Tecnologia** da Informação e da **Linguagem Humana** – TIL 2005
2005	SILVA, Cassiana Fagundes da ; VIEIRA, Renata ; OSORIO, Fernando Santos	Evaluating the Use of **Linguistic** Information in the **Preprocessing** Phase of **Text** Mining	Iberoamerican Journal of **Artificial Intelligence**.
2005	PILTCHER, Gustavo ; BORGES, Thyago ; LOH, S. ; LITCHNOW, Daniel ; SIMOES, Gabriel	Correção de **Palavras** em Chats: Avaliação de bases para **Dicionários** de Referência	III Workshop em **Tecnologia** da Informação e da **Linguagem Humana** – TIL 2005
2006	RINO, L. H. M. ; SENO, Eloize Rossi Marques	A importância do tratamento co-referencial para a **sumarização automática** de textos	Estudos Linguísticos
2006	CASELI, H. M. ; NUNES, M. G. V.	**Automatic** transfer rule induction from parallel **corpora**	3rd Workshop on Msc dissertations and PhD thesis in **Artificial Intelligence** – WTDIA'2006
2006	BALAGE FILHO, P. P ; UZEDA, V. R. ; PARDO, Thiago Alexandre Salgueiro ; NUNES, Maria das Graças Volpe	Experiments on applying a text **summarization system** for question answering	Cross **Language** Evaluation Forum – CLEF 2006
2006	ENEMBRECK, F ; SCALABRIN, E. E. ; TACLA, CESAR ; AVILA, BRAULIO COELHO	**Automatic** identification of teams based on **textual information retrieval**	Computer supported collaborative in design – cscwd 2001
2006	LEITE, DANIEL SARAIVA ; RINO, L. H. M.	Selecting a feature set to **summarize texts** in brazilian **portuguese**	Advances in artificial intelligence (iberamia/sbia)
2007	MORAES, S. M. W. ; STRUBE DE LIMA, V. L.	Um estudo sobre **Categorização** Hierárquica de uma grande coleção de **textos** em **Língua Portuguesa**	Workshop em **Tecnologia** da Informação e da **Linguagem Humana** – TIL 2007
2007	KINOSHITA, J. ; SALVADOR, L. N. ; MENEZES, C. E. D. ; SILVA, W. D. C. M.	COGROO - An Openoffice **Grammar** Checker	International Conference on **Intelligent Systems** Design and Applications – ISDA 2007

2007	SPECIA, L. ; STEVENSON, Mark ; NUNES, Maria das Gracas Volpe	Learning Expressive Models for **Word** Sense Disambiguation	Annual Meeting of the Association for **Computational Linguistics**
2007	SILVA, C. F. DA ; VIEIRA, RENATA	Categorização de textos da **língua portuguesa** com árvores de decisão, SVM e informações **linguísticas**	Workshop em **tecnologia** da informacao e da **linguagem humana** – TIL 2007
2007	MILIDIU, R. L. ; DUARTE, JULIO CESAR ; ROBERTO CAVALCANTE	Machine learning **algorithms** for **portuguese** named entity recognition	Simpósio Brasileiro de **Inteligência Artificial**
2008	CASELI, H. M. ; PARDO, T. A. S. ; GOMES, F. T. ; NUNES, M. G. V.	VisualLIHLA: the visual online tool for **lexical** alignment	VI Workshop em **Tecnologia** da Informação e da **Linguagem Humana** – TIL 2008
2008	AZIZ, W. F. ; PARDO, T. A. S. ; PARABONI, I.	An Experiment in Spanish-**Portuguese** Statistical Machine **Translation**	19th Brazilian Symposium on **Artificial Intelligence** – SBIA 2008
2008	MORAIS, E. A. M. ; AMBROSIO, A. P. .	Automatic Domain **Classification** of Jurisprudence **Documents**	Euroamerican Conference on Telematics and Information **Systems** - EATIS 2008
2008	CAMINADA, Nuno ; QUENTAL, V. S. D. B. ; GARRAO, Milena Uzeda	**Linguistic** Tools – Uma Plataforma Expansível de Funções de Consulta a **Corpus**	VI Workshop em **Tecnologias** da Informação e **Linguagem Humana** – TIL 2008
2008	SENO, Eloize R M ; NUNES, M. G. V.	Some Experiments on **Clustering** Similar Sentences of **Texts** in **Portuguese**	International Conference on **Computational Processing** of the **Portuguese Language** – PROPOR 2008
2009	AZIZ, W. F. ; PARDO, T. A. S. ; PARABONI, I.	Fine-tuning in **Portuguese**-English Statistical Machine **Translation**	7th Brazilian Symposium in Information and Human **Language Technology** - STIL 2009
2009	SENO, Eloize R M ; NUNES, M. G. V.	Fusão Automática de **Sentenças** Similares em **Português**	Simpósio Brasileiro em **Tecnologia** da Informação e da **Linguagem Humana** – TIL 2009
2009	SALLES, T. ; ROCHA, L. C. ; MOURAO, F. H. J. ; CUNHA, L. ; PAPPA, G. L. ; GONCALVES, Marcos Andre ; MEIRA JUNIOR, Wagner	**Classificação Automática** de **Documentos** Robusta Temporalmente	Simpósio Brasileiro de Banco de Dados
2009	BRAGA, I. A. ; MONARD, M. C. ; MATSUBARA, E. T.	Combining Unigrams and bigrams in Semi-supervised **Text Classification**	**Portuguese** Conference on **Artificial Intelligence**

2009	VILLAVICENCIO, A. ; CASELI, H. M. ; MACHADO, A.	Identification of **multiword** expressions in technical domains: investigating statistical and alignment-based approaches	Brazilian symposium in information and **human language technology – TIL 2009**

Os artigos analisados serão apresentados, em ordem cronológica, nas seções a seguir de acordo com as dimensões definidas, ou seja, de acordo com as categorias de análise definidas anteriormente. As referências incluídas em notas de rodapé não foram lidas pela autora desta tese, e sim citadas pelos artigos submetidos à análise de conteúdo. Para delimitar cada enunciado apresentado, optou-se por destacar em negrito as citações bibliográficas dos 68 artigos focalizados.

Na seção 4.2.1.1, são apresentados os problemas abordados e os objetivos propostos nas publicações analisadas. Procurou-se, neste momento, identificar quais as **problemáticas** que foram discutidas pelos autores ao longo dos anos. Na seção 4.2.1.2, os artigos foram analisados dentro da dimensão "**metodologia adotada**", ou seja, procurou-se identificar os métodos e as técnicas utilizados durante a realização de cada trabalho. Na seção 4.2.1.3, cada artigo foi analisado com o intuito de identificar o **material empírico usado**, assim como algumas características discriminantes tais como o idioma utilizado. Finalmente, na seção 4.2.1.4, são apresentados os **resultados** alcançados e discutidos nos trabalhos focalizados, juntamente com as perspectivas de continuidade dos mesmos.

4.2.1.1. Problemática apresentada nos artigos analisados

Nesta seção pretende-se apresentar, em ordem cronológica, as publicações analisadas sob a ótica dos problemas abordados, incluindo a problemática central de cada artigo, assim como os objetivos propostos. Na seção 4.2.2 é apresentado um mapa conceitual sintetizando todas as temáticas discutidas nos artigos focalizados. Os conceitos e relacionamentos apresentados em vermelho nesse mapa conceitual representam o núcleo semântico e, portanto, estão destacados em negrito nos parágrafos que se seguem.

O primeiro artigo selecionado foi **Semeghini-Siqueira, Costa e Cohn (1986)** publicado no Simpósio de Inteligência Artificial por autores da área da Linguística. Esta interdisciplinariedade é evidenciada pelos autores no início do artigo ao afirmarem que "a compreensão da linguagem natural, por computador, requer a cooperação interdisciplinar sobretudo de: linguistas (com informações sobre fonologia, sintaxe, semântica e pragmática); psicólogos (com dados sobre o processamento humano da informação: memória, atenção, percepção, etc.); filósofos (com sistemas de formalização do conhecimento); especialistas em um ramo do saber (para a montagem da base de conhecimento) e programadores (com domínio de uma linguagem de programação, como o Prolog)" (p. 113). Apesar dos autores passarem a proposta conexionista no título e citar a Inteligência Artificial logo no início, o que sugere a utilização de redes neurais, eles descrevem um sistema implementado em Prolog, com a finalidade de facilitar a **consulta a uma base de dados relacional usando linguagem natural**, interagindo o componente sintático e o componente semântico com incursões pelo componente pragmático.

O artigo de **Ziviani e Albuquerque (1987)** foi publicado num congresso da Sociedade Brasileira da Computação e teve como objetivo apresentar um novo método para identificação de termos indexadores através da **utilização de um índice** que reduz drasticamente a quantidade de dados a serem percorridos. Uma árvore Patrícia é construída sobre as assinaturas das palavras do texto, permitindo a detecção de termos em tempo proporcional ao logaritmo base dois do número de assinaturas obtidas do arquivo original. A assinatura de uma palavra é uma função que transforma palavras (cadeias de caracteres) em inteiros (p. 177).

Ripoll e Mendes (1988) apresentam a ambiguidade como problema central do tratamento das linguagens naturais, e propõem utilizar um modelo conexionista e uma gramática de casos para tratar a **ambiguidade léxica** de um subconjunto de **verbos** no português (p. 296). A proposta é escolher adequadamente o significado de uma palavra na frase. Segundo os autores, não se pretende discutir a questão do que é significado de uma palavra e as nuances que determinam vários significados para palavras ambíguas.

O artigo publicado em **Fusaro e Ziviani (1989)** apresenta a continuidade de um trabalho anterior (ZIVIANI; ALBUQUERQUE, 1987). Não apresenta como tema central o processamento de linguagem natural, e sim a **construção de um arquivo invertido** e a estrutura de dados usada. O objetivo principal deste artigo é

apresentar uma linguagem de consulta para sistemas de **recuperação de informação** em texto completo, comparável às linguagens mais modernas, baseando-se no sistema PatPlus (apresentado em Ziviani e Albuquerque, 1987).

O artigo de **Strube de Lima (1990)** apresenta uma revisão de literatura e visa prover ao leitor uma visão "panorâmica" no que se refere ao tema **correção ortográfica** automatizada, apresentando um resumo das técnicas e métodos empregados à época no tratamento da língua natural, abordando suas vantagens, suas deficiências e sua transposição para o português (p. 43).

O artigo de **Leffa (1991)** tem como objetivo principal comparar a **utilização do dicionário tradicional com o eletrônico**. A questão básica abordada neste trabalho, e que, segundo o autor, as investigações realizadas até o momento de sua publicação ainda não haviam sido respondidas, é, como um dicionário eletrônico, incorporando uma léxico-gramática e os recursos do computador, beneficiaria o leitor de uma língua estrangeira na **tradução** de textos autênticos. A hipótese principal desta investigação é que o dicionário eletrônico pode tornar o texto autêntico da língua estrangeira compreensível para o leitor de baixa proficiência nessa língua. O autor destaca que, em termos da quantidade de ajuda oferecida ao leitor, o pressuposto teórico foi de que "o dicionário não deveria oferecer nem de menos, deixando o texto incompreensível para o leitor, nem demais, abafando o texto a ponto de mudar a interação leitor/texto para leitor/dicionário" (p. 190). Em termos de qualidade, o autor complementa que "a ajuda deveria ser rápida (idealmente oferecida no momento em que o significado está sendo construído), discreta (nunca substituindo o texto lido ou colocando-se entre o leitor e o texto) e contextualizada (dando informação relacionada ao segmento do texto que está sendo lido)" (p. 190).

O artigo em **Rocha *et al.* (1992)** tem como objetivo apresentar um sistema com rede neural artificial evolutiva e hierárquica de três níveis, capaz de compreender o conteúdo de textos e **produzir listas de tópicos** a partir de registros de banco de dados. Os autores dedicam grande parte do artigo discutindo questões relacionadas aos atributos da rede neural construída: logo na introdução, os autores destacam que definir o número de camadas, assim como o número de neurônios por camada, pode ser uma tarefa difícil. Diante disso, os autores propõem um sistema composto de três diferentes redes: a primeira seria capaz de reconhecer as palavras; a segunda, responsável por reconhecer a associação entre estas palavras;

e finalmente, a terceira para apreender o principal conceito presente nos registros de banco de dados (p. 819).

Em **Rocha, Rocha e Huff (1993)**, observa-se que os autores tiveram como ponto de partida um problema e tentaram resolvê-lo, o que pode ser justificado pelo perfil dos autores: todos são da medicina. O problema apresentado é que a integração de sistemas clínicos ou médicos, segundo os autores, quase sempre requer uma etapa de **tradução**, onde vocabulários são comparados e os conceitos similares são combinados. Segundo os autores, o problema central que dificulta o desenvolvimento de qualquer sistema clínico é a ausência de métodos padronizados para representação de terminologia médica (p. 690). Assim, o principal objetivo do trabalho é traduzir termos expressos em diferentes **vocabulários médicos** usando um processo completamente automatizado.

Em **Robin (1994),** o autor apresenta o desenvolvimento de **sumarizadores** automáticos como sendo fundamental para administrar ou lidar com o volume de informações disponibilizadas *online*. Inicialmente, o autor identifica cinco aspectos a serem considerados na geração de sumários: a complexidade das sentenças; os conceitos flutuantes (*floating*); os fatos de cenário ***(historical background),*** que explicam algo ou que são relevantes; concisão (*conciseness*) e paráfrase (*paraphrasing*). Assim, o autor propõe a criação de um modelo que primeiro constrói um rascunho contendo somente os fatos essenciais do texto e depois vai incrementando-o com fatos de cenário (*historical background*) presentes em um limite de espaço. Segundo o autor, este modelo requer um novo tipo de conhecimento linguístico: as operações de revisão (*revision operations)*, especificando as várias maneiras nas quais um rascunho pode ser transformado de forma concisa, a fim de acomodar uma nova informação.

Julia, Seabra e Semeghini-Siqueira (1995) propõem um parser que realiza a **análise sintática e semântica** de afirmações sobre especificação de software expressas de maneira irrestrita em linguagem natural. O analisador proposto corresponde a uma estrutura (como definido por Piaget), que automaticamente gera regras semânticas durante a análise, orientada por um método **heurístico**. Segundo os autores, uma estrutura é um sistema de transformações caracterizadas por um grupo de regras. A parte sintática da gramática é expressa por meio de regras, tais como as regras de gramática proposta por **Chomsky**. O *parser* implementado é baseado em algoritmos de busca que tem

como objetivo encontrar um caminho da árvore até um nó folha que contenha uma categoria de significado. A categoria de cada palavra na sentença irá depender da ordem em que ela aparece na sentença.

O artigo de **Barros (1996)** descreve um mecanismo para **resolução de anáfora pronominal** sem a utilização de modelo do mundo (*world models*), para garantir a portabilidade e ainda oferecer uma **interface para consultas em banco de dados** em linguagem natural. Segundo a autora, o módulo de discurso (*discurse module*) incorporado não precisa ser customizado, garantindo assim a portabilidade do sistema, sendo esta a principal contribuição do seu trabalho.

Rosa (1997) propõe a construção de uma arquitetura conexionista para **mapear papéis temáticos** em regras **semânticas**. Os vetores de características são organizados com base nas relações temáticas entre o verbo e as outras palavras de uma frase. O principal objetivo do trabalho é fornecer um mecanismo que lida com as restrições do papel semântico sobre a atribuição do papel temático. O modelo tem de ser capaz de aprender com base na experiência com frases e suas representações temáticas, e tem de ser capaz de generalizar novas sentenças. O artigo teve como inspiração dois trabalhos da década anterior onde as palavras são representadas por um conjunto de características semânticas que possuem um significado associado. Assim, o objetivo do artigo é aplicar a ideia dessa representação para construir uma arquitetura capaz de analisar e aprender a atribuição correta dos relacionamentos temáticos das palavras nas sentenças. O autor destaca que o sistema não pretende resolver o problema de ambiguidade, mas contribui com ideias para torná-lo menos difícil, visto que informações semânticas são usadas para representar os significados.

Oliveira e Wazlawick (1998) discutem o problema da **ambiguidade** diante da **resolução de anáforas**. Segundo os autores, **o objeto ou a pessoa referenciada** é encontrado usando um modelo conexionista inspirado no modelo SPEC – *Subsymbolic Parser for Embedded Clauses* (proposto por R.P. Miikkulainen, em 1995). O pronome usado no trabalho foi o "ele" (*he*) e o "ela" (*she*). Segundo os autores, referência anafórica é um fenômeno linguístico que ocorre quando um pronome ou um sintagma nominal em uma frase está se referindo a alguém ou a um objeto já mencionado no texto. O problema então é saber quem é este pronome ou sintagma nominal, uma vez que podem haver vários objetos ou pessoas mencionadas até o momento no qual a referência é feita (p. 1.194).

Em **Carvalho e Strube de Lima (1999),** o objetivo do trabalho foi investigar o uso de **sistemas multi-agentes** para o processamento da língua natural. As autoras afirmam que existem no mínimo duas possibilidades diferentes de distribuição do conhecimento linguístico entre os agentes no campo do processamento da língua natural: distribuição léxico-estrutural: os agentes são associados às palavras da sentença, de acordo com a categoria **morfossintática** das mesmas e de acordo com uma série de princípios de associação; e distribuição linguístico-cognitiva: os agentes são associados a níveis de processamento linguístico (**morfológico, sintático, semântico**), ou a fenômenos linguísticos específicos (elipse, coordenação, anáfora, ambiguidade categorial).

Kinoshita (1999) propõe um sistema de **tradução** baseado em exemplos. Os **exemplos** foram **extraídos da Bíblia**, livro de Mateus, em grego, inglês e português, anotado de acordo com a anotação de Strong (*Strong's annotation*). Segundo o autor, a anotação de Strong provê uma informação importante que não foi usada no trabalho: todas as palavras com o mesmo radical (*stem*) recebem o mesmo código. O autor sugere que esta informação seja utilizada em trabalhos futuros. Segundo o autor, neste trabalho, os exemplos são organizados em palavras, bigramas e trigramas (*bigrams* e *trigrams*). Assim, o autor destaca que dada uma sentença, as n-gramas (com n entre 1 e 3) são traduzidas de acordo com os exemplos. A hipótese do autor é que usando bigramas e trigramas será possível identificar melhor o contexto e então obter uma tradução melhor.

Barcia *et al.* (1999) propõem a utilização da técnica de **Raciocínio baseado em Casos (RBC)** para **solução de problemas jurídicos**. Segundo os autores, quando um profissional do direito realiza uma pesquisa jurisprudencial, ele está buscando informações para reforçar o seu ponto de vista sobre a interpretação de uma norma jurídica e define argumentos persuasivos para fazer a analogia entre o seu problema atual e o anterior, já solucionado. Ainda segundo os autores, muitos destes textos jurídicos estão disponíveis em bancos de dados, inclusive acessíveis na Internet. No entanto, as buscas por informações jurídicas nesses sistemas requerem conhecimento jurídico e estão limitadas devido a problemas como a ambiguidade sintática e semântica, e também a incerteza existentes nos textos dos documentos. O interessante deste trabalho é o fato da equipe de autores ser intrinsecamente multidisciplinar envolvendo pesquisadores da área do Direito, da Computação e da Engenharia da Produção.

Berber Sardinha (1999) apresenta um trabalho teórico com relatos dos resultados de um estudo cujo foco é a **descrição de padrões lexicais** e colocações do português. O objetivo é iniciar o estudo destes aspectos da linguagem em uso na língua portuguesa. Os relatos apresentados no presente trabalho visam fornecer uma descrição dos perfis semânticos de várias palavras da língua portuguesa. Segundo o autor, o estudo da **colocação ou co-ocorrência** significativa **de itens lexicais**, verificada computacionalmente em um corpus eletrônico, já se firmou como uma prática metodológica fundamental na descrição lexical e gramatical do inglês. A pesquisa relatada neste trabalho tem como objetivo preencher uma lacuna no estudo da padronização lexical da língua portuguesa, através da busca de elementos lexicais co-ocorrentes em um corpus eletrônico de grandes proporções. Segundo o autor, a investigação da padronização lexical baseada em corpus, conforme proposta neste trabalho, pressupõe uma visão da linguagem como um sistema probabilístico. O autor destaca que esta visão da linguagem encontra seu contraponto na linguística Chomskyana, mas com algumas diferenças: "foco no desempenho linguístico, em vez de competência; foco na descrição linguística, em vez de universais linguísticos; foco numa visão mais empirista do que racionalista da pesquisa científica" (p. 5).

Villavicencio (1999) demonstrou como o uso de **unificação padrão na organização da informação lexical** pode fornecer descrição não redundante de tipos lexical. Segundo a autora, padrões foram usados na definição da morfologia, na especificação da semântica lexical, na análise de construções em aberto (*gapping constructions*) e elipses (*ellipsis*), dentre outros. Neste trabalho, utilizou-se padrões para estruturar o léxico, concentrando-se na descrição das informações de categorização verbal.

O trabalho de **Jose Neto e Menezes (2000)** propõe um método para a construção de um **etiquetador morfológico**, que possa ser usado **em várias línguas**. Apesar de testá-lo apenas para a língua portuguesa, o trabalho propõe que seja treinável com o uso de corpus e que possibilite uma boa precisão na anotação. Segundo os autores, um etiquetador morfológico tem como função associar, a cada palavra, uma etiqueta que corresponda a sua categoria morfológica. E complementam que a principal dificuldade está em lidar com a ambiguidade. Um etiquetador morfológico robusto deve levar em conta não apenas as informações lexicais da palavra a ser anotada, mas também informações a respeito do contexto

em que esta palavra se encontra (p. 53).

Berber Sardinha (2000) tem como objetivo focalizar o problema de **tradução** de padrões lexicais, mais especificamente a tradução de termos equivalentes do inglês para o português, segundo a ótica da manutenção ou da quebra da **prosódia semântica** (associação entre itens lexicais e conotação – positiva, negativa ou neutra). Segundo o autor, um tipo de padrão importante para a tradução é a prosódia semântica, ou a associação recorrente entre itens lexicais e um campo semântico, indicando uma certa conotação (negativa, positiva ou neutra).

Em **Padilha e Viccari (2000)** foram desenvolvidos **processadores para a morfologia** do português utilizando **máquinas de estados finitos**, particularmente transdutores. Segundo os autores, um transdutor é um autômato cujas transições de estado são marcadas por pares ou tuplas de símbolos, em vez de símbolos simples. Ainda segundo os autores, "enquanto um autômato representa uma linguagem regular, um transdutor representa uma relação regular entre duas linguagens, associando diretamente cada "palavra" de uma à outra" (p. 44).

Em **Larocca Neto *et al.* (2000)**, os autores apresentam a **sumarização** de texto como sendo o processo de reduzir o tamanho do texto, preservando o conteúdo informacional do mesmo. Segundo os autores, existem vários sistemas robustos de sumarização de textos que utilizam técnicas estatísticas e/ou técnicas baseadas em análise linguística superficial e independente de domínio. A grande maioria dos sistemas, disponíveis à época, realizava sumarização de extratos, que segundo os autores, é uma forma relativamente simples de se fazer sumarização, onde sentenças do documento original são selecionadas de acordo com algum critério pré-definido. Normalmente, isto é feito, organizando-se as sentenças originais, e selecionando as que apresentarem maiores avaliações (escore). No entanto, isto não garante que o sumário obtido terá uma narrativa coerente. Diante disso, este trabalho tem como objetivo propor um sistema treinável baseado em aprendizado de máquina para **sumarização de notícias**. Ainda segundo os autores, o objetivo do trabalho é obter uma estrutura argumentativa aproximada do texto, usando algumas heurísticas.

Dias-da-Silva *et al.* (2000) abordam inúmeras questões envolvidas no processo de **compilação de um Thesaurus Eletrônico** Básico para o Português do Brasil (TeP). Segundo os autores, um thesaurus eletrônico, acoplado a outras ferramentas computacionais de auxílio à expressão escrita, deve complementar

outras referências, em meio digital, tais como dicionários e gramáticas. "Essa ferramenta deverá oferecer ao usuário da língua portuguesa a oportunidade ímpar de escolher palavras sinônimas e antônimas que ele, por motivos de estilo, de precisão, de correção ou de aprendizagem, deseja substituir" (p. 2). São apresentados o **arcabouço teórico-metodológico** adotado, assim como os principais problemas enfrentados na elaboração de um **modelo de representação linguístico-computacional** adequado e eficiente. Além disso, os autores procuram caracterizar o termo thesaurus, visto que, segundo os autores, o mesmo tem sido empregado por diferentes especialistas para denotar objetos bastante diversos. E por fim, os autores apresentam uma solução para a implementação do modelo, incluindo o editor do thesaurus, projetado para auxiliar o linguista no processo de construção do thesaurus. Os autores finalizam enumerando os resultados alcançados até o momento da publicação deste trabalho, assim como as etapas seguintes.

Rossi *et al.* (2001) tem como objetivo identificar as sequências de expressões em um texto que se referem a uma mesma entidade. Mais especificamente, investiga-se a **correferência das descrições definidas**, que são os sintagmas nominais iniciados por artigo definido (a, o, as, os). Segundo os autores, um dos motivos para se trabalhar com descrições definidas é o fato de, normalmente, ocorrerem em grande quantidade nos textos da Língua Portuguesa. O presente trabalho faz parte de um projeto maior intitulado ANACORT (Anotação Automática de Correferência Textual) que tem como objetivo geral a construção e o tratamento computacional de um corpus linguístico visando à resolução da correferência em textos do português. Segundo os autores, uma cadeia de correferência nominal é uma sequência de substantivos (ou nomes) em um discurso que se referem a uma mesma entidade, e "podem melhorar a qualidade dos resultados em diversas aplicações de processamento de linguagem natural, como recuperação e extração de informações, geração automática de resumos, traduções automáticas, entre outros" (p. 1).

Gamallo, Agustini e Lopes (2001) tem como objetivo descrever um método baseado em corpus para a **extração de informação semântica**. Segundo os autores, o intuito é utilizar informações sintáticas para extrair as restrições de seleção e preferências semânticas ao invés de combinação de palavras. Em outras palavras, é apresentado um método não supervisionado "pobre de conhecimento"

(*knowledge-poor*) para adquirir restrições de seleção baseado em hipóteses de contexto e de co-especificação. Segundo os autores, métodos pobres de conhecimento necessitam apenas ter noção de informação linguística: co-ocorrência de palavras. O principal objetivo é calcular a frequência da co-ocorrência dentro de construções sintáticas, ou sequências de n-gramas, com o objetivo de extrair informações semânticas, tais como restrições de seleção e ontologias de palavras. Segundo os autores, o relacionamento sintático binário é constituído tanto pela palavra que impõe restrições linguísticas (o predicado) como pela palavra que deve preencher essas restrições (o seu argumento). Em uma relação sintática, cada palavra tem um papel fixo. O argumento é visto como a palavra que especifica ou modifica as restrições sintático-semânticas impostas pelo predicado, enquanto o útlimo é visto como a palavra especificada ou modificada pela primeira.

Gonzalez e Strube de Lima (2001) apresentam uma primeira avaliação dos resultados obtidos com a **expansão automática de consulta** em **recuperação de informação**. Foi utilizado um **thesaurus**, com estruturação semântica e operações gerativas, para gerar o campo lexical de cada termo da consulta e obter a expansão automaticamente. A seleção dos novos termos e o cálculo de seus pesos, na consulta expandida, depende da sobreposição dos campos lexicais e do nível de profundidade que se avança na busca de descritores dos termos considerados.

Souza, Pereira e Nunes (2001) tem como objetivo apresentar um ambiente para testes de estratégias de **sumarização** automática **extrativa [extratos]** de português, chamado SUMEX. Segundo os autores, a sumarização automática extrativa consiste da extração de sentenças relevantes do texto-fonte para a formação do sumário.

Orengo e Huyck (2001) apresentam o desenvolvimento de um algoritmo para realizar análise de radicais (*stemming*) para o português, ou seja, suprimir o sufixo (*suffix-stripping*) das palavras reduzindo-as à sua raiz (*stem*). Segundo os autores, essa técnica tem sido amplamente utilizada na fase de pré-processamento dos documentos para recuperação de informação, por reduzir a estrutura de indexação adotada. Diante disso, este trabalho tem como objetivo apresentar a implementação de um algoritmo simples, mas, segundo os autores, efetivo, para **remoção de sufixos** na língua portuguesa.

Jose Neto e Moraes (2002) tem como principal objetivo mostrar o potencial e a aplicabilidade de formalismos adaptativos - em particular, dos

autômatos adaptativos - para a resolução de alguns dos problemas tipicamente encontrados na representação e no processamento de linguagens naturais. Segundo os autores, dois importantes aspectos ligados à complexidade sintática das linguagens naturais que precisam ser tratados são: o não-determinismo e a **ambiguidade sintática**. Os autores definem não-determinismos como sendo quando duas ou mais construções sintáticas, que ocorrem em um determinado ponto das sentenças, apresentem prefixo comum, e ambiguidades como sendo fenômenos linguísticos em que uma sentença pode ter duas ou mais interpretações válidas na mesma linguagem. Segundo os autores, os reconhecedores de linguagens ambíguas, muitas vezes, podem lidar com as ambiguidades simplesmente buscando a aceitação de uma, talvez a mais usual, ou a mais facilmente identificada, ou a de tratamento mais simples, das interpretações possíveis para a sentença, sendo as demais interpretações ignoradas. No entanto, os autores destacam que, um dos problemas usualmente encontrados no processamento de linguagens naturais, corresponde à dificuldade de expressar, através de um formalismo legível e expressivo, as complexas nuanças estruturais sempre presentes nas linguagens naturais. Uma alternativa é efetuar uma redução inicial da complexidade da linguagem que se deseja definir, através da **elaboração de uma aproximação livre de contexto** da mesma. Os autores afirmam que esta técnica é "bastante conveniente na prática, uma vez que, para linguagens livres de contexto, estão disponíveis inúmeras técnicas simples e eficientes de reconhecimento e de análise" (p. 1). Esta restrição pode parecer, num primeiro momento, um pouco quanto sem propósito para um trabalho que propõe o "reconhecimento de linguagem natural". Os autores complementam ainda que "através da eliminação dos aspectos mais complexos da linguagem, tais como ambiguidades e dependências de contexto, pode-se obter uma boa aproximação da linguagem natural, que represente, de forma simples, mas com uma fidelidade aceitável, todos os seus aspectos sintáticos mais importantes" (p. 1).

Bidarra (2002) considera alguns aspectos básicos para a **construção de léxicos** para o PLN. Além disso, o autor considera a **afasia ou parafasia semântica** como indícios para compreender "como as palavras estariam, teoricamente, representadas no léxico mental" (p. 1). Segundo o autor, a afasia tem sido uma grande fonte de descobertas não só para a neurolinguística e medicina, mas também objeto de pesquisas para o desenvolvimento de modelos computacionais de

Processamento da Linguagem Natural (PLN).

> O comportamento linguístico de sujeitos afásicos tem dado mostras de que o sistema lexical humano - os processos cognitivos que subjazem tarefas tais como nomeação de objetos, enunciação de expressões linguísticas ou partes delas e compreensão das palavras – parece constituído por componentes de processamento relativamente independentes entre si, desde que podem ser seletivamente prejudicados em virtude de lesões cerebrais (p. 1).

Sendo assim, o principal objetivo do trabalho é, partindo-se do problema da afasia, trazer para o debate a pesquisa que o autor desenvolve desde 1997, quando ingressou no doutorado. Segundo o autor, na sua tese, ele propôs um modelo de descrição lexical para dar suporte a questões relacionadas com patologias da linguagem e consequentemente modelos computacionais para este fim. Dentre as questões de pesquisa que o autor pretende aborda estão: Um dano causado nas estruturas internas do léxico seria realmente capaz de provocar a perda da capacidade do sistema para a recuperação do conjunto de informações necessárias para a correta captura da palavra desejada? E se não confirmada a primeira alternativa, seria o caso de dizer que os mecanismos usados pelo sistema para a manipulação dessas informações estariam prejudicados?

Pardo e Rino (2002) apresentam a **sumarização** automática como sendo uma área promissora de pesquisa nos dias atuais, diante da crescente quantidade de informação disponível e do tempo cada vez mais reduzido que o leitor tem para apreender o máximo dessa informação. Sendo assim, esse trabalho explora a abordagem fundamental para a sumarização dirigida por **objetivos comunicativos**, propondo a implementação de um modelo discursivo desenvolvido na tese de doutorado de um dos autores (RINO, 1996[3]). Segundo os autores, as premissas deste trabalho são de que os sumários gerados automaticamente devem satisfazer o objetivo comunicativo e preservar a proposição central do texto-fonte. Segundo os autores, discussões preliminares foram realizadas em outro trabalho de mesma autoria (PARDO E RINO, 2002[4]). O objetivo comunicativo é o responsável por garantir a coerência dos sumários gerados automaticamente e selecionar as proposições do texto-fonte que se relacionarão à proposição central nos sumários, garantindo, portanto, sua preservação.

---

[3] Rino, L.H.M. (1996). Modelagem de Discurso para o Tratamento da Concisão e Preservação da Idéia Central na Geração de Textos. Tese de Doutorado. IFSC-USP. São Carlos – SP.

[4] Pardo, T.A.S. and Rino, L.H.M. (2002). DMSumm: Review and Assessment. In E. Ranchhod and N. J. Mamede (eds.), Advances in Natural Language Processing, pp. 263-273 (Lecture Notes in Artificial Intelligence 2389). Springer-Verlag, Germany.

Em **Schulz *et al.* (2002),** os autores, da área médica, afirmam que a quase totalidade da informação médica produzida é expressa por meio da linguagem natural, e que o volume de informações disponíveis está crescendo a ponto de dificultar a seleção e a leitura do que é, de fato, útil e de interesse. Os autores destacam que a terminologia médica exibe características próprias que dificultam o uso eficiente dos mecanismos de busca, e citam vários exemplos dentre variação ortográfica, derivação, sinonímia, dentre outros. Além disso, os autores complementam destacando a necessidade de lidar com documentos em línguas diferentes. Sendo assim, os autores apresentam dois projetos que encontram-se em desenvolvimento em cooperação entre o Departamento de Informática Médica da Universidade de Freiburg (Alemanha) e o Grupo de Tecnologia em Saúde do Programa de Pós-Graduação em Informática Aplicada (PPGIA) da Pontifícia Universidade Católica do Paraná (PUCPR). Como parte deste projeto, desenvolveu-se o MORPHOSAURUS, uma metodologia que tem como objetivo aperfeiçoar a busca em coleções multilíngues de documentos médicos, e que é apresentado neste artigo. Em outras palavras, a metodologia apresentada, segundo os autores, abandona os métodos tradicionais de recuperação e se baseia no **uso de morfemas médicos**, como unidades atômicas para indexação e recuperação de informação.

Bonfante e Nunes (2002) destacam a importância de se **recuperar a estrutura sintática das sentenças**. Segundo elas, muito esforço tem sido empregado na construção de **analisadores sintáticos**, mas que a dificuldade de se especificar uma gramática com poder de descrição abriu caminho para a pesquisa empírica. Assim, um conjunto de sentenças anotadas sintaticamente é usado, como dados de treinamento, num processo de aprendizado para realizar a anotação de uma sentença desconhecida. Dentre as abordagens empíricas, as autoras citam o aprendizado de máquina simbólico, conexionista e estatístico. Este trabalho apresenta parte da tese de doutorado em andamento da primeira autora, que visa investigar o comportamento de analisadores sintáticos, implementados seguindo cada uma das três abordagens descritas anteriormente. Mais especificamente, o presente trabalho descreve a experiência de implementação de um parser probabilístico, seguindo o modelo de Collins (1999[5]). Segundo as autoras, o modelo baseia-se na noção de núcleos lexicais, onde para cada regra observada no

[5] Collins, M. J. . Head-Driven Statistical Models for Natural Language Parsing. PhD thesis, University of Pennsylvania, 1999.

conjunto de treinamento, as palavras que não são núcleo são chamadas de modificadores, exercendo influência sobre ele.

Zavaglia (2003) inicia seu artigo destacando que o léxico tem sido constantemente relacionado a memória humana, visto que as entradas lexicais em um dicionário são como registros de memória. Além disso, a autora afirma que neste sentido, o computador está fadado a incompletudes, visto que a forma como os dados são armazenados na memória ainda é um mistério (segundo a autora). Como possibilidade de representação, a autora destaca a utilização de redes semânticas, organizadas por relações semânticas. Segundo a autora, dos fenômenos linguísticos que fazem parte da língua natural, a **homonímia e a polissemia** causam o fenômeno da **ambiguidade**. Ainda segundo a autora, no âmbito do léxico, bem como dos fenômenos linguísticos geradores de ambiguidade, os pesquisadores em PLN certamente encontrarão vários obstáculos e empecilhos (p. 80). A autora destaca ainda que, o problema da homonímia gramatical é resolvido facilmente por sistemas computacionais, mas o mesmo não acontece com outros problemas da ambiguidades, tais como homonímia semântica e a polissemia. Segundo a autora, isso se deve ao fato da máquina não ser capaz de relacionar semanticamente itens lexicais em meio a construções sintáticas ou inseridos no contexto. Assim, este trabalho tem como objetivo apresentar uma proposta para o tratamento de itens lexicais homônimos da língua portuguesa do Brasil, por meio da construção de uma base de dados conceitual (base de conhecimento lexical – BCL). Pressupõem-se que tal base irá suprir as necessidades de um analisador sintático, e que a homonímia poderá ser tratada, uma vez que será fornecido a máquina, subsídios linguísticos tais como relações semânticas de itens lexicais em redes de significação.

Martins, Monard e Matsubara (2003) iniciam o artigo discutindo sobre a tarefa de classificar textos em linguagem natural. Segundo os autores, métodos manuais são caros e algumas vezes impraticáveis, enquanto que a maioria dos métodos automáticos modernos utiliza técnicas de aprendizado de máquina (*machine learning*) para classificá-los a partir de exemplos. No entanto, os autores destacam que é necessário transformar o texto em um **formato apropriado**[6] para os algoritmos de aprendizado, o que inclui atribuir pesos aos termos assim como

[6] No mapa conceitual da seção 4.2.2, esse artigo foi alocado dentro do conceito de 'Pré-processamento dos documentos', visto que, normalmente é nesta etapa que os documentos são transformados em um "formato apropriado".

reduzir a dimensão adotada. Segundo os autores, a representação tem forte influência na eficiência do algoritmo de aprendizado. Diante disso, o presente artigo tem como objetivo descrever uma maneira de **reduzir a dimensão** da tabela de atributos-valor, usada para representar uma coleção de documentos. Os autores apresentam a ferramenta PreText, desenvolvida com o objetivo de realizar automaticamente a tarefa de pré-processamento de uma coleção de documentos, e inclui a funcionalidade de reduzir a dimensionalidade do conjunto de dados usando a lei de Zipf e os limiares de Luhn.

Pardo, Rino e Nunes (2003) destacam que a **sumarização** automática de textos é o processo de se produzir uma versão mais curta de um texto-fonte que pode ser um extrato (*extract*) ou um sumário (*abstract)*. Segundo os autores, o extrato corresponde à justaposição de sentenças do texto-fonte consideradas importantes, enquanto que o sumário "altera a estrutura e/ou o conteúdo das sentenças originais, fundindo-as e/ou reescrevendo-as, para generalizar ou especificar as informações" (p. 1). No presente artigo é apresentado o sumarizador NeuralSumm (*NEURAL network for SUMMarization*), que produz extratos utilizando uma técnica de aprendizado de máquina – uma rede neural SOM (*self-organizing map*) (KOHONEN, 1982[7]), para identificar as sentenças importantes de um texto-fonte que irão compor seu extrato. A classificação das sentenças em graus de importância é feita pela rede neural com base em **características (*features*) extraídas** durante o processo de sumarização.

Gasperin e Strube de Lima (2003) apresentam a continuação do trabalho de mestrado da primeira autora (GASPERIN, 2001[8]), onde uma lista de palavras semanticamente relacionadas é gerada automaticamente usando técnica pobre de conhecimento baseada em sintaxe (*syntax-based knowledge-poor technique*). Segundo as autoras, por ser difícil avaliar a qualidade dessa lista de uma maneira sistemática, optou-se por aplicá-la a uma tarefa visível ao usuário e então avaliar tal tarefa. Sendo assim, o presente trabalho tem como objetivo avaliar a utilização de uma **lista de palavras relacionadas semanticamente** como fonte de conhecimento semântico na tarefa de **expansão de consulta**.

Oliveira, Garrao e Amaral (2003) destacam que, um tipo de expressão que, no português, frequentemente contém nomes são as preposições compostas

[7] Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. Biological Cybernetics, Vol. 43, pp. 59-69.
[8] Gasperin, C. V. . Extração Automática de relações semânticas a partir de relações sintáticas. Master's thesis, PPGCC – PUCRS, Porto Alegre (2001).

ou locuções preposicionais (*complex prepositions*). Ainda segundo os autores, é importante investigar critérios que permita a correta identificação dessas expressões, para evitar que as mesmas sejam detectadas como sintagmas nominais. Em outras palavras, os autores apresentam as estruturas preposicionais como sendo padrões negativos para serem usados na extração de sintagmas nominais de textos. Diante disso, o presente trabalho tem como objetivo formular um critério sistemático para **reconhecimento das estruturas [locuções] preposicionais**, ou seja, um critério que possa ser formalizado e implementado em um sistema de computador.

Alves e Chishman (2004) afirmam que, um tradutor automático pode desempenhar melhor sua função se forem utilizados métodos adequados para processar ambiguidades, anáforas e ideias implícitas da língua natural. As autoras complementam que a **ambiguidade** é um desafio, tanto para a linguística como para a computação, sendo que “sob a ótica da Linguística teórica, (…) esse fenômeno é uma riqueza da língua”, já para a área da tradução automática “é um problema a ser superado” (p. 97). Diante disso, o presente trabalho, tem como objetivo principal, mostrar como **os tradutores automáticos** tratam o complexo fenômeno linguístico da ambiguidade. Além disso, as autoras propõem uma reorganização das nomenclaturas usadas para o tratamento teórico desse fenômeno, uma vez que, segundo as autoras, foram encontradas na literatura, definições imprecisas e sobrepostas.

Specia e Nunes (2004) afirmam que apesar da tradução automática ser uma das áreas mais antigas do PLN, ela ainda apresenta muitos problemas, e que um dos principais problemas é a **ambiguidade lexical**. Segundo as autoras, esse problema se mostra ainda mais complexo de ser tratado quando são identificadas apenas variações de significado (de sentido) nas opções de tradução, ou seja, todas as opções são da mesma categoria gramatical (chamada de ambiguidade lexical de sentido). Diante disso, o presente trabalho tem como objetivo apresentar discussões preliminares de um projeto que encontra-se em especificação que propõe a construção de um modelo híbrido linguístico-computacional de **desambiguação lexical de sentido** (*word sense disambiguation*), ou seja, baseado em conhecimento linguístico (dicionários) e em algoritmos de aprendizado de máquina (corpus de exemplos). Mais especificamente, ele contempla a ambiguidade de um conjunto de **verbos**, visto que são altamente ambíguos e porque da sua desambiguação pode depender a desambiguação de outras palavras da sentença, principalmente dos

seus argumentos.

Rino *et al.* (2004) apresentam o problema da **sumarização** automática, mais especificamente os métodos extrativos baseados em técnicas estatísticas ou empíricas, nos quais trechos do texto original são justapostos para compor o extrato produzido. O presente trabalho tem como objetivo **comparar cinco métodos** de sumarização automática encontrados na literatura.

Aluisio *et al.* (2004) apresentam alguns consórcios americanos e ingleses com o intuito de desenvolver pesquisas acadêmicas na área de PLN, e destacam que iniciativas como estas, que produza recursos para a língua portuguesa no Brasil, ainda são desconhecidas. Os autores tentam justificar esta carência argumentando que talvez isso se deva ao fato do português não ser o idioma difundido mundialmente nas pesquisas e negócios. Por outro lado, eles argumentam que a língua portuguesa é falada por cerca de 200 milhões de pessoas no mundo todo, e que portanto merece destaque (sendo a sexta língua mais falada). Neste sentido, o presente trabalho tem como objetivo discutir os **requisitos necessários** para se construir um grande **repositório de recursos e ferramentas**, e apresentar o corpora **Lácio-Web**, projeto em desenvolvimento desde 2002, na universidade de São Paulo (NILC, IME e FFLCH). O Lácio-Web foi projetado tanto para pesquisadores linguísticos teóricos como práticos, e para o desenvolvimento tanto de ferramentas linguísticas computacionais como de aplicações, tais como etiquetadores (*tagger*), analisadores (*parsers*), corretores gramaticais (*grammar checkers*), métodos de recuperação de informação e sumarização automática.

Matsubara, Monard e Batista (2004) destacam que o **aprendizado de máquina semi-supervisionado** é uma área de pesquisa, segundo eles, relativamente recente, relacionada com algoritmos que aprendem utilizando uma combinação das facilidades oferecidas pelo aprendizado supervisionado – no qual é fornecido um conjunto de exemplos de treinamento rotulado com a **classe associada** a cada exemplo – e das facilidades oferecidas pelo aprendizado não-supervisionado – no qual a classe de cada exemplo não é conhecida. Os autores complementam apresentando o algoritmo co-training, proposto originalmente por Blum e Mitchell (1998[9]) e, implementado e disponibilizado em Matsubara e Monard

[9] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In Proc. 11th Annu. Conf. on Comput. Learning Theory, pages 92–100. ACM Press, New York, NY, 1998.

(2004[10]), como sendo um algoritmo de aprendizado semi-supervisionado que utiliza um método interessante (de **múltiplas visões**) para rotular exemplos quando o número de exemplos originalmente rotulados é pequeno. Diante disso, os objetivos deste trabalho são propor uma maneira de obter duas ou mais descrições dos dados, por meio de n-gram, em qualquer base de texto, para serem utilizadas em algoritmos multi-visão; e avaliar experimentalmente o comportamento do algorimo *co-training* com essa nova proposta.

Pardo, Marcu e Nunes (2005) destacam que muitos esforços têm sido feitos para a criação de repositórios semanticamente anotados, e que a anotação dos argumentos dos verbos representa uma parcela significativa destes esforços. Em outras palavras, as **anotações semânticas** focadas neste artigo são as **estruturas argumentais dos verbos**, que indicam quantos e quais são os possíveis argumentos que os verbos requerem. Os autores destacam que o ideal seria que todas as possibilidades de estruturas argumentais fossem incluídas na especificação semântica dos verbos. Sendo assim, neste artigo é apresentada uma abordagem não supervisionada, completamente automática, para o aprendizado das estruturas argumentais, utilizando-se um modelo estatístico gerativo baseado no modelo *noisy-channel* (SHANNON, 1948[11]) e treinado por meio do algoritmo *Expectation-Maximization* (DEMPSTER *et al*., 1977[12]).

Caseli, Nunes e Forcada (2005) apresentam **alinhamento de sentenças** (*multiwords*) **e palavras** como um importante papel em várias aplicações em processamento de linguagem natural tais como **tradução automática** baseada em exemplos e métodos estatísticos, desambiguação lexical de sentido (*word sense desambiguation*), dentre outros. Segundo os autores, alinhamento de dois ou mais textos significa encontrar correspondência (traduções equivalentes) entre segmentos do texto fonte (parágrafos, sentenças, palavras, etc) e segmentos de suas traduções no texto alvo. O enfoque deste artigo é o alinhamento lexical, ou seja, o alinhamento entre palavras ou sentenças em português do Brasil, espanhol e inglês. Em outras palavras, o método apresentado, *Language-Independent Heuristics Lexical Aligner (LIHLA)* tem como ponto de partida o uso de alinhamentos **estatísticos**, e pela

[10] Edson Takashi Matsubara and Maria Carolina Monard. Projeto e implementação do algoritmo de aprendizado de máquina semi-supervisionado co-training. Technical Report 229, ICMC-USP, 2004. ftp://ftp.icmc.usp.br/pub/BIBLIOTECA/rel_tec/rel_tec229.zip.
[11] Shannon, C. (1948). A mathematical theory of communication. Bell System Technical Journal, Vol. 27, N. 3, pp. 379-423.
[12] Dempster, A.P.; Laird N.M.; Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, Ser B, Vol. 39, pp. 1-38.

utilização de **heurísticas** independente de linguagem, objetivando encontrar o melhor alinhamento entre palavras e sentenças.

Specia, Nunes e Stevenson (2005) apresentam o problema da **desambiguação lexical de sentido** (*word sense disambiguation*) no contexto da **tradução** automática como sendo quando são identificadas apenas variações de significado (de sentido) nas opções de tradução, ou seja, todas as opção são da mesma categoria gramatical (*part-of-speech*). Assim, os autores ressaltam que, neste contexto, sentido (*sense*) significa de fato, tradução. Os autores destacam que existem várias abordagens, mas que o foco do presente trabalho está nas abordagens híbridas, que minimizam o gargalho da aquisição de conhecimento, mas permite aprimorar o conhecimento adquirido. Os autores ressaltam que o presente trabalho faz parte de um projeto de pesquisa maior que visa a criação de uma nova abordagem híbrida simbólica de desambiguação lexical de sentido, a ser aplicado na tarefa de tradução automática Inglês-Português. Os autores destacam que a principal inovação nesta abordagem é a utilização de um formalismo relacional para representar o conhecimento contextual ou de fundo (*background knowledge*). Além disso, os autores ressaltam que um ponto chave nas abordagens híbridas e nas baseadas em corpus é a fonte de conhecimento (*knowledge source*) utilizada no processo de aprendizado de máquina. Assim, o modelo será aplicado a várias fontes de conhecimento para que seja possível compará-las, e identificar a que apresente os melhores (o que, segundo os autores foi realizado em trabalhos anteriores). Tendo feito isto, o objetivo desse trabalho é extrair regras do modelo predito que possam ser usadas como fonte de conhecimento no processo de aprendizado de máquina.

Silva, Vieira e Osorio (2005) iniciam discutindo que os métodos de mineração de dados podem ser adaptados a textos em linguagem natural (mineração de texto ou *text mining*), com o intuito de extrair padrões úteis para organizar e **recuperar a informação** contida em coleções de documentos. Segundo os autores, a primeira etapa da tarefa de mineração de textos é o pré-processamento quando os documentos são representados de uma maneira mais estruturada. Como resultado dessa etapa, usualmente, os documentos são representados por uma lista de palavras (*bag-of-words*), sendo que as palavras sem importância (*stopwords*) são eliminadas e as palavras são reduzidas ao seu radical (*stemming*). Neste trabalho, os autores propõem uma nova técnica de pré-

processamento utilizando informações linguísticas, selecionando as palavras pela sua categoria (nomes, adjetivos, nomes próprios, verbos) e usar a sua forma canônica. Sendo assim, este trabalho tem como objetivo avaliar o **uso de informações linguísticas** no **pré-processamento** de textos tendo em vista as tarefas de classificação e clusterização de documentos[13].

Piltcher *et al.* (2005) afirmam que o presente trabalho trata da correção de palavras dentro de um ambiente de *chat* (salas de bate-papo). As técnicas apresentadas neste trabalho consideram cada palavra separadamente, não sendo avaliados aspectos relacionados à concordância verbal ou nominal. A abordagem utilizada é probabilística (estatística) pois não requer a utilização de analisadores sintáticos (*parsers*). Diante disso, o presente trabalho tem como objetivo utilizar as métricas Levenshtein, Metaphone e Soundex como função de similaridade para **correção automática** de **erros de digitação**.

Rino e Seno (2006) discutem a cerca das duas abordagens principais de **sumarização** automática de textos: abordagem profunda ou rica em conhecimento, baseada em informações linguísticas; e abordagem superficial ou pobre em conhecimento, baseada em informações estatísticas ou empíricas. Segundo as autoras, as abordagens profundas tendem a produzir sumários textuais, ou resumos (p. 1180). Segundo as autoras, o problema de co-referenciação anafórica é evidente na sumarização, diante da possibilidade de uma sentença anafórica referencial ser escolhida para compor um texto sem que sua sentença correspondente antecedente também o seja. As autoras concluem que a ausência de resolução anafórica pode induzir à descontinuidade referencial e, assim, a mensagens incompatíveis de um sumário, se comparado a seu texto-fonte. Diante disso, o presente trabalho tem como objetivo propor a implementação do protótipo RHeSumaRST (Regras Heurísticas de Sumarização de estruturas RST), baseado nos **modelos de estruturação retórica** do discurso da *Rhetorical Structure Theory* – RST (MANN & THOMPSON, 1987[14]) e de coerência global do discurso da Teoria de Veins (CRISTEA *et al*., 1998[15]).

---

[13] No mapa conceitual da seção 4.2.2, esse artigo foi alocado dentro do conceito de 'Pré-processamento dos documentos', visto que, os autores utilizaram as tarefas de classificação e clusterização para avaliar o uso de informação linguística na fase de pré-processamento dos documentos.

[14] MANN, W.C.; THOMPSON, S.A. Rhetorical Structure Theory: A Theory of Text Organization. Technical Report ISI/RS-87-190, 1987.

[15] CRISTEA, D.; IDE, N.; ROMARY, L. Veins Theory: A Model of Global Discourse Cohesion and Coherence. In the Proceedings of the Coling/ACL'1998, pp. 281-285. Montreal, Canada, 1998.

Caseli e Nunes (2006) apresentam a **tradução automática** (*machine translation*) como sendo uma tarefa difícil principalmente por precisar de conhecimento linguístico profundo de duas ou mais linguagens. Além disso, métodos estatísticos podem também ser complicados por requererem a utilização de grandes corporas paralelos e alinhados. Segundo as autoras, torna-se necessário transformar a pouca informação multilíngue em conhecimento linguístico útil para tradução automática. Neste sentido, as autoras destacam que vários métodos têm sido propostos com o intuito de obter automaticamente correspondências estruturais, sintáticas ou lexicais a partir de textos paralelos. Essas correspondências são generalizadas para construir gramáticas de tradução (conjunto de regras de transferência ou tradução) e outros recursos úteis (tais como dicionários bilíngues) para sistemas de tradução automática. Diante disso, o presente trabalho pretende apresentar o projeto ReTraTos, que tem como objetivo **induzir** conhecimento linguístico para tradução – **regras de transferência** e dicionário bilíngue – combinando técnicas diferentes para o português do Brasil.

Balage Filho *et al.* (2006) apresentam os resultados obtidos durante a participação no CLEF 2006, da aplicação de um sistema de **sumarização** na tarefa de **responder perguntas** em um único idioma (*monolingual Question Answering*) para textos em português, usando o método orientado a tópicos. Assim, cada pergunta será considerada um tópico a partir do qual o sumário deverá ser construído, esperando conter a resposta apropriada.

Enembreck *et al.* (2006) afirmam que um problema comum nas organizações é identificar pessoas ou competências para formar uma equipe especializada tanto em ambientes acadêmicos como industriais. Assim, o objetivo deste trabalho é, dada a descrição de um projeto de pesquisa e desenvolvimento, extrair informações que permitam **identificar pessoas que têm competências**[16] para participar do projeto, utilizando a base de currículos Lattes, e técnicas de recuperação de informação textual.

Leite e Rino (2006) apresentam como alternativas de extração em **sumarização** automática, métodos empíricos ou técnicas estatísticas que têm como objetivo identificar segmentos relevantes do texto que devem ser incluídos no sumário (extrato) final. Ao mesmo tempo, os trechos irrelevantes devem ser filtrados

[16] No mapa conceitual da seção 4.2.2, esse artigo foi alocado dentro de exemplos de aplicação do PLN com “Identificação automática de equipes (times)”.

e descartados. Segundo os autores, algumas **[seleção de] características** são usadas para produzir regras neste processo, que referenciam diferentes tipos de informações, como gramatical, funcional ou posicional (como palavras do título). Em um trabalho anterior da segunda autora, e também analisado nesta tese, o sumarizador Supor foi comparado com outros sistemas e obteve o melhor desempenho. No entanto, os autores afirmam que não foi possível identificar, dentre as características usadas pelo Supor, qual ou quais apresentaram papel mais significante. Diante disso, o presente trabalho tem como objetivo avaliar o impacto dessas características no processo de sumarização automática, usufruindo do fato do sistema Supor ser configurável pelo usuário. Como consequência, os autores propõem o desenvolvimento de uma nova versão da ferramenta que utilize características mais informativas e apresente melhor desempenho e escalabilidade.

Moraes e Strube de Lima (2007) discutem a importância de se definir métodos eficazes de classificação, e consequente recuperação de informação, de grandes quantidades de documentos. Além disso, diante da diversidade dos textos, um número elevado de categorias pode ser definido, dificultando assim a navegação e a busca de informações sobre essas categorias. Segundo as autoras, uma alternativa comumente adotada é a organização dessas categorias em hierarquias. Sendo assim, o presente trabalho tem como objetivo experimentar a **categorização hierárquica de textos** em língua portuguesa em uma escala maior, analisando os resultados obtidos e as principais dificuldades encontradas. Os parâmetros usados por Langie (2004[17]) no processo de classificação serão analisados, bem como a eficácia do processo de categorização em relação ao pequeno número de documentos usados para treinar os classificadores.

Kinoshita *et al.* (2007) destacam que, apesar do OpenOffice ser um projeto código aberto, multi plataforma e idiomas, utilizado por inúmeros usuários e organizações que apóiam software livres, opções de corretores gramaticais para esse editor de texto ainda é um grande problema. Diante disso, o presente artigo tem como objetivo descrever o CoGrOO – **Corretor Gramatical** para OpenOffice, um projeto de um corretor ortográfico para a língua portuguesa do Brasil. Dentre os erros normalmente observados nos textos, os autores destacam erros de digitação (*spelling*), gramaticais (quando as regras de gramática não são observadas), de

[17] Langie, L. C. (2004) “Um Estudo sobre a Aplicação do algoritmo k-NN à Categorização Hierárquica de Textos”. Dissertação de Mestrado. Faculdade de Informática, PUCRS, 126 p.

estilo e semânticos (que são fortemente dependentes do contexto).

Specia, Stevenson e Nunes (2007) apresentam o problema da **desambiguação lexical de sentido** (*word sense disambiguation*) como sendo a correta identificação do significado de palavras ambíguas dentro do contexto. Segundo os autores, abordagens comumente utilizadas para desambiguação de nomes consideram o contexto próximo à palavra ambígua fazendo-se uso de uma lista de palavras (*bag-of-words*), ou, no caso da desambiguação de verbos, de conhecimento mais específico, como relação do verbo com outros itens da sentença. Diante disso, este trabalho tem como objetivo apresentar uma abordagem de desambiguação lexical de sentido baseada em corpus com conhecimento de fundo (*background knowledge*). Em outras palavras, é apresentada uma estratégia híbrida que combina conhecimento e evidências baseadas em corpus, e aplica um formalismo para permitir a representação do conhecimento profundo sobre os exemplos. Neste trabalho foi dada ênfase a **tradução** de dez **verbos** ambíguos do Inglês para o Português.

Silva e Vieira (2007) apresentam como tema central do artigo o problema da **categorização[18] de textos** onde os textos (ou documentos) são organizados em categorias pré-definidas, de acordo com o conteúdo que os compõem. Segundo as autoras, o objetivo deste trabalho é avaliar a utilização de **informações linguísticas** para seleção de características na etapa de pré-processamento de categorização de textos da língua portuguesa e comparar seus efeitos em relação a dois métodos de aprendizado distintos, tais como árvores de decisão e *Support Vector Machine (SVM)*.

Milidiu, Duarte e Cavalcante (2007) apresentam o problema de encontrar os nomes próprios *(named entity recognition)* em um texto e classificá-los em categorias usualmente utilizadas: personalidades (pessoas), organizações e localidades. Neste trabalho, foi considerado somente o problema de reconhecimento de nomes próprios livre de contextos, ou seja, ambiguidades geradas pela omissão de certas palavras não foram consideradas. Assim, o objetivo deste trabalho é avaliar a utilização de três técnicas de aprendizado de máquina no processo de **reconhecimento de nomes próprios**: modelos de cadeias de Markov (*Hidden*

[18] Neste artigo, as autoras definem categorização como sendo o processo de alocar documentos em categorias pré-definidas. No entanto, no mapa conceitual construído no escopo desta tese, adotou-se categorização como sendo sinônimo de clusterização, que ocorre quando os documentos são agrupados em função das suas similaridades, visto que não se tem conhecimento prévio das categorias.

Markov Models – HMM), aprendizado baseado em transformações e *Support Vector Machine (SVM).*

Caseli *et al.* (2008) apresentam uma **ferramenta visual gratuita** desenvolvida baseada no alinhador lexical híbrido LIHLA, proposto anteriormente pelos próprios autores. O método implementado tenta encontrar o melhor **alinhamento entre palavras e sentenças** baseado em heurísticas independente da linguagem, e alinhamento estatístico entre palavras simples definido em dicionários probabilísticos bilíngue construído automaticamente. Assim, usando dois dicionários bilíngues construídos pelo NATolls e heurísticas, o LIHLA tentat encontrar o melhor alinhamento lexical em um par de sentenças paralelas por meio de um processo iterativo.

Aziz, Pardo e Paraboni (2008) discutem sobre a ampla utilização que os métodos estatísticos tem encontrado em várias tarefas do processamento de linguagem natural, dentre elas a de **tradução automática**. Segundo os autores, essas abordagens têm apresentado resultados superiores que os baseados em regras para as linguagens ditas distantes, tais como inglês e árabe. Diante disso, este trabalho tem como objetivo **comparar** a utilização da **abordagem estatística** com a baseada em **regras de transferência** compartilhada[19] para linguagens próximas (*closely-related*), tais como o espanhol e o português.

Morais e Ambrosio (2008) ressaltam que várias pesquisas têm destacado a importância da categorização de documentos para auxiliar na recuperação: alguns utilizam técnicas de mineração de texto (*text-mining*) para identificar o contexto semântico dos documentos, enquanto que outros representam o contexto usando ontologias. No entanto, os autores destacam que não foi encontrado na literatura alguma pesquisa que combine ambas as estratégias – ontologia e técnicas de mineração de texto – para desenvolver mecanismos de **categorização automática de documentos**. Diante disso, este trabalho tem como objetivo descrever o sistema, desenvolvido no mestrado do primeiro autor[20], que dado um documento e um domínio representado por uma **ontologia**, analisa automaticamente se o documento é relevante, utilizando técnicas de **mineração de texto**.

[19] Corbí-Bellot, A.M.; Forcada, M.L.; Ortiz-Rojas, S.; Pérez-Ortiz, J.A.; Ramírez-Sánchez, G.; Sánchez-Martínez, F.; Alegria, I.; Mayor, A.; Sarasola, K. An open-source shallow-transfer machine translation engine for the romance languages of Spain. 10th Annual Conference of the European Association for Machine Translation (2005) 79-86.

[20] MORAIS, E. Contextualização de Documentos em Domínios Representados por Ontologias Utilizando Mineração de Textos. Master Thesis, Universidade Federal de Goiás, 2007.

Caminada, Quental e Garrao (2008) destacam que a identificação de expressões multivocabulares é uma das tarefas lexicográficas que mais se beneficiam de métodos e aplicações computacionais. Neste sentido, o presente trabalho tem como objetivo apresentar uma ferramenta extensível para a busca e **identificação de bigramas e trigramas** multivocabulares da língua portuguesa, baseados em padrões gramaticais definidos por um processo de anotação. A ideia central é quebrar os multivocábulos previamente identificados pelo parser PALAVRAS e validar estes multivocábulos, assim como identifica novas expressões.

Seno e Nunes (2008) destacam que identificar passagens ou trechos similares de textos tem desempenhado um importante papel em várias aplicações de processamento de linguagem natural, tais como geração de parágrafo, sumarização automática, construção de ontologias, bibliotecas digitais, dentre outras. Diante disso, o presente trabalho tem como objetivo propor a **avaliação do framework SiSPI** – *Similar Short Passages Identifier*, baseado num método não supervisionado e incremental de clusterização. A hipótese desse artigo é que uma abordagem de **clusterização** incremental apresenta resultados melhores que usando somente métricas estatísticas de similaridades.

Aziz, Pardo e Paraboni (2009) destacam que, nos últimos anos, a pesquisa em **tradução automática** usando **métodos estatísticos** tem alcançado resultados de qualidade com a utilização de técnicas baseadas em frase, ou seja, uso de modelos de tradução que consideram o **alinhamento** de sequências de palavras, motivado ou não liguisticamente. Esse processo envolve a escolha de heurísticas adequadas de alinhamento, algoritmos de decodificação (*decoding)*, modelos adequados, e um grande número de opções de configurações. Ainda segundo os autores, para alcançar resultados melhores, os sistemas desse tipo devem ser customizado para um determinado par de linguagens ou domínios. Sendo assim, o presente trabalho tem como objetivo investigar inúmeros parâmetros envolvidos no treinamento e no processo de decodificação (número de palavras que podem ter a ordem invertida durante a tradução), e comparar os resultados obtidos nos experimentos.

Seno e Nunes (2009) apresentam como problemática a **fusão de sentenças**, que consiste em produzir, dadas duas ou mais sentenças similares, uma única sentença que combina informações daquelas sentenças, e ao mesmo tempo elimina as informações redundantes. Segundo as autoras, neste artigo é

apresentado um modelo inédito para a fusão de sentenças em português.

Salles *et al*. (2009) destacam que apesar do tempo ser uma dimensão importante para qualquer espaço informacional, a maioria das técnicas atuais de **classificação automática de documentos** não consideram a **evolução temporal dos documentos**, ou seja, que a variação na definição dos termos e das classes ao longo do tempo, tende a tornar o conjunto de treinamento muito confuso, impactando negativamente nos classificadores que negligenciam esta evolução. Diante disso, o presente trabalho propõe uma nova abordagem para o tratamento dos efeitos temporais em algoritmos de classificação, segundo os autores, já conhecidos, derivando classificadores robustos temporalmente.

Braga, Monard e Matsubara (2009) afirmam que nos últimos dez anos, a **classificação de textos** tem recebido grande atenção da comunidade de aprendizado de máquina, e que inicialmente a ênfase estava nos algoritmos supervisionados, apesar de haver algum interesse em abordagens semi-supervisionada. Segundo os autores, no aprendizado semi-supervisionado um classificador pode ser obtido a partir de dados rotulados e não rotulados. Os autores destacam que, na classificação, a maneira mais comumente utilizada para representar os documentos é extraindo palavras (ou unigramas). Além disso, os autores ressaltam que **combinar unigramas e bigramas** não tem apresentado melhorias significativas na classificação supervisionada de textos. Diante disso, o presente trabalho tem como objetivo verificar se essa constatação é observada também no aprendizado semi-supervisionado.

Villavicencio, Caseli e Machado (2009) apresentam algumas abordagens para **identificação de** expressões **multi-palavras** (*multiwords expressions*) em corpora técnicos. Segundo os autores, estas expressões podem ser definidas como sendo a combinação de palavras que apresentam idiossincrasias lexical, sintática, semântica, pragmática ou estatística. O objetivo deste trabalho é determinar a influência de diferentes fontes de conhecimento na tarefa de identificação de expressões multi-palavras. Dentre as fontes avaliadas estão informações linguísticas e técnicas estatísticas.

4.2.1.2. Metodologia Adotada nos artigos focalizados

Nesta seção, procurou-se apresentar a metodologia utilizada em cada artigo selecionado para análise, com o intuito de identificar os métodos e as técnicas utilizados nestes trabalhos. Em outras palavras, esta categoria tem como objetivo revelar o que os pesquisadores da área têm usado em termos de ferramentas, tanto computacional como linguística, para resolver os problemas apresentados. As técnicas e métodos que emergiram a partir desta análise foram incluídos nos mapas conceitual construídos, que serão apresentados na próxima seção, e estão destacados em negrito nos parágrafos que se seguem.

Semeghini-Siqueira, Costa e Cohn (1986) propõem que a compreensão de linguagem natural por computador seja feita interagindo o componente sintático (formalizações de Chomsky) e o componente semântico (contribuições da semântica estrutural, gerativa e argumentativa) com incursões pelo componente pragmático. Assim, segundo os autores, o ponto de partida deste trabalho é a **gramática gerativo-transformacional** mas que, em consonância com a corrente semanticista, não seria possível compreender a sintaxe desvinculada da semântica e recorreram a Fillmore (1968) para postular que a informação semântica faz parte do "plano de composição da frase" (p. 116). Segundo os autores, a compreensão da linguagem seria feita numa rede de ativação onde os componentes são tratados paralelamente: o conhecimento específico da língua sobre a estrutura dos sintagmas, as relações entre as unidades lexicais[21] e os conhecimentos sobre o universo linguístico escolhido (no caso do trabalho, Astronomia) são trabalhados concomitantemente. Os autores complementam que após muitas confrontações, foi possível depreender um conjunto de relações lógico-semânticas entre as unidades lexicais, ou seja, associações hierárquicas entre os significados. Estas relações lógico-semânticas foram apresentadas utilizando-se as notações da linguagem de programação Prolog (p. 119). A análise sintático-semântica é realizada utilizando-se regras de gramáticas e sintagmas nominais, preposicionais, adjetivos e frases relativas (onde os verbos são analisados para gerar conexões entre os itens lexicais) (p. 122).

O artigo **Ziviani e Albuquerque (1987)** apresenta um novo método para detecção de termos através da **utilização de um índice** que reduz drasticamente a

[21] Os autores ressaltam que os termos linguísticos palavra, vocábulo, signo, item lexical e unidade lexical foram considerados equivalentes, apesar de não serem iguais.

quantidade de dados a serem percorridos. O índice é criado através da **obtenção de um arquivo de assinaturas**. Segundo os autores, um arquivo de assinaturas de palavras é uma sequência de inteiros, onde cada inteiro é obtido aplicando-se uma função de transformação sobre as palavras do texto. Uma árvore Patricia é construída sobre as assinaturas das palavras do texto, permitindo a detecção de termos em tempo proporcional ao logaritmo base dois do número de assinaturas obtidas do arquivo original. Os autores apresentam inúmeras vantagens em converter a palavra em um número inteiro, dentre elas a compreensão do tamanho (p. 177 e 178). Apesar dos autores apresentarem um apelo computacional maior, eles também fazem uso de estratégias sabidamente da área de PLN: elimina stopwords (apesar de não usar esta denotação), consultando uma lista "negativa" de palavras e citando van Rijsbergen (1979); elimina sufixos, citando Lovins (1968) e van Rijsbergen(1979); resolve problemas de grafias ambíguas usando o metodo Soundex (Knuth, 1973); e tratam erros de digitação utilizando medidas de semelhança entre duas cadeias (p. 178).

Ripoll e Mendes (1988) destacam que existem duas hipóteses de **resolução da ambiguidade léxica** pelo homem: na primeira todos os possíveis significados de uma palavra são ativados em paralelo e depois, **através do contexto é feita a escolha do significado adequado**[22]. A outra possibilidade, segundo os autores, é que **o contexto restrinja a priore a escolha precisa do significado adequado** (p. 297). Quanto a representação dos significados das palavras, segundo os autores, existem duas alternativas utilizando um modelo conexionista: a **abordagem localizada**, onde cada conceito (significado de uma palavra) corresponde a uma unidade da rede; e a **abordagem distribuída**, onde cada conceito corresponde a um padrão de ativação de determinadas unidades da rede, que são também compartilhadas por outros conceitos (p. 298). Para ilustrar, os autores citam algumas **características** que normalmente são usadas para representar verbos e substantivos, tais como se existe um agente, se o verbo é causal, qual a natureza da mudança, dentre outras (p. 298). O sistema proposto resolve ambiguidade léxica de verbos em **três níveis**: no **nível léxico**, cada unidade corresponde a uma palavra que consta no dicionário da linguagem. As unidades do nível léxico ativam as do **nível de significado** das palavras; neste nível, cada

[22] No mapa conceitual, adotou-se as expressões “Identificar o significado pelo contexto” e “Contexto restringe os significados possíveis” para representar as duas hipóteses de resolução de ambiguidades.

unidade representa um possível significado das palavras; cada unidade do nível léxico ativa uma ou mais unidades do nível de significado. A decisão de qual é o significado correto é tomada através das ativações vindas do **nível dos casos verbais**; no nível de casos verbais, cada unidade corresponde aos possíveis casos verbais de cada significado do verbo (dentre as inspirações, utilizou-se os **casos de Fillmore**) (p. 300).

Fusaro e Ziviani (1989), assim como em Ziviani e Albuquerque (1987), tem como enfoque o método de recuperação de informação textual: codificação dos termos, construção do **arquivo invertido** e definição de uma linguagem de consulta. A implementação da linguagem apresentada envolveu a construção de um interpretador para a gramática da linguagem (que pode ser feita manual ou automaticamente), e a criação de rotinas semânticas para cada sentença válida da gramática (responsáveis por localizar as ocorrências dos termos da consulta e retornar os documentos que apresentam tais termos) (p. 294 e 295).

Strube de Lima (1990) apresentou uma revisão de literatura quanto a métodos e técnicas aplicados no tratamento da língua natural, mais especificamente na correção ortográfica automática de textos. Quanto aos tipos de erros e estratégias de correção, a autora destaca que do ponto de vista linguístico clássico, os **erros** podem ser distribuídos em três níveis diferentes: **o nível léxico, o nível sintático e o nível semântico**, em outras palavras, erros a nível da palavra, da construção ou do sentido, respectivamente. Segundo a autora, no nível léxico estão os **erros de ortografia**, os **erros fonéticos** (troca de letras que apresentam som similares) e os **erros de geração** (uso incorreto de uma desinência ao construir um plural, por exemplo). Os erros tipográficos (**erros de digitação**) são incluídos nesta categoria. No nível sintático, considera-se a correção no que diz respeito às **regras de construção da frase** e à **concordância entre seus componentes**. Já no nível semântico, uma **interpretação da frase** é concebida de tal maneira que seja possível verificar o seu significado (p. 44 e 45).

Em **Leffa (1991),** o objetivo da pesquisa foi comparar a utilização de um **dicionário tradicional com um eletrônico**. O dicionário **eletrônico deveria incluir** em cada verbete não apenas **informação lexical** como também **informação gramatical**. A seleção dos itens para compor o dicionário eletrônico foi feita de acordo com dois critérios básicos: frequência dos termos no uso escrito da língua inglesa e contrastividade com o termo correspondente da língua portuguesa. Para

levantamento dos termos mais frequentes, usou-se listas já existentes, enquanto que a constrastividade foi baseada num trabalho anterior do próprio autor. A lista final chegou a 4.700 verbetes com cerca de 10.000 valores semânticos (bem acima dos 3.300 termos considerados apropriados para compreensão geral de um texto, de acordo com o autor) (p. 192).

Rocha *et al.* (1992) apresentam um sistema com **rede neural artificial evolutiva e hierárquica** de três níveis, para identificar o conteúdo em comum de um grupo de textos de um banco de dados, e produzir sumários (lista de tópicos). A primeira rede neural (***word net***) é responsável por escanear os textos e **identificar as palavras mais frequentes** e mais significativas. Esta rede provê a entrada da próxima rede: rede de frases (***phrase net***). Esta rede tem como objetivo **identificar as associações (frases) entre as palavras mais frequentes** no banco de dados. Estas frases são usadas como entradas da próxima rede: rede de texto (***text net***). Esta última rede é responsável por **encontrar possíveis padrões de textos** no banco de dados (p. 822). Os autores destacam que redes neurais evolutivas são redes que utilizam aprendizado para ajustar a estrutura de seus neurônios para representar eventos no ambiente externo. Desta forma, é possível representar as palavras, as frases ou textos no banco de dados, assim como modificar os neurônios para acomodar pequenas variações das mensagens recebidas, tais como as promovidas por erros de digitação, no caso de palavras ou pequenas variações na composição de frases e textos (p. 819).

Rocha, Rocha e Huff (1993) destacam a importância de se traduzir termos expressos em diferentes vocabulários médicos usando um processo completamente automatizado. Segundo os autores, durante vinte anos tem se concentrado em desenvolver modelos **alternativos para representação de dados clínicos de pacientes**. Segundo os autores, a **melhor opção** é o **modelo de definição de eventos (*Event Definition Model*)** que apresenta uma **visão conceitual dos registros médicos** como sendo uma sequência de eventos clínicos. Os dados clínicos são **semanticamente representados** por um **quadro (*frame*) de atributos**. Segundo os autores, a criação e a manutenção dessas representações requerem um léxico, onde os conceitos são representados na sua forma canônica (p. 690). Este **léxico** foi **criado manualmente** avaliando-se todos os termos presentes nos vocabulários usados: o dicionário de dados de um **sistema especialista** (*Iliad - OpenClinical AI Systems in Clinical Practice*) como vocabulário de origem e o ***UMLS***

Metathesaurus como vocabulário alvo (*target*).

Robin (1994) propõe um modelo de sumarização automática que primeiro **constrói um rascunho** contendo somente **os fatos essenciais do texto**, e depois vai **incrementando-o com fatos anteriores (*historical background*)** presentes em um limite de espaço. Segundo o autor, este modelo requer um novo tipo de **conhecimento linguistico: operações de revisão** (*revision operations)*, especificando as várias maneiras nas quais uma rascunho pode ser transformado de forma concisa, a fim de acomodar uma nova informação. O sistema desenvolvido tem cinco componentes principais: um gerador de fatos, um planejador de frases, um lexicalizador (*lexicalizer*), um revisor e um unificador. Internamente, um rascunho é representado como uma estrutura de características em três camadas: **especificação semântica profunda, especificação semântica de superfície e especificação gramatical profunda**.

Julia, Seabra e Semeghini-Siqueira (1995) propõem a construção de um analisador (*parser*) que gera automaticamente regras semânticas. O analisador proposto corresponde a uma estrutura (como definido por Piaget), que automaticamente gera **regras semânticas** (abstrações *lambda*) durante a análise, orientada por um **método heurístico**. A parte sintática da gramática é expressa por meio de regras, tais como as **regras de gramática proposta por Chomsky**. Em outras palavras, o analisador corresponde a uma estrutura representada por um sistema formal cujos axiomas são abstrações de casos e as inferências são regras de redução. O analisador implementado é baseado em **algoritmos de busca** cujo objetivo é descobrir um caminho da árvore no qual a folha seja a categoria de significado (p. 806 e 807). Segundo os autores, no decorrer do tempo, diversas teorias linguísticas têm tentado explicar os aspectos relacionados com palavras, frases e textos considerando apenas os seres humanos como interlocutores. No entanto, os autores afirmam que, por não ser possível **compreender o significado de um texto** com base apenas nas **palavras** e na **estrutura sintática**, volta-se o estudo para as operações linguísticas e **processos cognitivos** que estão implícitos na **produção e recepção do texto**. Coesão e coerência são então apresentados como princípios fundamentais. Segundo os autores, vários fatores podem ser usados para explicar o que faz uma produção verbal tornar-se um texto: a coesão, a coerência, a situação, a informação, a intenção, a intertextualidade, a aceitabilidade, a inferência, a pertinência, o conhecimento do mundo, dentre outros. Neste estudo,

os autores propõem que a coesão é uma espécie de avaliação do texto realizada pelo leitor. Em outras palavras, a coesão é objetiva e pode ser reconhecida automaticamente, já a coerência é subjetiva, pois cada autor a identifica de uma maneira diferente, de acordo com sua visão do mundo (p. 807).

Barros (1996) propõe a construção de um modelo de **resolução de anáfora pronominal** sem a utilização de modelo de mundo (*world model*) para assim garantir a portabilidade do sistema. O modelo proposto, chamado de **módulo de discurso** (*discurse module*), é baseado em uma **lista de candidatos** a resolução da anáfora. Esta lista é **incrementada dinamicamente** para cada consulta analisada. Assim, quando uma anáfora é encontrada numa consulta, os candidatos são selecionados nesta lista tendo como base **informações sintáticas e de domínio**. Quando várias candidatas são selecionadas, as opções são apresentadas ao usuário para que possa escolher uma, ou inclusive rejeitar todas. Assim, a autora conclui afirmando que este modelo provê um processo semi-automático de resolução de anáforas independente do domínio.

Rosa (1997) propõe **mapear papéis temáticos** em regras **semânticas** usando vetores de **características** organizados com base nas relações temáticas entre o verbo e as outras palavras de uma frase. Cada palavra é representada por um vetor de bits no qual cada subconjunto tem um significado associado. O objetivo do trabalho é utilizar a ideia de representação de **características semânticas** com o intuito de construir uma estrutura capaz de analisar e aprender a correta **atribuição de relacionamentos temáticos das palavras na sentença** (p. 241). A rede neural construída tem três camadas, sendo que a entrada da rede é o vetor de características, enquanto que a saída é a estrutura temática da sentença. O autor destaca que, para vários problemas de Inteligência Artificial, é impossível fornecer para a rede, todos os valores possíveis de entrada. Segundo ele, esta deficiência é resolvida pelas **redes *backpropagation*** usando-se mecanismos de generalização, ou seja, a rede tem condições de fazer uma espécie de interpolação dos dados fornecidos e prover uma saída para aquelas situação, até então desconhecidas (p. 242).

Oliveira e Wazlawick (1998) propõe a resolução de anáforas usando **redes neurais artificiais**. As sentenças utilizadas tanto no treinamento como no teste do modelo, são constituídas por padrões tais como "sujeito verbo objeto. Ele/ela verbo objeto". O modelo é composto por duas redes neurais artificiais: o

***parser* (rede simples recorrente)** e o **segmentador (rede multicamadas *feedforward*),** cada uma com função específica. O *parser* recebe como entrada uma sequência de palavras e como saída a representação dos papéis (*case role representation*) (p. 1195). Já no segmentador, a entrada é a próxima palavra da sequência mais a saída do parser (p. 1196).

Carvalho e Strube de Lima (1999) propõem a utilização de **sistemas multi-agentes** para o processamento de língua natural e apresentam que existem várias maneiras de distribuir o conhecimento linguístico entre os agentes. Segundo as autoras, quando os agentes são associados somente a níveis linguísticos, o número de agentes é bastante reduzido, se comparado à abordagem onde existe um agente por palavra da sentença. Entretanto, a complexidade dos agentes envolvidos no processo cresce consideravelmente. Para testar essas abordagens, as autoras propõem o desenvolvimento de dois sistemas: no primeiro, os **agentes** foram associados à **categoria morfossintática** das palavras e no segundo, os agentes foram associados a **níveis de conhecimento e a fenômenos linguísticos**. As autoras complementam que embora os dois sistemas tenham sido desenvolvidos para a língua portuguesa, as ideias principais podem ser generalizadas para outras línguas. O modelo é inspirado em um trabalho de 1995 para a língua francesa, mas se difere em vários aspectos, sendo a principal diferença, o fato de usar conhecimento semântico. Sob o ponto de vista linguístico, o sistema faz análise léxico-morfológica, sintática e semântica. Os agentes contém ainda dicionários, gramáticas e redes conceituais. Segundo as autoras, o tratamento da frase começa com uma **análise léxico-morfológica**, através do **agente morfológico**, que envia seus resultados para o **agente sintático**, para que este possa construir a **árvore de derivação**; o agente sintático, por sua vez, envia seus resultados para o **analisador semântico** para a construção da **estrutura semântica**.

Kinoshita (1999) propõe um **sistema de tradução baseado em exemplos** que foram **extraídos da Bíblia** em grego e suas traduções para inglês e português. Segundo o autor, os exemplos são anotados com números de acordo com a **anotação de James (*James Strong annotation*).** Segundo o autor, **palavras, bigramas e trigramas** são extraídos dos exemplos, juntamente com suas traduções e uma **máquina de estados finitos é construída**. A tradução de novas sentenças então é feita identificando-se as palavras, bigramas e trigramas que já são conhecidas e atribuindo-se a tradução correspondente.

Barcia *et al.* (1999) propõem a utilização da técnica de Inteligência Artificial de **Raciocínio baseado em Casos** (RBC) para solução de problemas jurídicos. Segundo os autores, Raciocínio Baseado em Casos usa experiências anteriores e semelhantes para a solução de um novo problema, baseando-se no princípio de analogia, assumindo-se que problemas semelhantes tem soluções semelhantes. Por esta razão, os autores afirmam que o RBC é uma técnica muito adequada ao domínio jurídico, pois utiliza o mesmo tipo de raciocínio utilizado pelos juristas na solução de um problema. Segundo os autores, os casos jurídicos são representados na forma de um caso que consiste no texto do documento original e um conjunto de índices na forma de pares atributo-valor. Os atributos dos documentos textuais, usados como índices para a recuperação, têm que indicar a utilidade das informações do caso na situação presente. Para reforçar esta forma de representação, o conhecimento do domínio é incluído na forma de um **vocabulário** jurídico **controlado** e um dicionário de termos. É este conhecimento de domínio que permite a recuperação dos documentos e o processo de extração automático, através da identificação de expressões indicativas e relevantes dos textos jurídicos em linguagem natural, juntamente com a modelagem explícita da semelhança destes termos jurídicos. O vocabulário controlado define o valor dos índices usados, enfocando no domínio de aplicação específico (no caso, jurisprudência criminal). Segundo os autores, este vocabulário controlado é constituído de termos jurídicos que são usados “nos Tribunais para representar fatos enquadrados normativamente e que, por sua vez, são definidos através dos termos-chave de uma norma”. Segundo os autores, o dicionário foi desenvolvido por profissionais do direito com base na experiência deles. Segundo os autores, o processo de recuperação é dividido em etapas, sendo que inicialmente um problema jurídico é descrito em linguagem natural para iniciar o processo de recuperação pelo usuário; em seguida, a similaridade do problema inicial com cada caso na base de casos é determinado por uma métrica da similaridade (relativo a importância de cada atributo para a pesquisa); e finalmente, os dez melhores casos são ordenados de acordo com o grau de similaridade. Os casos ordenados são apresentados ao usuário de modo que ele possa visualizar um resumo de todos os casos, com algumas informações sobre o documento jurídico ou, se o usuário quiser, o documento completo. Assim, o usuário poderá escolher entre os melhores casos, aquele documento que é o mais adequado.

Berber Sardinha (1999) apresenta um trabalho teórico onde algumas questões no estudo da padronização são levantadas: quais os padrões lexicais dos quais a palavra faz parte; a palavra se associa regularmente com outros sentidos específicos; em quais estruturas ela aparece; há uma correlação entre o uso/sentido da palavra e as estruturas das quais ela participa; e a palavra está associada com (uma certa posição na) organização textual? Para verificar se dois ou mais itens lexicais formam um padrão é necessário saber se a **co-ocorrência é significativa**, isto é, se ocorre mais vezes do que o esperado por acaso. Segundo o autor, para saber se a co-ocorrência entre os itens pesquisados é significativa, é necessária a obtenção de **estatísticas de co-ocorrência**, através da aplicação de fórmulas matemáticas especializadas. Desta forma, a função das estatísticas é apontar se os itens formam colocações ou se são co-ocorrências espúrias. Segundo o autor, o trabalho se estrutura em torno de **estudos de caso. Quatro cenários** foram selecionados por serem potencialmente relevantes para a área. O primeiro discute, do ponto de vista de sua prosódia semântica (associação entre itens lexicais e conotação – positiva, negativa ou neutra), a expressão **'tocando para a frente'**. O segundo e o terceiro centram-se na prosódia semântica de dois verbos comuns do português: **'causar'** e **'acontecer'**, respectivamente. O quarto e último estudo enfoca o advérbio **'absolutamente'**. O autor ressalta que a motivação para esses estudos é diferente. O primeiro estudo é de caráter exploratório e visa buscar evidências no corpus que confirmem ou desconfirmem a intuição do falante nativo. Em contrapartida, o autor destaca que os três últimos estudos possuem uma orientação contrastiva, nos quais o intuito é comparar os resultados do português com o de outras línguas.

Em **Villavicencio (1999)** é usada uma **rede ortogonal de heranças múltiplas** para representar informação lexical. Segundo autora, diferentes redes são usadas para representar diferentes tipos de conhecimento linguístico. Assim, as **regularidades linguísticas** são representadas próximo ao topo da rede, enquanto que os nodos mais abaixo da rede são usados para representar as **sub-regularidade ou exceções**.

Jose Neto e Menezes (2000) apresentam a arquitetura básica do **etiquetador morfológico treinável** proposto neste trabalho. Segundo os autores, o modelo é dividido em três módulos: o primeiro responsável pela **etiquetação inicial de palavras conhecidas**; o segundo que realiza a **etiquetação inicial de palavras**

desconhecidas, e um terceiro e último, que promove um **refinamento contextual** (p. 55). No primeiro modulo, ocorre a obtenção da etiqueta mais provável para as palavras conhecidas: a estrutura de dados utilizada neste módulo é uma **árvore n-ária de letras**, utilizada para armazenar o **léxico**, contendo uma lista associada a cada uma de suas folhas. Esta lista é utilizada para armazenar as várias etiquetas morfológicas possíveis, em ordem decrescente de frequência de aparecimento. Segundo os autores, uma vantagem, inerente a esta estrutura em forma de árvore, é que ocorre naturalmente uma compressão do tamanho da base de dados pelo fato de todos os prefixos serem armazenados apenas uma vez na estrutura. O segundo módulo atribui uma etiqueta para as palavras desconhecidas, **com base em sufixos**. Os autores complementam que com base nas últimas letras dos itens lexicais encontrados no corpus de treinamento e nas etiquetas morfológicas associadas a cada um deles, este módulo infere um mapeamento que é usado na etiquetação de itens lexicais que nunca apareceram no corpus de treinamento (palavras desconhecidas). Segundo os autores, a heurística por trás deste módulo tem um embasamento linguístico: sabe-se que, nas línguas cujas palavras apresentam a estrutura prefixo + radical + sufixo, o sufixo de uma palavra tem uma forte correlação com a sua categoria morfológica. O terceiro módulo é o refinador contextual. Segundo os autores, ele é responsável por escolher, dentre as várias etiquetas possíveis para uma dada palavra, aquela que mais se adapte ao contexto em que esta palavra se encontra. Os autores complementam que a ideia central do método baseia-se na utilização de uma janela de três posições (etiqueta já consumida anteriormente, a da posição e a próxima).

Berber Sardinha (2000) tem como objetivo analisar a **manutenção da prosódia semântica** dentro do contexto da **tradução** de padrões lexicais. Segundo o autor, o primeiro **item lexical analisado** para o estudo foi o **'commit',** por acreditar que o mesmo possui prosódia semântica negativa. Assim a questão em foco seria verificar se o equivalente 'cometer' teria também uma prosódia semântica negativa. O segundo item selecionado como estudo de caso foi a expressão verbal (*phrasal verb)*, **'set in'**, pelo fato dele apresentar associações negativas. Para cada tradução do item lexical selecionado para estudo, foi feita uma busca no corpus para observar a frequência de ocorrência dos mesmos. Os mais bem colocados, em termos da frequência de ocorrência, e valores aceitáveis de informação mútua e o teste de associação T-score) eram analisados.

Padilha e Viccari (2000) descrevem o processamento de **transformações morfológicas** do português, utilizando **transdutores**, um tipo de máquina de estados finitos. Através de uma extensão da **morfologia de dois níveis**, **gramáticas** para a ortografia e **fonologia** portuguesa foram desenvolvidas. Segundo os autores, essas gramáticas são "traduzidas nos transdutores que efetivamente realizam as transformações de forma simples, eficiente, bidirecional e localizada, sem misturar etapas linguísticas diversas" (p. 51).

Larocca Neto *et al.* (2000) afirmam que as notícias representam um campo importante de aplicação de sumarização automática. O objetivo deste trabalho é desenvolver um sistema treinável baseado em aprendizado de máquina para sumarização de notícias. Segundo os autores, para obter a **estrutura aproximada do texto**, combinou-se a saída de um **algoritmo de clusterização aglomerativa** com a detecção de **sentenças** que são **essenciais** (capturam a ideia principal do documento) ou **cenário** (*background* que contém informação adicional mas não essencial). Segundo os autores, a detecção das sentenças de cenário é baseada em heurísticas. Em seguida, os autores complementam que, usando a representação aproximada da estrutura do texto é possível selecionar um conjunto de **características** para usar num sistema de sumarização automática (similares às **heurísticas de Strzalkowski**, 1998), dentre elas: indicador de conceitos principais, usando **medidas estatísticas** tais como frequência do termo ou o *tfidf* (*term frequency inverse document frequency*); ocorrência de nomes próprios; ocorrências de anáforas; ocorrência de marcadores de discursos, tais como 'por que', 'além disso', dentre outros; conexão com outras sentenças, identificada de maneira similar aos mapas de relacionamento textual (*Text Relationship Maps* de Mitra, 1997), usando representação vetorial; a profundidade da sentença na árvore gerada pelo algoritmo de clusterização aglomerativa; dentre outras. Os autores destacam que todas as sentenças são analisadas por um **software de *part-of-speech***, amplamente usado na literatura (BRILL, 1992[23]).

Dias-da-Silva *et al.* (2000) ressaltam que o processo de desenvolvimento do thesaurus eletrônico para o português (TeP) foi realizado em oito etapas: **análise** da forma e do conteúdo **de obras de referência** disponíveis (os mais variados tipos de dicionários do português e inglês, publicados em papel ou disponíveis em meio

[23] Brill, E. A simple rule-based part-of-speech tagger. In Proccedings of the Third Conference on Applied Computational Linguistics.Association for Computational Linguistics. 1992.

digital), com vistas à delimitação do objeto thesaurus e, sobretudo, à utilização dessas obras como fontes de conhecimento lexical; seleção das obras de referência, enquanto fontes garantidas de conhecimento lexical, e estabelecimento de critérios de filtragem da informação lexical extraída; especificação do conteúdo e da forma da base do thesaurus; **implementação de um editor** para a construção dessa base; inserção dos dados na base do thesaurus por linguistas; aplicação de **testes de consistência global** da base e de sua completude relativa às fontes de conhecimento lexical selecionadas e ao léxico do ReGra (NUNES *et al.*, 1996); conversão da base do thesaurus no TeP; análise de questões referentes à apresentação e disseminação do TeP, bem como ao seu modo de integração a outros aplicativos.

Rossi *et al.* (2001) definem as descrições definidas conforme trabalho anterior, desenvolvido para o tratamento da correferência em Língua Inglesa (VIEIRA, 1998[24]). A classificação adotada no presente trabalho distingue os usos que seguem um antecedente textual, daqueles que introduzem novos elementos no discurso. Assim, quatro classes foram definidas: **Anáforas diretas**, que são aquelas antecedidas por uma expressão (definida ou não) que tem o mesmo nome-núcleo e referem-se à mesma entidade no discurso; **Anáforas indiretas**, que não têm o mesmo nome-núcleo do seu antecedente (pode ser um sinônimo do antecedente ou mesmo uma elipse); **Associativas**, que introduzem um referente novo no discurso, mas que tem uma relação semântica com algum antecedente já introduzido; e **Novas no discurso**, que são aquelas que introduzem um novo referente no texto que não se relaciona com nenhum antecedente no discurso, ou seja, não tem uma âncora em que possa se apoiar semanticamente. A fim de auxiliar a análise das descrições definidas, foi desenvolvida uma **interface para a anotação de correferência** em corpus da língua portuguesa, classificação e contabilização dos tipos usados. Através dessa interface, três diferentes sujeitos realizaram o trabalho de anotação de correferência, armazenando e totalizando os resultados alcançados de forma padronizada. Para cada texto selecionado, são exibidas as sentenças numeradas, com as descrições destacadas. O usuário então classifica-a em nova no discurso, anáfora direta, anáfora indireta, associativa ou não classificada. Se a descrição definida for classificada, o usuário deverá informar ainda alguns dados,

[24] Vieira, R. . “Definite description processing in unrestricted text”. Centre for Cognitive Science, Edinburgh University. Edinburgh, UK., 1998. (Dissertação de Doutorado).

tais como o número da sentença onde o antecedente se encontra e, o sintagma nominal antecedente desta descrição. Tendo feito esta classificação manual, os autores apresentam um sistema (em Prolog) para o tratamento automático de correferência nominal, baseado no estudo feito manualmente. A versão apresentada no artigo, por ser um trabalho em andamento, efetua a classificação das descrições como novas no discurso ou anáforas. Assim, para cada sintagma nominal, o sistema compara o seu núcleo com os armazenados na lista de antecedentes potenciais (todos os extraídos do corpus). Se existir, a descrição definida é classificada como anáfora direta. Se, não houver antecendente, procura-se investigar indícios de que ela possa ser uma descrição nova no discurso. Segundo os autores, no sistema desenvolvido anteriormente para a Língua Inglesa (VIEIRA, 1998), a identificação de descrições definidas novas no discurso, é baseada numa série de heurísticas (referentes à estrutura sintática da descrição definida), e que estudos comparativos entre o Inglês e o Português estavam sendo desenvolvidos para adaptar essas heurísticas. Por fim, caso todos os testes falhem, a descrição definida é dada como não classificada. Os textos foram submetidos a uma **análise sintática automática**, através do software interativo do projeto ***Visual Interactive Syntax Learning – VISL*** (http://visl.hum.ou.dk/itwebsite/visl/visltop.html).

Gamallo, Agustini e Lopes (2001) descrevem um **método não supervisionado pobre de conhecimento** para a aquisição de **restrições de seleção**, baseado nas **hipóteses contextual e de co-especificação**. Segundo os autores, na maioria das **abordagens pobres de conhecimento** para aprendizagem de restrição de seleção, o processo de indução e generalização da informação semântica a partir de **frequências de co-ocorrência de palavras** consiste em agrupar automaticamente palavras consideradas similares. Segundo os autores, a melhor estratégia conhecida para se medir a similaridade entre palavras é a baseada na **hipótese de distribuição de Harris**. De acordo com esta teoria, palavras que co-ocorrem em contextos sintáticos semelhantes são semanticamente similares e, portanto, devem ser agrupados em uma mesma classe semântica. No entanto, os métodos de aprendizagem baseados na hipótese de distribuição podem levar, a concentrar numa mesma classe, palavras que atendem a diferentes restrições de seleção. Segundo os autores, para contornar este problema, deve-se extrair classes contextuais de palavras a partir de construções sintáticas adequadas, considerando que **contextos sintáticos similares compartilham as mesmas restrições**

semânticas nas palavras. Os autores destacam que dois contextos sintáticos que ocorrem com (quase) as mesmas palavras são semelhantes e, então as mesmas restrições semânticas sobre essas palavras são aplicadas (chamada de **hipótese contextual**). Estratégias de extração semântica baseada na hipótese contextual podem explicar a variação semântica das palavras em diferentes contextos sintáticos. Uma vez que estas abordagens estão preocupadas com a extração de similaridades semânticas entre os contextos sintáticos, as palavras serão agrupadas de acordo com a sua distribuição sintática. Esses agrupamentos representam classes semânticas dependentes do contexto. O processo de restrição mútua entre duas palavras relacionadas é chamado por Pustejovsky[25] como "co-especificação" ou "co-composição". Co-especificação ocorre quando duas expressões sintaticamente dependentes deixam de ser interpretadas como "predicado-argumento", onde o predicado é a função ativa, que institui as preferências semânticas em um argumento passivo, que corresponde a essas preferências. Pelo contrário, cada palavra de uma dependência binária é considerada simultaneamente como um predicado e um argumento. Ou seja, cada palavra, tanto impõe restrições semânticas como atende a requisitos semânticos. Segundo os autores, para avaliar as hipóteses apresentadas, foi desenvolvido um sistema para realizar aquisição automática de restrições semânticas. O sistema é constituído por quatro módulos: ***Parsing***, onde o texto é anotado por palavra (MARQUES, 2000[26]) e por blocos parcialmente analisados (ROCIO *ET AL.*, 2001[27]). Uma heurística é então utilizada para identificar dependências binárias. O resultado é uma lista de tuplas de co-ocorrência contendo a relação sintática e os lemas das duas palavras relacionadas; **Extração**, quando as dependências binárias são utilizadas para extrair os contextos sintáticos; **Filtragem**, quando cada par de palavras contextuais são comparados estatisticamente usando uma variação da medida de pesos de Jaccard[28]; **Clusterização**, onde classes bases são sucessivamente agregadas pelo método de agrupamento conceitual para induzir a classes mais gerais, que representem as restrições de seleção dos contextos sintáticos.

[25] James Pustejovsky. The Generative Lexicon. MIT Press, Cambridge, 1995.
[26] Nuno Marques. Uma Metodologia para a Modelação Estatística da Subcategorização Verbal. PhD thesis, Universidade Nova de Lisboa, Lisboa, Portugal, 2000.
[27] Rocio, V.; Clergerie, E. de la; and Lopes, J.G.P. . Tabulation for multi-purpose partial parsing. Journal of Grammars, 4(1), 2001.
[28] Gregory Grefenstette. Explorations in Automatic Thesaurus Discovery. Kluwer Academic Publishers, USA, 1994.

Gonzalez e Strube de Lima (2001), com o intuito de viabilizar a **expansão automática de consulta**, utilizaram um **thesaurus (denominado T-Lex),** que possui uma estruturação semântica para implementar relacionamentos lexicais, considerando fundamentos da **Teoria do Léxico Gerativo (TLG)** de Pustejovsky (1995[29]). A TLG introduz um conjunto de recursos para análise semântica de expressões em linguagem natural, incluindo operações gerativas que possibilitam derivações composicionais das palavras dependentes de contexto. Um léxico semântico, de acordo com a TLG, é caracterizado como um sistema computacional onde a estrutura Qualia é um dos níveis de representação, com quatro campos (ou papéis) de descrição. Assim, na representação de um item lexical, o papel Formal distingue-o num amplo domínio, o papel Constitutivo descreve o que faz parte dele, o papel Agentivo especifica como ele passou a existir e o papel Télico explica qual a sua função ou propósito. No T-Lex, cada categoria gramatical possui estruturas Qualia específicas. Assim, os novos termos expandidos nas consultas são obtidos pela utilização de critérios semânticos para compor relacionamentos lexicais, conforme estabelecidos no T-Lex. Segundo os autores, os **termos obtidos** na expansão de uma consulta são selecionados entre os descritores contidos no T-Lex, a partir de **operações gerativas**. Essas operações tem como objetivo compor um campo lexical (ou campo semântico) de um item lexical. As operações utilizadas são (GONZALEZ, 2000[30]): **especialização, co-herança, associação, equivalência, decomposição e agregação**.

Souza, Pereira e Nunes (2001) apresentam um ambiente para testar estratégias de sumarização automática extrativa de português. Segundo os autores, as técnicas utilizadas "consideram a **seleção de sentenças relevantes** a partir da existência de palavras-chave geradas automaticamente, e/ou palavras do título, de sua localização, de palavras sinalizadoras e palavras-chaves fornecidas pelo autor do documento, quando disponíveis" (p. 1). Foram implementados dois **métodos de extração de palavras-chaves** (PEREIRA, 2001[31]): baseado na **frequência de** determinados **padrões morfossintáticos**, e baseado na **frequência de radicais,** baseado no **algoritmo Extractor** (TURNEY, 1999[32]). Os textos sumarizados são

[29] PUSTEJOVSKY, J. The Generative Lexicon. Cambridge: The MIT Press, 1995. 298 p.
[30] GONZALEZ, M. O Léxico Gerativo de Pustejovsky sob o Enfoque da Recuperação de Informações. Trabalho Individual I, PPGCC, Faculdade de Informática, PUCRS, maio 2000. 52 p.
[31] PEREIRA, M. Algoritmos de Extração de Palavras-chaves em Português. NILC-TR-01-06, Setembro, 2001.
[32] TURNEY, Peter (1999). Learning to Extract Keyphrases from Text, Tech. Report Number NRC-41622, National Research Council Canada, Institute for Information Technology

submetidos a dois programas (ALUÍSIO; AIRES, 2000[33]): ***Tokennizer***, que separa as pontuações das palavras, e o ***Tagger***, que etiqueta todas as palavras do texto (classe morfossintática). Várias estratégias de seleção de sentenças foram avaliadas, combinando-se as palavras-chave, as palavras do título, a localização de palavras sinalizadoras (tais como objetivo, resultado, neste artigo, este artigo, dentre outras), e as palavras-chave definidas pelo autor. O sistema implementado utiliza um programa de **extração de radicais** para o português baseado no algoritmo de Porter (PORTER, 1980[34]).

Orengo e Huyck (2001) apresentam um algoritmo composto por oito passos, que devem ser executados em sequência. Dentre os passos implementados estão a redução do plural e do feminino, a redução adverbial, do aumentativo ou diminutivo, a redução nominal e verbal, dentre outros. Cada **passo** tem um conjunto de **regras**, sendo que cada regra contém uma **lista de exceções**. Para avaliar o algoritmo apresentado, utilizou-se o método de Paice[35], que determina o cálculo dos índices de *overstemming,* quando parte do radical é removida pelo algoritmo, e *understemming,* quando o sufixo não é removido completamente.

Jose Neto e Moraes (2002) exploram as propriedades dos autômatos adaptativos para a elaboração de reconhecedores que tratem o não-determinismo e a ambiguidade sintática. Segundo os autores, não-determinismo ocorre quando duas ou mais construções sintáticas apresentam prefixo comum, enquanto que ambiguidades ocorre quando uma sentença tem duas ou mais interpretações válidas. Segundo os autores, adotou-se, para a descrição da sintaxe desta simplificação da linguagem natural, a **notação de Wirth**, uma **meta-linguagem** apropriada para a elaboração de descrições **gramaticais livres de contexto**. Para cada regra da gramática, aplicam-se transformações de substituição, de forma que seja reduzido, a um mínimo, o número de não-terminais presentes na gramática. A cada um desses não-terminais remanescentes corresponderá uma sub-máquina específica. Em cada regra identificam-se todas as construções sintáticas que correspondam a sequências de símbolos que devem figurar obrigatoriamente na sentença (p. 2 e 3). Os autores destacam ainda que o método adotado para a construção de um autômato adaptativo a partir da gramática consiste em desenhar

[33] ALUÍSIO, S.M.; AIRES, R.V. Etiquetação de um Corpus e Construção de um Etiquetador de Português. Relatórios Técnicos do ICMC-USP, 107 (NILC-TR-00-2). Março 2000, 18p.
[34] PORTER, M. F. An algorithm for suffix stripping. Program, 14(3):130--137, 1980.
[35] Paice, C.D. An Evaluation Method for Stemming Algorithms. In: ACM SIGIR Conference on Research and Development in Information Retrieval, 1994, pp. 42-50.

uma máquina de estados inicial que reconheça qualquer cadeia válida de símbolos representada pelo conjunto disponível de regras gramaticais. E complementam que não é usado nenhum método de construção de reconhecedores convencional a partir de gramáticas livres de contexto, mas explora-se a característica adaptativa do modelo de reconhecimento adotado.

Bidarra (2002) apresenta alguns aspectos básicos para **construção de léxicos** computacionais baseados em dados de **parafasia semântica**. Segundo o autor, parafasia é um tipo de erro recorrente na fala de pacientes - problema de acesso lexical – não apenas restrito às substituições de palavras, mas também marcado por erros de escolha de som. Assim, o léxico foi elaborado a partir das chamadas **estruturas de traços tipadas** (CARPENTER, 1992[36]), no contexto das **gramáticas de unificação**, fortemente baseado nas **teorias do léxico gerativo** (PUSTEJOVSKY, 1991[37]) e da HPSG (POLLARD; SAG, 1987[38]). Como **ambiente cognitivo de execução** para o processamento lexical foi considerado o **modelo proposto por Levelt (1992[39]).** Segundo eles, as palavras no léxico seriam estruturadas por meio de traços funcionais, perceptivos e semântico conceituais, alguns genéricos, outros mais especializados, esses últimos também chamados de traços distintivos. Os traços seriam, em tal abordagem, os responsáveis pela identificação e consequente indexação da palavra correta no léxico. No modelo, notam-se cinco processos envolvidos na geração de fala fluente: o módulo conceitualizador, responsável pela geração e monitoração das mensagens emitidas e recebidas pelo falante; o de processamento linguístico, responsável por realizar dois tipos de codificação (gramatical e fonética) nos fragmentos de mensagem oriundos do módulo conceitualizador; os sistemas de articulação e de audição, que produzem a entrada do próximo módulo; e o módulo de compreensão, que fecha o ciclo retornando a informação processada ao conceitualizador, aceitando-a como correta ou rejeitando-a, e corrigindo-a se possível. Segundo o autor, para determinar as condições necessárias para a implementação do léxico (eletrônico), o modelo teórico lexical proposto tem sido confrontado com dados extraídos de seções clínico-terapêuticas de acompanhamento longitudinal de pacientes afásicos, submetidos a

[36] CARPENTER, B. The Logic of Typed Feature Structures. Cambridge University Press: NY, 1992.
[37] PUSTEJOVSKY, J. The Generative Lexicon. Cambridge: The MIT Press, 1995. The Gererative Lexicon. Computational Linguistics, n. 17, p. 409-441, 1991.
[38] POLLARD, C. e SAG, I. A. Information-Based Syntax and Semantics, Vol. 1: Fundamentals. CSLI Lecture Notes n. 13. Center for the Study of Language and Information: Stanford, 1987.
[39] LEVELT, W. J. M. Accessing words in speech production: stages, processes and representations. Cognition, n. 42, p. 1-22, 1992. Speaking: from intention to articulation. Cambridge: MIT Press, 1989.

testes de conversação livre ou de nomeação de objetos.

Pardo e Rino (2002) apresentam o sumarizador **DMSumm (*Discourse Modeling SUMMarizer*),** que possui os três processos clássicos da geração automática de textos: **seleção de conteúdo, planejamento textual e realização linguística**. A mensagem-fonte, que será sumarizada pelo DMSumm, é constituída de três componentes: o **objetivo comunicativo** (objetivo que o escritor pretende alcançar com seu texto), a **proposição central** (a informação principal que o escritor deseja transmitir com a veiculação do texto) e a **base de conhecimento** (conteúdo informativo do texto-fonte). Os autores destacam que na seleção de conteúdo, o DMSumm recebe a mensagem-fonte como entrada e tem a função de diminuir o conteúdo informativo para a produção dos sumários, **podando a base de conhecimento**. São usadas 12 **heurísticas de poda**, definidas em (RINO E SCOTT, 1994[40]), para selecionar o conteúdo relevante; no planejamento textual, a mensagem-fonte reduzida é usada pelo planejador para **produzir as estruturas retóricas** dos possíveis sumários, considerando-se a base de conhecimento já podada e o objetivo comunicativo e a proposição central originais (com base no modelo de discurso de Rino, fazendo um mapeamento de relações semânticas e intencionais nas relações retóricas dos planos de texto, utilizando, para isso, operadores de plano, os quais são um artifício computacional para a montagem da estrutura e do conteúdo textual); por fim, a **realização linguística produz os sumários**, propriamente ditos, a partir dos planos de texto, utilizando, para isso, templates definidos em função de uma gramática e de um léxico de uma determinada língua natural.

Schulz *et al.* (2002) tem como objetivo apresentar uma metodologia que, segundo os autores, abandona os métodos tradicionais de recuperação e se baseia no **uso de morfemas médicos**, como unidades atômicas para indexação e recuperação de informação. Segundo os autores, a ideia central é extrair de um documento somente **informações relevantes** para a busca, as quais geralmente estão contidas nas **raízes das palavras (radicais)** para construir o **índice do documento**, em vez de usar a “superfície” do texto, ou seja, as **palavras originais**. Assim, segundo os autores, todos os documentos seriam submetidos a um processo de normalização morfossemântica antes de serem automaticamente indexados para,

[40] Rino, L.H.M. and Scott, D. (1994). Automatic generation of draft summaries: heuristics for content selection. ITRI Techn. Report ITRI-94-8. University of Brighton, England.

assim, melhorar o desempenho de motores de busca. O **sistema MORPHOSAURUS** faz uso do que os autores chamaram de **bases terminológicas** e de **rotinas de normalização** de textos. Segundo os autores, as bases terminológicas são constituídas por vários repositórios: "*subwords*" classificados como radicais, prefixos e sufixos, nomes próprios, thesaurus (dicionário de sinônimos) e um mapeamento do repositório de *subwords* para o **MeSH (*Medical Subject Headings*)**. Ainda segundo os autores, a terminologia interage com rotinas de normalização de textos: segmentador, que extrai *subwords* do texto para, então, substituí-las pelo seu identificador, usando o repositório de *subwords*; e processador de acrônimos, onde acrônimos e abreviações são identificados e expandidos usando a base de acrônimos.

Bonfante e Nunes (2002) apresentam a implementação de um **parser probabilístico**, seguindo o **modelo de Collins (1999[41])**. O modelo, implementado originalmente para a língua inglesa, baseia-se na **noção de núcleos lexicais**, onde para cada regra observada no conjunto de treinamento, as **palavras** que não são **núcleo** são chamadas de **modificadores**, exercendo influência sobre ele. A formação da estrutura sintática de uma sentença se dá através de um processo *bottom-up* comandado pela **probabilidade de um núcleo e um modificador se juntarem para formar um sintagma[42]**. Utilizou-se um conjunto de sentenças obtidas do corpus NILC e **anotadas** sintaticamente pelo **parser do Bick (2000[43])**.

Zavaglia (2003) destaca que o modelo de representação proposto contém informações semânticas (relações semânticas para resgatar o significado de cada item lexical) e morfossintáticas (classe gramatical, gênero e número das palavras). O modelo proposto não pretende definir, de modo direto, o significado de um item lexical homógrafo, mas somente sugerir o significado para cada item, assim como para suas ocorrências polissêmicas. A **base de conhecimento lexical** construída é constituída de vários componentes: **informação ontológica** (esta ontologia foi construída na tese de Doutorado da própria autora), **informação Qualia** (baseada na **Teoria do Léxico Gerativo de Pustejovsky**), **informação morfossintática**, **informação definicional** (extraída de um dicionário base), e **informação pragmática** (exemplos do uso do item homônimo extraídos de um corpus de 11

[41] Collins, M. J. . Head-Driven Statistical Models for Natural Language Parsing. PhD thesis, University of Pennsylvania, 1999.
[42] No mapa conceitual da seção 4.2.2, esta expressão foi representada como “Probabilidade de ocorrência”.
[43] Bick, E. . The Parsing System. "Palavras" – Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus University Press, 2000.

milhões de palavras do Laboratório de Estudos Lexico-gráficos da Unesp de Araraquara).

Martins, Monard e Matsubara (2003) discutem que tarefas como sumarização e clusterização de documentos estão relacionadas ao processo de mineração de texto (*text mining*), dependendo da natureza da aplicação e das categorias serem conhecidas. Dentre as fases clássicas do processo de mineração de textos estão: a obtenção dos documentos, pré-processamento, extração de conhecimento e avaliação e interpretação dos resultados. Segundo os autores, na fase de **pré-processamento**, os documentos podem ser transformados em um vetor de termos (*bag-of-words*) que ocorrem no documento, ou em representações mais elaboradas (e cita Montes-y-Gómez *et al.*, 2001[44]). No entanto, os autores destacam que experimentos anteriores revelaram que as representações mais sofisticadas apresentaram resultados inferiores. Os termos que compõem o vetor podem ser palavras simples ou compostas (2, 3, …, n-gram) que ocorrem no documento. Cada termo é usada como um atributo do conjunto de dados representado na forma de atributo-valor. O valor atribuído a cada termo pode ser binário (0 ou 1) ou medidas estatísticas que levam em consideração a frequência com o termo aparece tanto no documento, como nos demais. Os termos com alta frequência em todos os documentos, ou pelo menos na maioria, tendem a não apresentarem informação útil para discriminar um documento. Segundo os autores, vários **critérios** podem ser utilizados para reduzir o número de atributos (dimensionalidade) do conjunto de dados. Um método amplamente utilizado, de acordo com os autores, é reduzir cada termo ao seu radical, utilizando **algoritmos de *stemming*** (e cita o algoritmo de Porter, de 1980, para o inglês). Outra maneira de reduzir a dimensionalidade do vetor de termos é **eleger os termos mais representativos** para discriminar os documentos. Neste momento, os autores citam a **lei de Zipf** (Zipf's Law de 1949[45]), como sendo uma alternativa para eliminar os termos que não são representativos numa coleção de documentos. Além disso, os autores complementam citando o **trabalho de Luhn** (1958[46]) que usou esta lei para especificar um limiar mínimo e máximo de corte, para excluir os termos irrelevantes e o trabalho de Van Rijsbergen

[44] Montes-y-Gómez, M., A. Gelbukh, A. López-López, & R. Baeza-Yates (2001). Flexible comparison of conceptual graphs. In H. Mayr, J. Lazansky, G. Quirchmayr, & P. Vogel (Eds.), Proc. DEXA-2001, 12th International Conference and Workshop on Database and Expert Systems Applications. LNCS 2113, pp. 102–111. Springer-Verlag.
[45] Zipf, G. (1949). Human Behaviour and the Principle of Least Effort. Addison-Wesley.
[46] Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal of Research and Development 2 (2), 159–165.

(1979[47]) que afirma que a correta determinação desse limiar é obtida por um processo de tentativa e erro. Utilizaram três **algoritmos de aprendizado**: dois algoritmos simbólicos de **regras de associação**; e um baseado em técnicas estatísticas de aprendizado – ***Support Vector Machines (SVM)*** para ilustrar o uso de redução de dimensionalidade do conjunto de treinamento. Os autores apresentam a **ferramenta PreText**, desenvolvida com o objetivo de realizar automaticamente a tarefa de pré-processamento de um coleção de documentos, e inclui a funcionalidade de reduzir a dimensionalidade do conjunto de dados usando a lei de Zipf e os limiares de Luhn. A ferramenta permite considerar somente os radicais dos termos, com frequências entre um intervalo definido pelo usuário ou pelo limiar de Luhn. As configurações padrão do PreText é usar a frequência do termo, não reduzir ao radical e usar como termos, 1, 2 e 3-grams. De uma maneira geral, a PreText funciona da seguinte maneira: extrai os radicais de palavras em português, inglês ou espanhol; elimina as *stopwords*; contabiliza a frequência de cada radical no documento e na coleção; atribui um peso aos termos (binária, frequência absoluta ou *tfidf*); aplica a lei de Zipf e os limiares de Luhn; define usar palavras simples ou compostas (n-grams); apresenta gráficos da frequência dos radicais; e salva a tabela de atributos-valor em um arquivo (para ser usado, por exemplo, como entrada de um outro projeto da equipe do LABIC/USP – Discover).

Pardo, Rino e Nunes (2003) apresentam o **NeuralSumm**, sumarizador que utiliza uma **rede neural *SelfOrganizing (SOM)*** para classificar cada sentença de um texto-fonte de acordo com seu grau de importância. A rede Som foi comparada com os **algoritmos de classificação Naive-Bayes** e o de **árvore de decisão C4.5**. Considerou-se que as **sentenças** podem ser classificadas como **essenciais** (ideia principal de um texto), **complementares** (acrescentam conteúdo a ideia principal) ou **supérfluas** (não acrescentam conteúdo algum). As **sentenças essenciais** devem sempre estar no extrato **[sempre são incluídas]**, **as complementares podem** ou não **ser incluídas** enquanto que as **supérfluas** devem ser descartadas **[não são incluídas]**. O extrato produzido pelo NeuralSumm considera a classificação obtida pelas sentenças, a taxa de compressão desejada e a pontuação das sentenças dada pela distribuição de palavras do texto (uma das características consideradas pela rede neural). As **características (*features*)** utilizadas são: o tamanho da sentença; a posição da sentença no texto; a posição da

[47] Van Rijsbergen, C. J. (1979). Information Retrieval, 2nd edition. Dept. of Computer Science, Univ. of Glasgow.

sentença no parágrafo a que pertence; a presença de palavras-chave na sentença; a presença de palavras da “gist sentence” na sentença[48], isto é, a sentença que melhor expressa a ideia principal do texto (PARDO, 2002[49]).; a pontuação da sentença com base na distribuição das palavras do texto; o TF-ISF da sentença; a presença de palavras indicativas na sentença. Os autores destacam que as *features* utilizadas para classificação das sentenças são, quase na totalidade, independentes de língua, gênero textual e domínio.

Gasperin e Strube de Lima (2003) destacam que a técnica desenvolvida na dissertação de mestrado (GASPERIN, 2001[50]), para criar listas de palavras relacionadas semanticamente, inclui três passos: **extrair o contexto sintático de cada palavra** do corpus; **comparar cada par de palavras usando seus contextos sintáticos por meio de uma medida de similaridade[51]**; e construir **listas com as palavras mais similares** para cada nome (termo) do corpus. Segundo as autoras, para encontrar a similaridade entre as palavras, seus contextos sintáticos são comparados usando a **medida binária de Jaccard** (1994[52]). Neste trabalho, gerou-se listas somente para os nomes e adjetivos, sendo que os verbos não foram considerados. Utilizou-se uma **ferramenta para expansão de consulta – QET** (PIZZATO, 2002[53]) que, comparando várias listas, utiliza grafos e um processo iterativo, para retornar pesos para cada palavra, em função do caminhamento adotado na consulta (GASPERIN; STRUBE DE LIMA, 2003, p. 4).

Oliveira, Garrao e Amaral (2003) propuseram um **conjunto de critérios** que devem ser aplicados às expressões para a **detecção de locuções preposicionais**. No entanto, os autores ressaltam que é possível que uma locução seja positiva para um dado teste e negativo para outro, ou seja, que os testes não sejam nem necessário, nem tão pouco suficientes para determinar a implementação de uma dada expressão. O primeiro critério definido consiste em **reconhecer uma locução** preposicional como uma expressão **fixa e não ambígua** (*unambiguously frozen expression*), o que significa que a sequência de lexemas em questão é sempre interpretada como uma locução preposicional. O segundo critério definido é

[48] Segundo os autores, a gist sentence de um texto é determinada pela aplicação do método GistKey, que também utiliza a distribuição de palavras no texto.
[49] Pardo, T.A.S. (2002a). GistSumm: Um Sumarizador Automático Baseado na Idéia Principal de Textos. Série de Relatórios do NILC. NILC-TR-02-13. Available for download in www.nilc.icmc.usp.br/nilc/~thiago
[50] Gasperin, C. V. . Extração Automática de relações semânticas a partir de relações sintáticas. Master's thesis, PPGCC – PUCRS, Porto Alegre (2001).
[51] No mapa conceitual da seção 4.2.2, adotou-se a expressão “Comparar usando medidas de similaridades”.
[52] Gregory Grefenstette. Explorations in Automatic Thesaurus Discovery. Kluwer Acad. Publishers, USA, 1994.
[53] Pizzato, L. A. . Estrutura multitesauro para recuperação de informações. Master's thesis, PPGCC – PUCRS, Porto Alegre (2002).

o da substituição, baseado na noção de que uma locução preposicional normalmente pode ser **substituída por uma preposição simples** ou por outra locução preposicional. O terceiro critério é o que identifica se a preposição **existe em função de um verbo precedente**, ou seja, se ocorre anteriormente à expressão. E o quarto e último critério é se **existe a possibilidade de inserir um determinante** (artigo definido).

Alves e Chishman (2004) apresentam uma visão crítica sobre os tradutores automáticos e o fenômeno da ambiguidade, procurando mostrar como os tradutores automáticos processam esse complexo fenômeno linguístico. Inicialmente, as autoras fizeram uma apresentação dos tradutores automáticos, e em seguida propuseram uma reorganização das nomenclaturas envolvidas no tratamento da ambiguidade. Finalmente, as autoras compararam o desempenho de quatro tradutores automáticos: **Systran(SYS) e Free Translator (FTR) – disponíveis na Web livremente, L&H Power Translator Pro (PTP) e Micro Power Delta Translator 2.0 (DT) – comercializados.** Segundo as autoras, os tradutores foram avaliados de acordo com sua capacidade de tradução de casos ambíguos utilizando como língua fonte o Português e como língua alvo, o Inglês.

Specia e Nunes (2004) procuraram identificar os **casos mais problemáticos** de ambiguidade e assim delimitar a proposta do modelo de desambiguação lexical de sentido. Essa pesquisa apresentada com detalhes em Specia and Nunes (2004[54]), consistiu de um experimento com o córpus BNC (*British National Corpus*) (BURNARD, 2000[55]) utilizando três **sistemas de tradução automática** inglês-português: **Systran, FreeTranslation e Globalink Power Translator Pro**. Foram submetidas aos tradutores, as sentenças que continham os 15 verbos mais frequentes do BNC. As traduções foram, então, manualmente analisadas para verificar a ocorrência da ambiguidade, seus efeitos na tradução das sentenças e o comportamento dos sistemas diante desse fenômeno. Desse estudo, foram selecionados sete verbos (*to go, to get, to make, to take, to come, to look e to give*) que foram considerados os mais problemáticos, por terem sido usados em sentenças que não foram traduzidas corretamente por nenhum tradutor (75% do total). As autoras destacam que em uma etapa de pré-processamento do corpus de exemplos, podem ser levantadas informações linguísticas (ou extralinguísticas)

[54] Specia, L. and Nunes, M.G.V. (2004) “A ambiguidade lexical de sentido na tradução do inglês para o português – um recorte de verbos problemáticos”, Série de Relatórios do NILC, NILC-TR-04-01, São Carlos, Março, 30p.
[55] Burnard, L. (2000) “Reference Guide for the British National Corpus (World Edition)”, Oxford University Press.

sobre a sentença, como a categoria gramatical das palavras, o lema do verbo e a tradução (sentido) desse verbo (classe de cada exemplo, ou seja, a característica que deve ser aprendida). Para submeter esse corpus a um **algoritmo de aprendizado de máquina**, é preciso indicar que **características (*features*)** das sentenças serão consideradas no processo de aprendizado para a geração de regras para classificar novos casos de ambiguidade. Segundo as autoras, quanto mais informações sobre o uso da palavra ambígua (*features* significativas) são fornecidas ao algoritmo, maiores são as chances de o conhecimento obtido ser consistente e útil. A partir da estrutura atributo-valor gerada pode-se utilizar um algoritmo de aprendizado de máquina simbólico convencional para gerar um modelo preditivo para classificar, isto é, identificar o sentido (classe) dos verbos em novas sentenças. Como formalismo de representação proposicional, as autoras optaram por utilizar a **Programação Lógica Indutiva** para o aprendizado baseado em exemplos, por fazer uso de um formalismo relacional, cujo poder de expressividade é, segundo as autoras, similar ao da Lógica de Primeira Ordem. Assim, com base nos exemplos e no conhecimento representado de maneira relacional, bem como nas restrições sobre o modelo a ser gerado, **regras seria geradas automaticamente**[56] que relacionem diversos tipos de conhecimento e de informações das sentenças de exemplo. Segundo as autoras, por serem simbólicas, tais regras podem ser facilmente compreendidas e, com isso, o modelo pode ser manualmente ajustado, se necessário.

Rino *et al.* (2004) compararam o desempenho de cinco sistemas de sumarização **[sumarizadores (extratos)] automática** encontrados na literatura: **GistSumm, TF-ISF-Summ, NeuralSumm, ClassSumm e SuPor**. Utilizou-se 100 textos de jornais brasileiros disponibilizados no corpus TeMário, para os quais, os sumários foram produzidos manualmente por consultores da língua portuguesa. Não houve comparação dos sumários gerados pelos sistemas com os produzidos por especialistas da língua portuguesa no corpus TeMário, visto que não são extratos. Para efeitos de comparação, foram usadas as medidas precisão, cobertura (*recall*) e f-measure[57].

Aluisio *et al.* (2004) apresentam o **Lácio-Web** como sendo um **repositório de recursos** para o desenvolvimento de pesquisas da língua

[56] No mapa conceitual da seção 4.2.2, foi incluída a expressão “Regras de Identificação de sentido”.
[57] Para calcular tais medidas, os sumários ideais (extratos dos sumários manuais) foram gerados utilizando-se http://www.nilc.icmc.usp.br/~thiago.

portuguesa do Brasil **e de ferramentas** linguísticas e computacionais. Segundo os autores, o Lácio-Web integra **diferentes tipos de corporas** disponíveis, que podem ser acessíveis por usuários especialistas e leigos, e por este motivo, uma interface foi desenvolvida para permitir que os usuários exponham os seus objetivos. Segundo os autores, o Lácio-Web é composto por: um **corpus de referência**, chamado **Lácio-Ref**; parte do Lácio-Ref foi manualmente validado por *tags* morfosintáticas, chamado Mac-Morpho; parte do Lácio-Ref foi também anotado automaticamente com lemas, partes do discurso (POS) e *tags* sintáticas; por **textos não anotados**, chamado de **Lácio-Dev**; corporas de **textos Português-Inglês**, chamados **Par-C** (**para alinhamento** de sentenças e palavras) e **Comp-C** (para métodos de **extração de termos**). Além disso, os autores discutem sobre os dados mantidos no corpus de referência Lácio-Ref, e sobre os requisitos necessários para que um texto seja incluído no corpora.

Matsubara, Monard e Batista (2004) tem como objetivo propor avaliar a utilização de duas ou mais **descrições dos dados** em **algoritmos multi-visão**. Cada descrição é composta por todos os termos constituídos por um determinado número de palavras, ou seja, por todos os n-gram (com n fixo) formados a partir do texto. Neste trabalho, avaliou-se utilizar **unigrama** para a **primeira visão**, e **bigrama** para a **segunda visão**. Segundo os autores, no Laboratório de Inteligência Computacional – LABIC[58] foi desenvolvida a **ferramenta computacional PreTexT**[59] que tem como objetivo realizar o **pré-processamento de textos** utilizando a abordagem ***bag-of-words*.** Foram escolhidas duas bases de textos para realizar os experimentos. Para ambas as bases, utilizou-se a lista de *stopwords* padrão do PreTexT e os termos foram transformados em *stems*, removendo prefixos ou sufixos de um termo, ou mesmo transformando um verbo para sua forma no infinitivo. Para representar os textos foi utilizada a medida *tf*, que representa o valor de cada *stem* (termo ou atributo) do documento como o número de vezes que o termo aparece no documento. Para cada base, foram criadas as duas descrições (usando 1-gram e 2-grams) para aplicar o algoritmo *co-training.*

Pardo, Marcu e Nunes (2005) apresentam o aprendizado de **estruturas argumentais dos verbos** modelado probabilisticamente com base no **modelo**

[58] http://labic.icmc.usp.br
[59] Edson Takashi Matsubara, Claudia Aparecida Martins, and Maria Carolina Monard. Pretext: Uma ferramenta para pré-processamento de textos utilizando a abordagem bag-ofwords. Technical Report 209, ICMCUSP, 2003. ftp://ftp.icmc.sc.usp.br/pub/BIBLIOTECA/rel_tec/RT_209.zip.

noisy-channel **(SHANNON, 1948**[60]), considerado um modelo gerativo pelo fato de se basear em uma história gerativa de como os dados são produzidos ou transformados. Neste trabalho, assume-se que **sentenças** em língua natural são produzidas por um **processo gerativo estocástico**, onde o verbo da sentença é escolhido, assim como os seus argumentos, com certas probabilidades. O **algoritmo *Expectation-Maximization* (EM)** (DEMPSTER *et al*., 1977[61]) é então usado para estimar os parâmetros do modelo, ou seja, estimar estas probabilidades (que são inicializadas uniformemente).

Caseli, Nunes e Forcada (2005) apresentam o método ***Language-Independent Heuristics Lexical Aligner (LIHLA)*** que usa **alinhamentos estatísticos**, e pela utilização de **heurísticas** independente de linguagem, objetivando encontrar o melhor alinhamento entre palavras e sentenças. **Dois léxicos bilíngues** gerados a partir da **ferramenta NATools** foram utilizados. Para gerar este léxico, os textos paralelos devem ser alinhados. Neste trabalho, utilizou-se o ***Translation Corpus Aligner -TCA***[62], mas segundo os autores, qualquer outro método de alinhamento poderia ser utilizado. Dado dois arquivos de sentenças alinhadas, o alinhador de palavras NATools contabiliza as co-ocorrências das palavras em todos os pares de sentenças alinhadas e constrói uma **matriz esparsa de probabilidades de palavra-palavra**, usando o **algoritmo** iterativo **Expectation-maximization**. Finalmente, os elementos com maiores valores na matriz são escolhidos para compor os dois léxicos bilíngue probabilísticos (fonte-alvo e alvo-fonte). Para cada palavra do corpus, são dados também o número de ocorrências no corpus (frequência absoluta) e as traduções mais prováveis, juntamente com as probabilidades.

Specia, Nunes e Stevenson (2005) apresentam como objetivo obter e **avaliar as regras [de identificação de sentido]** geradas, e afirmam que utilizou-se os mesmos sete verbos ambíguos e as mesmas **características (*features)*** como fonte de conhecimento (utilizados no trabalho anterior). Para produzir as regras, utilizou-se o **algoritmo de árvore de decisão C4.5**, considerando cada ramo como sendo uma regra, e usando a implementação original do **sistema Sniffer** (BATISTA;

[60] Shannon, C. (1948). A mathematical theory of communication. Bell System Technical Journal, Vol. 27, N. 3, pp. 379-423.
[61] Dempster, A.P.; Laird N.M.; Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, Ser B, Vol. 39, pp. 1-38.
[62] Hofland, K. (1996). A program for aligning English and Norwegian sentences. In Hockey, S., Ide, N., and Perissinotto, G., editors, Research in Humanities Computing, pages 165–178, Oxford. Oxford University Press.

MONARD, 2004[63]), como parte do **ambiente de mineração de dados Discover** (PRATI *et al*., 2003[64]). As regras produzidas foram analisadas usando critérios subjetivos (manualmente) e medidas objetivas tais como erro (versão negativa da precisão), cobertura (*coverage*), suporte (mesmo que revocação) e novelty (relação entre a premissa e a conclusão da regra), fornecidas pelo **sistema Rulee** (PAULA, 2003[65]), que também faz parte do ambiente Discover.

Silva, Vieira e Osorio (2005) afirmam que**,** o conhecimento linguístico utilizado foi produzido a partir dos resultados obtidos pelo **analisador sintático Palavras (BICK, 2000**[66]) juntamente com o **Palavras Xtractor** (GASPERIN *et al.,* 2003[67]). Sete experimentos foram realizados com as seguintes combinações de termos: somente os nomes, nomes e adjetivos, nome e nomes próprios, nomes, adjetivos e nomes próprios, adjetivos e nomes próprios, somente verbos, e verbos e nomes. As palavras irrelevantes foram eliminadas com base na lista de *stopwords* de Paulo Quaresma (da Universidade de Évora), contendo 476 termos. A análise de radicais foi feita baseado no **algoritmo de Porter**[68]. A frequência relativa foi usada para obter os termos relevantes. O modelo espaço vetorial foi usado para representar os documentos. Utilizou-se como **algoritmos de aprendizado árvore de decisão e redes neurais** para a tarefa de **classificação**, e o ***k-means***, para a **clusterização**. Todas as implementações foram obtidas no **Weka**.

Piltcher *et al.* (2005) destacam que o corretor ortográfico foi desenvolvido para fazer a mediação entre a interface (*chat*) e o módulo que identifica o assunto da mensagem (utilizando uma ontologia de domínio) do sistema de recomendação. As palavras das mensagens são enviadas para o corretor separadamente, e a função de similaridade é calculada. Caso uma correção seja sinalizada, de acordo com um limiar, uma nova palavra (corrigida) é repassada. A **função de similaridade** implementada no presente trabalho é um híbrido das **métricas Levenshtein** (LEVENSHTEIN, 1966[69]), **Metaphone** (substituição de um caractere por outro de

[63] Batista, G.E.A.P.A. and Monard, M.C. (2004) “Sniffer: um Ambiente Computacional para Gerenciamento de Experimentos de Aprendizado de Máquina Supervisionado”. In: Proceedings of the I WorkComp Sul, Florianópolis.
[64] Prati, R.C, Geronimi, M.R., and Monard, M.C. (2003) “An Integrated Environment for Data Mining”. In: Proceedings of the IV Congress of Logic Applied to Technology (LAPTEC-2003), Marília.
[65] Paula, M.F. (2003) “Ambiente para exploração de regras”. Dissertação de Mestrado em Ciência da Computação. Instituto de Ciências Matemáticas e de Computação, USP, São Carlos.
[66] Bick, E. . The Parsing System. "Palavras" – Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus University Press, 2000.
[67] Gasperin et al., Extrating XML Syntactic Chunks from Portuguese Corpora. In: Proceeding of the Workshop TALN 2003 Natural Language Processing og Minority Language and Small Languages – France June 11 – 14 (2003).
[68] Disponível em http://snowball.sourceforge.net.
[69] Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions and reversals. Soviet Physics Doklady, 10(8).

som igual ou semelhante) e **Soundex** (KNUTH, 1973[70]). Em outras palavras, calcula-se o custo da distância de Levenshtein entre a palavra analisada e a palavra do banco; calcula-se o custo da distância de Levenshtein entre a Metaphone Key de ambas as palavras; e finalmente, obtém-se a diferença de bits entre a Soundex Key gerada a partir de cada palavra. Para reduzir o tempo de processamento, a função de similaridade era calculada somente para as palavras que tinham a mesma inicial da procurada.

Rino e Seno (2006) propõem a implementação do protótipo **RHeSumaRST (Regras Heurísticas de Sumarização de estruturas RST**), baseado nos modelos de estruturação retórica do discurso da ***Rhetorical Structure Theory* (RST)** (MANN & THOMPSON, 1987[71]) e de **coerência global do discurso da Teoria de Veins (VT)** (CRISTEA *et al*., 1998[72]). Segundo as autoras, o problema do RHeSumaRST é **evitar quebras de cadeias de co-referências** nos sumários. Assim, **heurísticas** são escolhidas para identificar informações supérfluas em uma estrutura retórica (ou estrutura RST) de um texto-fonte, e garantir que sua exclusão não prejudicará a recuperação de possíveis elos co-referenciais. Assim, um antecedente só será incluído diante da inclusão do seu referente. O sistema admite como entrada somente a estrutura RST do texto-fonte a ser sumarizado, e como saída, ou a estrutura RST do sumário ou um texto cuja realização linguística é elementar. Em outras palavras, as heurísticas do RHeSumaRST se baseiam em duas hipóteses principais, associando a RST à VT, respectivamente: (a) os satélites de relações RST podem ser supérfluos e, portanto, excluídos de uma estrutura RST de um sumário; (b) os satélites que contêm antecedentes de termos anafóricos já inclusos na estrutura de um sumário não podem ser excluídos. Foram utilizados dois outros sumarizadores automáticos: o modelo de saliência de Marcu (1997[73]) e um *baseline*, cujos sumários são construídos pela poda de todo satélite das estruturas RST.

Caseli e Nunes (2006) tem como objetivo **induzir** conhecimento linguístico para tradução – **regras de transferência [processo de indução de regras]** e dicionário bilíngue – para o português do Brasil. O corpus utilizado foi pré-

[70] Knuth, D. (1973). The Art of Computer Programming. Addison-Wesley, 3 edition.
[71] MANN, W.C.; THOMPSON, S.A. Rhetorical Structure Theory: A Theory of Text Organization. Technical Report ISI/RS-87-190, 1987.
[72] CRISTEA, D.; IDE, N.; ROMARY, L. Veins Theory: A Model of Global Discourse Cohesion and Coherence. In the Proceedings of the Coling/ACL'1998, pp. 281-285. Montreal, Canada, 1998.
[73] MARCU, D. 1997. The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts. PhD Thesis, Department of Computer Science, University of Toronto.

processado fazendo o alinhamento das sentenças, etiquetando o discurso e alinhando as palavras. As sentenças dos artigos do corpusFAPESP foram automaticamente alinhadas usando a **ferramenta *Translation Corpus Aligner (TCAalign[74])***; os pares de sentenças alinhadas também foram **etiquetadas** usando **ferramentas** disponíveis **em Apertium[75]**. Finalmente, os exemplos de tradução tiveram suas palavras alinhadas usando o **alinhador lexical LIHLA[76]**, proposto pelas autoras (e analisado posteriormente). O processo de indução de regras de transferência do projeto ReTraTos é dividido em três passos: **identificação de padrões**, baseada nos **algoritmos *Sequential Pattern Mining (SPM)* e PrefixSpan[77]**; **geração de regras**, criando restrições[78] entre os valores das características de um dos lados do padrão identificado, e generalizando essas restrições; e finalmente, a **seleção das regras**. De acordo com as autoras, além das regras de transferência, um dicionário bilíngue é induzido baseado nos alinhamentos observados nos exemplos de tradução: criação de um dicionário bilíngue para cada direção de tradução (da fonte para o alvo e do alvo para a fonte); união dos dois dicionários; generalização das entradas do dicionário; tratamento das diferenças sintáticas quando o valor do gênero ou número tiver sido determinado; e tratamento das sentenças.

Balage Filho *et al.* (2006) avaliam a utilização do **sistema de sumarização GistSumm**, desenvolvido por parte dos autores (PARDO *et al.,* 2003[79]), na tarefa de responder perguntas em um determinado idioma. Segundo os autores, o GistSumm é um sumarizador automático que utiliza o método extrativo **[extratos]** baseado em gist (ideia principal do texto). Em outras palavras, o sumarizador assume que por meio de estatísticas simples é possível identificar a sentença gist (*gist sentence*), e a partir dela construir o sumário. O sumarizador GistSumm é composto por três etapas: segmentação do texto; ranking das

[74] Hofland, K.: A program for aligning English and Norwegian sentences In Hockey, S., Ide, N., Perissinotto, G. (eds) Research in Humanities Computing, Oxford, Oxford University Press (1996) 165–178.
[75] Disponível em http://www.apertium.org.
[76] Caseli, H. M., Nunes, M. G. V., Forcada, M. L.: Evaluating the LIHLA lexical aligner on Spanish, Brazilian Portuguese and Basque parallel texts In Proceedings of the XXI Congreso de la Sociedad Espaola para el Procesamiento del Lenguaje Natural (SEPLN) (2005) 1–8
[77] Pei, J., Han, J., Mortazavi-Asl, B., Wang, J., Pinto, H., Chen, Q., Dayal, U., Hsu, M.: Mining Sequential Patterns by Pattern-Growth: The PrefixSpan Approach IEEE Transactions on Knowledge and Data Engineering 16(10) (2004) 1–17
[78] Carbonell, J., Probst, K., Peterson, E., Monson, C., Lavie, A., Brown, R., Levin, L.: Automatic Rule Learning for Resource-Limited MT In Proceedings of the 5th Conference of the Association for Machine Translation in the Americas (AMTA 2002) (2002) 1–10.
[79] Pardo, T.A.S., Rino, L.H.M., Nunes, M.G.V.: GistSumm: A Summarization Tool Based on a New Extractive Method. In: Mamede, N.J., Baptista, J., Trancoso, I., Nunes, M.G.V. (eds.) PROPOR 2003. LNCS, vol. 2721, pp. 210–218. Springer, Heidelberg (2003)

sentenças, baseado em Luhn (1958[80]), que soma a frequência de cada palavra da sentença no documento todo; e produção do extrato. Uma nova versão do GistSumm[81] permite gerar um extrato a partir de um tópico definido pelo usuário. Durante a participação no **CLEF 2006**, **dois experimentos** foram executados: no primeiro, **a pergunta era usada como tópico**, e o sistema retornava as sentenças gist de maior avaliação; e no segundo experimento, foi desenvolvido um filtro **(heurística) para encontrar [a resposta]**, dentre as sentenças retornadas pelo GistSumm, aquela que fosse mais apropriada. Tendo como base as palavras que compõem as sentenças, a heurística tenta identificar o tipo da pergunta feita, e consequentemente, o tipo da resposta desejada, para, então, eleger a mais adequada.

Enembreck *et al.* (2006) compararam a descrição de um projeto de pesquisa e desenvolvimento com o perfil dos possíveis candidatos. Estes perfis foram gerados automaticamente a partir dos itens de conhecimento presente nos currículos da **plataforma Lattes**, tais como artigos e textos escritos pelos candidatos. Utilizou-se o **TFIDF** para atribuir valores às **características** dos currículos e abordagem baseada em centróide para calcular a similaridade entre os perfis e a descrição do projeto. Para normalizar os perfis e identificar os termos que melhor discriminam os candidatos, utilizou-se o índice Gini[82].

Leite e Rino (2006) analisaram a contribuição das características implementadas no sistema Supor no processo de sumarização automática. O sumarizador Supor utiliza um corpus de treinamento composto dos textos fonte e dos extratos ideais. Todas as sentenças do texto fonte são representadas no conjunto de treinamento como um tupla de valores binários para as características, selecionadas pelo usuário. Cada tupla é rotulada com a classe que sinaliza se a sentença está ou não no sumário (tendo como base o sumário ideal). Este conjunto de treinamento é usado para calcular as probabilidades usadas no classificador Bayesiano para ranquear as sentenças que deverão compor o sumário. Assim, as melhores sentenças são incluídas no sumário, tendo em vista o fator de compressão definido pelo usuário. Dentre as **características** implementadas pelo **Supor**, que

[80] Luhn, H.: The automatic creation of literature abstracts. IBM Journal of Res. and Develop. , 159–165 (1958)
[81] Balage Filho, P.P., Uzêda, V.R., Pardo, T.A.S., Nunes, M.G.V.: Estrutura Textual e Multiplicidade de Tópicos na Sumarização Automática: o Caso do Sistema GistSumm. Technical Report 283. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo (2006).
[82] S. Shankar, and G. Karypis, "A Feature Weight Adjustment Algorithm for Document Categorization", KDD-2000 Workshop on Text Mining, Boston, USA, August 2000.

podem ser usadas em separado ou em combinação, estão o método da cadeia lexical (*lexical chains method*[83]*),* que verifica a existência de palavras relacionadas (por exemplo, sinônimos/antônimos ou hipônimos/hiperônimos); o método do mapa de relacionamento textual (*text relationship map method*[84]*)*, semelhante ao método anterior, mas considera parágrafos ao invés de sentenças, e constrói um grafo chamado de mapa de relacionamento do texto fonte que representa o seu grau de coesão; o método da importância dos tópicos (*importance of topics method*[85]), que tem como objetivo identificar os principais tópicos do texto fonte, que deverão orientar a seleção de sentenças; o tamanho e a localização da sentença; a presença de nomes próprios; e a frequência das palavras. Segundo os autores, com o intuito de melhorar a expressividade desse conjunto de características, os autores assumiram **valores categóricos e numéricos** e não mais binários, como na versão original do Supor. Como **algoritmos de classificação**, foram utilizados o **modelo Naive-Bayes** e o algoritmo de **árvore de decisão C4.5**, ambos usando a **ferramenta Weka**. Os dados precisaram ser discretizados, visto que estes algoritmos não trabalham com valores numéricos.

Moraes e Strube de Lima (2007) utilizaram a hierarquia definida por Langie (2004), que estrutura as categorias na forma de uma árvore, analisando um volume bem menor de documentos. O categorizador hierárquico desenvolvido é formado por vários classificadores multicategoriais, que implementam o **algoritmo *k-Nearest Neighbor (k-NN).*** No presente trabalho, os **textos** utilizados da PLN-BR CATEG foram lematizados, ou seja, os verbos foram colocados no infinitivo e substantivos na forma masculina singular, pelas **ferramentas CHAMA e FORMA**, desenvolvidas por Gonzalez *et al*. (2006[86]). A preparação dos textos foi feita por classificador e, consistiu em identificar os termos (únicos), uma vez que já estavam lematizados, selecionar os mais relevantes e remover as *stopwords*. Os documentos foram representados no formato de **bag-of-words**, e a técnica *tfidf* foi utilizada para calcular os pesos dos termos. Os classificadores implementam o algoritmo k-NN, que, dado um texto do conjunto de documentos de teste (não rotulados), encontra os

[83] Barzilay, R., Elhadad, M.: Using Lexical Chains for Text Summarization. In: Mani, I., Maybury, M. T. (eds.): Advances in Automatic Text Summarization. MIT Press (1997) 111-121.
[84] Salton, G., Singhal, A., Mitra, M., Buckley, C.: Automatic Text Structuring and Summarization. Information Processing & Management 33 (1997) 193-207.
[85] Larocca Neto, J., Santos, A.D., Kaestner, A.A., Freitas, A.A.: Generating Text Summaries through the Relative Importance of Topics. In: Monard, M. C., Sichman, J. S. (eds.): Proc. Of 15th Brazilian Symposium on Artificial Intelligence (SBIA'00). Lecture Notes in Artificial Intelligence, No. 1952, Springer-Verlag (2000) 300-309.
[86] Gonzalez, M., Lima, V.L.S. e Lima, J.V. (2006) "Tools for Nominalization: an Alternative for Lexical Normalization", In: Workshop on Comp. Proc. Of Portuguese Lang – Written and Spoken, 7; PROPOR, 2006, Springer-Verlag, p.100-109.

k documentos vizinhos a esse texto que pertencem ao conjunto de treino, de acordo com uma métrica (co-seno) que avalia a similaridade entre os termos dos documentos.

Kinoshita *et al*. (2007) apresentam o **CoGrOO – Corretor Gramatical para o OpenOffice,** que visa fazer a correção gramatical de diversos erros, tais como concordância verbal e nominal, crase, uso indevido dos adjetivos 'mau' e 'mal', dentre outros encontrados no português do Brasil. De acordo com os autores, duas características do CoGrOO devem ser ressaltada: arquitetura híbrida e a utilização de regras e estatísticas. Neste artigo, os autores apresentam as diferenças entre as duas versões disponibilizadas para o corretor, e comparam o desempenho CoGrOO, com o corretor gramatical do editor de texto Microsoft Word, chamado ReGra (NUNES E OLIVEIRA, 2000[87]). O CoGrOO é composto pelos seguintes módulos: detector de **delimitador** (*boundary*) **de sentenças**; **"tokenizador"**, que separa as sentenças em palavras; **identificador de** potenciais **nomes próprios**; **etiquetador de discurso** (*part-of-speech tagger*); *chunker*, que identifica os sintagmas nominais e verbais; identificador do sujeito e do verbo da frase; e finalmente o detector de erros gramaticais. Para as etapas, onde conhecimento linguístico é exigido, utiliza-se um dicionário criado a partir do **corpus anotado CETENFOLHA**. O etiquetador utilizado foi similar ao de Brill (1992[88]). De acordo com os autores, utilizou-se como referência a biblioteca OpenNLP[89], que como foi desenvolvida originalmente para o inglês, precisou passar por algumas modificações.

Specia, Stevenson e Nunes (2007) apresentam uma **abordagem híbrida** de desambiguação lexical de sentido que utiliza um **formalismo de representação** expressivo, **fontes de conhecimento de fundo (cenário)** e compartilhado, e **a programação lógica indutiva** como técnica de aprendizado. A programação lógica indutiva utiliza técnicas de aprendizado de máquina e programação lógica para construir teorias de primeira ordem a partir de conhecimento de fundo e de exemplos, os quais são representados usando cláusulas de primeira ordem. Utilizou-se o **sistema Aleph ILP** (SRINIVASAN, 2000), que provê um sistema completo de inferência e pode ser customizado de várias

[87] M.G.V. Nunes, O.N. Oliveira Jr., "O processo de desenvolvimento do Revisor Gramatical ReGra", Proc. of XXVII SEMISH, Vol. 1, p. 6, Curitiba, Brazil, 2000.
[88] E. Brill, "A Simple Rule-Based Part Of Speech Tagger", Proceedings of ANLP-92, 3rd Conference of Applied Natural Language Processing, Trento, Italy, 1992.
[89] OpenNLP, an open-source framework to develop natural language applications http://opennlp.sourceforge.net

maneiras. Os corpus utilizados foram lematizados, usando **Minipar** (LIN, 1993[90]), e etiquetados (*part-of-speech*), usando **Mxpost** (RATNAPARKHI, 1996[91]). Como fonte de conhecimento de fundo usada nos algoritmos de aprendizado, os autores experimentaram 12 alternativas diferentes de **características**, contendo *bag-of-words*, bigramas, palavras a direita e a esquerda do verbo, etc. Além do conhecimento de fundo, o sistema aprende também a partir de um conjunto de exemplos. Dois experimentos foram realizados para um cenário multilíngue (inglês-português) e monolíngue (inglês). Para comparar a abordagem proposta, avaliou-se também os algoritmos de aprendizado: **árvore de decisão (C4.5)**, **Naive-Bayes** e ***Support Vector Machine*** (***SVM),*** usando o Weka.

Silva e Vieira (2007) apresentaram a comparação entre duas técnicas de **aprendizado de máquina**: **árvore de decisão e *support vector machine* (*SVM*)** para **categorização de textos** com seleção de características baseada em **informações linguísticas**. Na fase de pré-processamento dos documentos os termos irrelevantes (*stopwords*) foram removidos e utilizou-se o algoritmo de Martin Porter, desenvolvido para a língua portuguesa, para redução dos radicais[92]. As informações linguísticas foram adquiridas utilizando-se o **analisador sintático PALAVRAS (BICK, 2000).** As informações linguísticas consideradas foram as categorias gramaticais. Os classificadores baseados em árvore de decisão e SVM foram treinados usando a **ferramenta WEKA**.

Milidiu, Duarte e Cavalcante (2007) utilizaram três **algoritmos de aprendizado de máquina** para reconhecimento de nomes próprios: **cadeia de Markov (*Hidden Markov Models*)**, aprendizado baseado em **transformações** e ***Support Vector Machine (SVM).*** Para o algoritmo de Markov e o baseado em transformações utilizou-se implementações realizadas pelo laboratório LEARN da Puc-Rio, enquanto que para o algoritmo SVM utilizou-se uma implementação chamada libsvm (CHANG; LIN, 2001[93]). Os autores criaram manualmente um sistema de referência, chamado de *baseline system*, composto com nomes de localidades, personalidades e organizações extraídas da Web. No modelo de Markov, os autores usaram as etiquetas de reconhecimento de nomes próprios

[90] Dekang Lin. 1993. Principle based parsing without overgeneration. Proceedings of the 31st Meeting of the Association for Computational Linguistics (ACL- 93), Columbus, pages 112-120
[91] Adwait Ratnaparkhi. 1996. A Maximum Entropy Part-Of-Speech Tagger. Proceedings of the Conference on Empirical Methods in Natural Language Processing, New Jersey, pages 133-142.
[92] Disponível em http://snowball.sourceforge.net
[93] Chang, C. and Lin, C. (2001). Libsvm: a library for support vector machines, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

(NER): PER (para personalidades); ORG (para organizações); LOC (para localidades); e O (outros), como estados escondidos e as etiquetas de discurso (*pos-tag*) como observações. Assim, cada sentença é mapeada em uma sequência de etiquetas de discurso, e então o algoritmo de Viterbi encontra a melhor etiqueta NER. No modelo baseado em transformações, os autores usaram como classificadores iniciais, o sistema de referência e o modelo de Markov criado. Já no SVM, os autores precisaram mapear as sentenças num espaço vetorial. Assim, considerou-se uma janela de tamanho 5, composta com a palavra corrente, dois vizinhos antes e dois depois. Para cada elemento, foram elucidadas a própria palavra, a etiqueta de discurso e a classificação inicial, quando for o caso, representados por valores categóricos (zeros ou uns).

Caseli *et al.* (2008) apresentam uma **ferramenta visual gratuita** desenvolvida baseada no alinhador lexical híbrido LIHLA, proposto anteriormente pelos próprios autores

Aziz, Pardo e Paraboni (2008) destacam que a **abordagem estatística** de tradução consiste em encontrar a sentença em português que maximiza a **probabilidade de ser a tradução** de uma dada sentença em pspanhol. Dado um par de sentenças, assume-se que elas tem uma tradução mútua se existir pelo menos um **alinhamento possível** entre elas. Para **segmentar as sentenças** utilizou-se o **SENTER**[94], tanto para os textos em português como para espanhol (com algumas modificações). Para as tarefas de **alinhamento das sentenças**, utilizou-se o método ***Translation Corpus Aligner***[95] ***(TCAalign***[96]***)***. O modelo de tradução foi gerado usando o ***Cambridge Tool Kit (CMU***[97]***)***, enquanto que as **probabilidades foram obtidas** usando o ***ISI ReWrite Decoder tool***[98]. Para efeitos de avaliação, comparou-se os escores BLUE[99] do método estatístico proposto neste trabalho, com os obtidos pelo sistema baseado em **regras Apertium**.

[94] Pardo, T. A. S. SENTER: Um Segmentador Sentencial Automático para o Português do Brasil. NILC Technical Reports Series NILC-TR-06-01. University of São Paulo, São Carlos, Brazil (2006).
[95] Hofland, K. and Johansson, S. The Translation Corpus Aligner: A program for automatic alignment of parallel texts. In: Corpora and Cross-linguistic research Theory, Method, and Case Studies. S. Johansson & S. Oksefjell (Eds.): Rodopi, Amsterdam (1998).
[96] Caseli, H. M. Indução de léxicos bilíngues e regras para a tradução automática. Doctoral thesis, University of São Paulo, São Carlos, Brazil (2007).
[97] Clarkson, P. R. and Rosenfeld, R. Statistical Language Modeling Using the CMU - Cambridge Toolkit In Proceedings of ESCA Eurospeech (1997).
[98] Germann, U.; Jahr, M.; Knight, K.; Marcu, D., and Yamada, K. Fast Decoding and Optimal Decoding for Machine Translation. Proceedings of he 39th Annual Meeting of the Association for Computational Linguistics (2001).
[99] Papineni, K.; Roukos, S.; Ward, T. and Zhu, W. BLEU: a Method for Automatic Evaluation of Machine Translation. Proceedings of he 40th Annual Meeting of the Association for Computational Linguistics (2002) 311-318.

Morais e Ambrosio (2008) apresentam um sistema que tem como objetivo associar semanticamente documentos em domínios representados por ontologias. O sistema recebe a **ontologia**, definindo o contexto, e um documento texto, e usando técnicas de **mineração de dados**, compara-os, para obter o grau de relevância do documento no domínio. No presente trabalho foi escolhido o domínio 'acidentes de trânsito' (*traffic accident*), para o qual uma ontologia foi definida com o auxílio de especialistas que trabalham no TJGO e categorizam documentos, assim como os advogados que usam o repositório para consulta. A ontologia foi desenvolvida seguindo a metodologia proposta por Guizzardi (2000[100]) e usando o ambiente Protégé[101]. O **processo de categorização** é composto por três passos: **extração dos termos mais relevantes da ontologia**, identificando seus conceitos, atributos relações e instâncias; **extração dos termos mais relevantes dos documentos**, produzindo um vetor de termos com as frequências relativas (mineração de texto); e o **cálculo do grau de similaridade entre eles**, usando os **coeficientes de Jaccard** e ***Overlap***[102]. O sistema é avaliado usando as métricas precisão (*precision*), revocação (*recall*) e *fall-out.*

Caminada, Quental e Garrao (2008) destacam que o processo de identificação é baseado na localização de preposições no corpus a partir de sua anotação morfossintática. Sendo assim, ambos os corpora foram anotados morfossintaticamente pela **ferramenta PALAVRAS** e sofreram um processo de atomização para reverter as expressões multivocabulares já identificadas por esta ferramenta e assim poder ter seus multivocábulos re-identificados e avaliados contra a evidência de corpus. Esta ferramenta implementa cinco **algoritmos** amplamente descritos e utilizados na literatura **de multivocábulos e colocações**[103]: o **Teste-T**, o ***Chi-Square***, o ***Log Likelihood***, o ***Mutual Information*** e o ***Dice Coefficient***. Quando uma preposição é identificada, a palavra anterior e as duas subsequentes são armazenadas, compondo uma janela de análise.

Seno e Nunes (2008) destacam que para calcular a distância semântica entre uma sentença e um cluster, o **sistema SiSPI** implementa três **medidas**

---

[100] GUIZZARDI, G. Desenvolvimento para e com reuso: Um estudo de caso no domínio de vídeo sob demanda. Master's thesis, Universidade Federal do Espírito Santo, 2000.
[101] MUSEN, M. A. Protege-ii: Computer support for development of intelligent systems from libraries of components. MEDINFO 95 - World Congress on Medical Informatics, 8 1995.
[102] LOH, S. Descoberta de conhecimento em textos. Available in http://atlas.ucpel.tche.br/~loh/, last access in Sep/07, Dez 2005.
[103] MANNING, C. D., SCHUTZE, H. Foundations of Statistical Natural Language Processing, The MIT Press, Cambridge, Massachusetts, Londres – Inglaterra, 1999.

estatísticas de similaridade: ***Word Overlap***[104], ***TF-IDF*** (*Term Frequency Inverse Document Frequency*[105]) e ***TF-ISF*** (*Term Frequency Inverse Sentence Frequency*[106]). O sistema é composto por dois módulos: **segmentação de sentenças** (*Sentence Splitting*), desempenhado pelo **SENTER**[107], e **clusterização de sentença** (*Sentence Clustering*), que usa o método de **clusterização incremental Singlepass**[108]. Neste método, cada sentença de entrada deve ser inserida num cluster já existente ou em um novo cluster. Essa decisão é baseada num limiar de similaridade.

Aziz, Pardo e Paraboni (2009), com o intuito de avaliar diferentes parâmetros de treinamento e decodificação, vários experimentos de tradução do português do Brasil para o inglês americano foram realizados. Segundo os autores, foram utilizadas as **heurísticas de Moses** (KOEHN *et al.,* 2007[109]) e **trigramas** no sistema de **tradução automática estatística** descrito em um trabalho anterior (AZIZ *et al.,* 2009[110]). As opções analisadas foram a heurística de alinhamento, o tamanho máximo da frase, o uso de pesos de importância lexical e tuning.

Seno e Nunes (2009) afirmam que a fusão sentencial pode ser de duas formas: **por interseção**, que combina na sentença de saída somente as informações que se repetem nas sentenças de entrada; e **por união** de informações, que preserva todas as informações das sentenças de entrada na sentença de saída. A fusão de sentenças é feita em três passos: **identificação de informações comuns (alinhamento), fusão e linearização**. O sistema recebe de entrada um conjunto de sentenças similares previamente processadas pelo **parser Palavras (BICK, 2000[111]).** Para cada sentença, o parser fornece informações de discurso (*part-of-speech*) e de dependência sintática entre palavras e chunks, além do lema de cada palavra. Durante o alinhamento e fusão, o sistema faz uso da **base de sinônimos Tep1 (MAZIERO *et al*., 2008[112]),** de uma stoplist, para a identificação

---

[104] Radev, D., Otterbacher, J.: Zhang, Zhu.: Cross-document Relationship Classification for Text Summarization. In: Computational Linguistics (to appear, 2008).
[105] Salton, G., Allan, J.: Text Retrieval Using the Vector Processing Model. In: 3rd Symposium on Document Analysis and Information Retrieval. In: 3rd Symposium on Document Analysis and Information Retrieval. University of Nevada, Las Vegas (1994).
[106] Larocca Neto, J., Santos, A.D., Kaestner, C.A.A., Freitas, A.A.: Document Clustering and Text Summarization. In: 4th International Conference Practical Applications of Knowledge Discovery and Data Mining – PAAD 2000, pp. 41–55 (2000).
[107] Pardo, T.A.S.: SENTER: Um Segmentador Sentencial Automático para o Português do Brasil. Technical Report NILC-TR-06-01, São Carlos-SP, Brazil, 6p (2006).
[108] Van Rijsbergen, C.J.: Information Retrieval, 2nd edn. Butterworths, Massachusetts (1979).
[109] Koehn, Philipp et. al. (2007) “Moses: Open Source Toolkit for Statistical Machine Translation”. ACL-2007.
[110] Aziz, Wilker Ferreira, Thiago Alexandre Salgueiro Pardo and Ivandré Paraboni (2009) “Statistical Phrase-based Machine Translation: Experiments with Brazilian Portuguese”. VII Encontro Nacional de Inteligência Artificial (ENIA-2009).
[111] Bick, E. (2000). The Parsing System “Palavras” - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework, Aarhus University Press.
[112] Maziero, E.G., Pardo, T.A.S., Di Felippo, A., Dias-da-Silva, B.C. (2008). A Base de Dados Lexical e a Interface Web do TeP

de palavras irrelevantes ao alinhamento. Durante a linearização, um **gerador de formas superficiais,** desenvolvido no contexto do trabalho de **Caseli (2007[113]),** é usado para auxiliar na realização da sentença. A linearização envolve os aspectos gramaticais da sentença a ser gerada. Assim, a árvore resultante da fusão é percorrida, gerando todas as sentenças possíveis. As sentenças são pontuadas usando o **sistema jNina** (PEREIRA E PARABONI, 2007[114]).

Salles *et al*. (2009) tem como objetivo propor uma nova abordagem para o tratamento dos efeitos temporais em **algoritmos de classificação**. Definiu-se **o fator de ajuste temporal**, que faz uma ponderação dos documentos de treino, em função do momento de sua criação e do momento da criação do documento a ser classificado. A definição desse fator envolve dois aspectos: o **ponto de referência**, que define o momento de interesse da classificação, e a **distância temporal**, que define a separação em unidade de tempo do momento de criação do documento ao ponto de referência. Neste trabalho, os autores propõem **dois classificadores temporais** baseados no **modelo de espaço vetorial**: ***Rocchio* e *Knearest Neighbours* (KNN),** e compararam os resultados obtidos pelas versões temporais desses algoritmos com as versões tradicionais.

Braga, Monard e Matsubara (2009) destacam que os **algoritmos de aprendizado semi-supervisionado** pode ser de uma única visão (*single-view*) ou de múltiplas visões (*multi-view*), No presente trabalho, utilizou-se o algoritmo de uma **única visão *Self-Training*[115]**, e o de **múltiplas visões**, ***Co-Training*[116]**, ambos com **algoritmo de aprendizado supervisionado** Multinomial ***Naive Bayes*** (MNB)[117]. As base de dados foram representadas no formato de atributos-valores, onde unigramas foram usadas como primeira visão, e bigramas como segunda visão. Foi usado o **pré-processador PreTexT II**[118], as *stopwords* foram removidas e a análise de radicais realizada[119].

Villavicencio, Caseli e Machado (2009) avaliaram a influência de duas

---

2,0 - Thesaurus Eletrônico para o Português do Brasil. In: Anais do VI Workshop em Tecnologia da Informação e da Linguagem Humana - TIL, pp, 390-392.

[113] Caseli, H.M. (2007). Indução de Léxicos Bilíngues e Regras para a Tradução Automática. Tese de Doutorado. ICMC-USP, 158 p.

[114] Pereira, D.B. and Paraboni, I. (2007). A Language Modelling Tool for Statistical NLP. In: Anais do V Workshop em Tecnologia da Informação e da Linguagem Humana – TIL, pp. 1679-1688.

[115] Chapelle, O., Sch olkopf, B., Zien, A.: Introduction to semi-supervised learning. In: Semi-Supervised Learning (Adaptive Computation and Machine Learning). (2006) 1-12.

[116] Blum, A., Mitchell, T.: Combining labeled and unlabeled data with Co-Training. In: COLT '98: Proceedings of the 11th Annual Conference on Computational Learning Theory. (1998) 92-100.

[117] McCallum, A., Nigam, K.: A comparison of event models for naive bayes text classification. Technical Report WS-98-05, Association for the Advancement of Artificial Intelligence (1998).

[118] Disponível em http://www.icmc.usp.br/~caneca/pretext.htm

[119] Porter, M.F.: An algorithm for suffix stripping. Program: electronic library and information systems 40(3) (2006) 211-218.

abordagens para identificação de expressões multi-palavras: a primeira baseada em **estatísticas**, utilizando a medida de *pointwise mutual information* (PMI) e informação mútua (MI), como implementado no pacote estatístico Ngram[120]; e a segunda, baseada em **alinhamento lexical** entre o português e o inglês, gerado pelo **alinhador estatístico de palavras GIZA++**[121]. As sentenças do corpus usado foram inicialmente alinhadas usando o **alinhador *Translation Corpus Aligner (TCA[122])*** e anotadas usando o **Apertium**[123], enquanto que as ngramas candidatas foram anotadas usando o *Tree Tagger*[124].

4.2.1.3. Material empírico utilizado

Nesta seção, procurou-se destacar os materiais empíricos utilizados em cada artigo selecionado para análise, ou seja, identificar as bases de dados utilizadas e o idioma foco destes trabalhos. Espera-se verificar se a comunidade científica nacional foi capaz de criar um arcabouço experimental a partir do qual as pesquisas podem ser desenvolvidas, constituído tanto por ferramentas como por bases de documentos (*corpus*). Em outras palavras, pretende-se identificar os recursos disponíveis e utlizados ao longo dos anos, e elaborar um catálogo de possibilidades para pesquisas futuras na área de PLN.

Na TAB. 10 é apresentada a síntese dos resultados obtidos pela análise de conteúdo, de acordo com os materiais empíricos usados. Para cada artigo analisado, procurou-se identificar o material empírico, e se o mesmo havia sido construído especialmente para o trabalho em questão e qual o idioma dos documentos que o compõe. Considerou-se o fato de ter sido o material empírico CONSTRUÍDO, caso o mesmo tenha sido criado com o propósito inicial de ser usado nos experimentos do trabalho, e de ser CONHECIDO, caso o mesmo tenha sido reutilizado. Assim, se ao apresentar o material empírico, o autor incluir

---

[120] Banerjee, S. and Pedersen, T. (2003). The design, implementation and use of the ngram statistics package. In In Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics, pages 370–381.

[121] Och, F. J. and Ney, H. (2000). Improved statistical alignment models. In Proceedings of the 38th Annual Meeting of the ACL, pages 440–447, Hong Kong, China.

[122] Hofland, K. (1996). A program for aligning English and Norwegian sentences. In Hockey, S., Ide, N., and Perissinotto, G., editors, Research in Humanities Computing, pages 165–178, Oxford. Oxford University Press.

[123] Armentano-Oller, C., Carrasco, R. C., Corbí-Bellot, A. M., Forcada, M. L., Ginestí-Rosell, M., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Ramírez-Sánchez, G., Sánchez-Martínez, F., and Scalco, M. A. (2006). Open-source Portuguese-Spanish machinetranslation. In Proceedings of the VII Encontro para o Processamento Computacionalda Língua Portuguesa Escrita e Falada (PROPOR-2006), pages 50–59, Itatiaia-RJ, Brazil.

[124] Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In International Conference on New Methods in Language Processing.

referências a outros trabalhos, seu material empírico será considerado conhecido. Se o autor fez uso de um *corpus* de outro trabalho e fez um recorte, também será considerado conhecido. Considerou-se a opção 'Não se aplica', para os trabalhos que não apresentaram experimentos envolvendo base de documentos.

TABELA 10
Análise de conteúdo das publicações: dimensão Material empírico

Autores	Material empírico utilizado	Construído ou Conhecido	Idioma
Semeghini-Siqueira, Costa e Cohn (1986)	universo linguístico em 'sistema solar', com o auxílio de dois astrônomos da USP, e recorrendo-se a literatura específica	Construído	Português
Ziviani e Albuquerque (1987)	---	Não se aplica	Não se aplica
Ripoll e Mendes (1988)	---	Não se aplica	Português
Fusaro e Ziviani (1989)	---	Não se aplica	Genérico
Strube De Lima (1990)	---	Não se aplica	Genérico (exemplos em português)
Leffa (1991)	cinco pequenos textos de narrativas jornalísticas em inglês, de aproximadamente 100 palavras cada um	Construído	Português e Inglês (tradução)
Rocha *et al.* (1992)	utilizou-se uma base de 45 textos descrevendo técnicas de enfermagem.	Construído	Inglês
Rocha, Rocha e Huff (1993)	dicionário de dados de um sistema especialista (Iliad - OpenClinical AI Systems in clinical practice) e o UMLS Metathesaurus	Construído	Inglês
Robin (1994)	---	Não se aplica	Inglês
Julia, Seabra e Semeghini-Siqueira (1995)	---	Não se aplica	Inglês
Barros (1996)	---	Não se aplica	Inglês
Rosa (1997)	---	Não se aplica	Português
Oliveira e Wazlawick (1998)	utilizou-se textos compostos por três sentenças, mas não explicitaram o formato destes arquivos (somente que obedece o padrão "sujeito verbo objeto. Ele/ela verbo objeto")	Construído	Inglês
Carvalho e Strube De Lima (1999)	---	Não se aplica	Genérico (desenvolvido para o português).

Kinoshita (1999)	exemplos extraídos da Bíblia, livro de Mateus, em grego, inglês e português	Construído	Grego para inglês e português
Barcia *et al.* (1999)	---	Não se aplica	Português
Berber Sardinha (1999)	32 milhões de palavras oriundas de jornais, conversas informais, artigos e outros	Construído	Português
Villavicencio (1999)	---	Não se aplica	Inglês
Jose Neto e Menezes (2000)	corpus Penn Treebank, com textos em inglês, e o corpus Tycho Brahe, com textos em português	Conhecido.	Inglês e Português
Berber Sardinha (2000)	corpus com mais de 140 milhões de palavras	Construído	Português
Padilha e Viccari (2000)	---	Não se aplica	Exemplos em Português
Larocca Neto *et al.* (2000)	base de documentos TIPSTER, contendo textos em inglês de revistas sobre computação, hardware, software, etc.	Conhecido.	Inglês
Dias-Da-Silva *et al.* (2000)	---	Não se aplica	Português
Rossi *et al.* (2001)	corpus constituído por 15 textos do Jornal Correio do Povo (de Porto Alegre)	Construído	Português
Gamallo *et al.* (2001)	textos em Português P.G.R. (*Portuguese General Attorney Opinions*), constituído por documentos de jurisprudência	Conhecido	Português (genérico)
Gonzalez e Strube De Lima (2001)	corpus de teste com 7.095 palavras, constituído por 34 resumos de dissertações do PPGCC/PUCRS	Construído	Português
Orengo e Huyck (2001)	foram utilizadas de 2.800 palavras escolhidas aleatoriamente	Construído	Português
Souza, Pereira e Nunes (2001)	artigos científicos de computação retirados da Revista Brasileira de Informática na Educação e dos anais do Simpósio Brasileiro de Informática na Educação	Construído	Português
Jose Neto e Moraes (2002)	---	Não se aplica	Português
Bidarra (2002)	---	Não se aplica	Português
Pardo e Rino (2002)	Theses Corpus	Conhecido (desenvolvido por um dos autores)	Português
Schulz *et al.* (2002)	---	Não se aplica	Português e Inglês
Bonfante, A. G. ; Nunes, M. G. V.	corpus NILC	Conhecido	Português

Zavaglia (2003)	---	Não se aplica	Português
Martins, Monard e Matsubara (2003)	corpus do NILC	Conhecido	Português
Pardo, Rino e Nunes (2003)	CorpusDT	Conhecido	Português
Gasperin e Strube De Lima (2003)	corpus do NILC	Conhecido	Português
Oliveira, Garrao e Amaral (2003)	corpus do NILC	Conhecido	Português
Alves e Chishman (2004)	CETENfolha e ZERO Hora (Brasil) e CETENpublico e COMPARA (Portugal)	Conhecido	Português
Specia e Nunes (2004)	---	Não se aplica	Inglês e Português (tradução)
Rino *et al.* (2004)	corpus TeMário	Conhecido	Português
Aluisio *et al.* (2004)	---	Não se aplica	Português
Matsubara, Monard e Batista (2004)	base news e a base lnai	Conhecido	Inglês
Pardo, Marcu e Nunes (2005)	*Text REtrieval Conference (TREC 2002)*	Conhecido	Inglês
Caseli, Nunes e Forcada (2005)	CorpusFAPESP	Conhecido	Português, Espanhol e Inglês
Specia, Nunes e Stevenson (2005)	Compara	Conhecido	Inglês
Silva, Vieira e Osorio (2005)	corpus do NILC	Conhecido	Português
Piltcher *et al.* (2005)	três bases distintas: históricos de sessões, documentos da biblioteca digital e a ontologia de domínio	Construído	Independe (Português e Inglês)
Rino e Seno (2006)	corpus TeMário e corpus Rhetalho	Conhecido	Português (independente)
Caseli e Nunes (2006)	CorpusFAPESP	Conhecido	Português e Espanhol (tradução)
Balage Filho *et al.* (2006)	coleção de documentos em Português disponibilizada pelo CLEF	Conhecido	Português
Enembreck *et al.* (2006)	uma amostra de 52 projetos de mestrado escritos por membros do curso de Ciência da Computação da PUC-PR	Construído	Português
Leite e Rino (2006)	corpus TeMário	Conhecido	Português

Moraes e Strube De Lima (2007)	corpus PLN-BR CATEG	Conhecido	Português
Kinoshita *et al.* (2007)	corpus contendo informações do site do Metrô-SP, contendo 16.536 palavra em 800 sentenças	Construído	Português
Specia, Stevenson e Nunes (2007)	corpus inglês-português contendo 500 sentenças e o Senseval-3.	Construído e Conhecido (dois experimentos)	Inglês-Português e Inglês
Silva e Vieira (2007)	Corpus NILC	Conhecido	Português
Milidiu, Duarte e Cavalcante (2007)	SNR-CLIC (Mac-Morpho)	Conhecido	Português
Caseli *et al.* (2008)	---	Não se aplica	Português-Inglês e Português-Espanhol
Aziz, Pardo e Paraboni (2008)	revista eletrônica de Pesquisa da FAPESP	Conhecido	Português-Espanhol
Morais e Ambrosio (2008)	documentos de jurisprudência do Tribunal de Justiça do estado de Goiás – TJGO	Construído	Português
Caminada, Quental e Garrao (2008)	corpus Jornalístico, composto de textos de um jornal e o corpus Internet, construído a partir da ferramenta WebBootCat	Construído	Português
Seno e Nunes (2008)	corpus composto por 20 coleções de notícias, coletado manualmente a partir de vários sites de agências de notícias	Construído	Português
Aziz, Pardo e Paraboni (2009)	revista eletrônica de Pesquisa da PAFESP	Conhecido	Português-Inglês
Seno e Nunes (2009)	corpus NILC	Conhecido	Português
Salles *et al.* (2009)	duas coleções de documentos: da Biblioteca Digital da ACM e da base de dados MedLine	Conhecido	Inglês (Genérico)
Braga, Monard e Matsubara (2009)	foram utilizadas cinco bases de documentos:	Construído	Inglês
Villavicencio, Caseli e Machado (2009)	corpus paralelo português-inglês contendo 283 textos em português e sua versão em inglês, extraídos do Jornal de Pediatria	Construído	Português-inglês

Semeghini-Siqueira, Costa e Cohn (1986) delimitaram o universo linguístico em 'sistema solar' e com o auxílio de dois astrônomos da USP, e recorrendo-se a literatura específica (glossário específico e explicações técnicas sobre Astronomia em português), explicitaram os significados e estabeleceram os relacionamentos entre as unidades lexicais. No artigo, sugere-se que esta inspeção, assim como a consequente formulação das regras foram feitas manualmente.

Em **Ziviane e Albuquerque (1987)**, os autores apresentam a árvore Patrícia construída para uma frase exemplo, não fazendo assim uso de uma coleção de textos.

Ripoll e Mendes (1988) apresentam algumas frases em português para ilustrar a definição do significado adequado do verbo "bater", também não fazendo uso de uma base de documentos.

Fusaro e Ziviani (1989) não apresentaram experimentos envolvendo exemplos de aplicação da linguagem de consulta desenvolvida.

Strube de Lima (1990) apresentou uma revisão de literatura sobre métodos e técnicas empregadas a correção ortográfica automática e portanto não apresentou nem experimentos, nem resultados práticos.

Leffa (1991) destaca que antes de se testar o dicionário com usuários, era necessário fazer um levantamento da cobertura dos 4.700 termos inserido no dicionário criado, em textos de diferentes áreas. Foram selecionados aleatoriamente 6 segmentos de textos de 500 palavras cada um, produzindo um corpus de 30.000 palavras. Segundo o autor, os resultados obtidos para a cobertura destes 4.700 verbetes justificam um trabalho mais amplo de avaliação envolvendo leitores verdadeiros interagindo com textos autênticos. Assim, cinco pequenos textos de narrativas jornalísticas em inglês, de aproximadamente 100 palavras cada um, foram usados para o teste de compreensão de leitura. Um grupo de 43 alunos foram classificados de acordo com a proficiência em língua inglesa: iniciantes e intermediários. Somente os iniciantes foram usados na pesquisa. Dois testes de compreensão foram administrados para cada sujeito: um usando o dicionário tradicional e o outro usando o eletrônico.

Em **Rocha *et al.* (1992)**, a base de dados foi criada para ilustrar o funcionamento das redes neurais utilizadas. Com o objetivo de extrair conhecimento em interfaces de banco de dados em LN, o trabalho avaliou a possibilidade de fazer uso de diversos bancos de dados de termos médicos (ou clínicos), utilizou-se uma base de 45 textos descrevendo técnicas de enfermagem.

Em **Rocha, Rocha e Huff (1993)**, diante do objetivo de demostrar a abordagem adotada, os autores utilizaram um dicionário de dados de um sistema especialista (*Iliad - OpenClinical AI Systems in clinical practice*) como vocabulário de origem e o *UMLS Metathesaurus* como vocabulário alvo (*target*). O léxico usado tem 4.351 entradas representando mais de três mil conceitos (p. 691).

Em **Robin (1994),** tendo em vista que o objetivo do trabalho era o desenvolvimento de sumarizadores automáticos usando fatos históricos, foi apresentada uma análise de corpus de sumários sobre esportes escritos por humanos. O autor apresenta, para dois exemplos de narrativas de um jogo de basquete, as operações de revisão geradas, assim como o processo de geração de maneira incremental, utilizando estas operações, para gerar as frases que compõem o sumário.

Julia, Seabra e Semeghini-Siqueira (1995) não apresentaram experimentos envolvendo exemplos do analisador desenvolvido, e sim inúmeros exemplos de regras de produção utilizadas (expressões *lambda*s) (p. 809 e 810).

Barros (1996) não apresentou experimentos envolvendo exemplos do modelo desenvolvido, e sim um exemplo detalhado de como uma consulta (*query*) seria processada.

Rosa (1997) apresentou tabelas ilustrando os vetores de características semânticas extraídas para alguns substantivos e verbos de sentenças em português, e discute os resultados alcançados para situações específicas (p. 242).

Oliveira e Wazlawick (1998) discutem vários critérios de configuração das redes neurais utilizadas durante os experimentos, e afirmam que os resultados foram obtidos utilizando-se textos compostos por três sentenças, mas não explicitaram o formato destes arquivos (somente que obedece o padrão "sujeito verbo objeto. Ele/ela verbo objeto"). Os exemplos apresentados estão no idioma inglês.

Carvalho e Strube de Lima (1999) discutem as principais diferenças entre as distribuições léxico-categorial e a linguística-cognitiva, e algumas constatações a cerca dos modelos construídos, mas não realizaram experimentos explícitos usando bases de documentos.

Kinoshita (1999) propõe um sistema de tradução baseado em exemplos extraídos da Bíblia, livro de Mateus, em grego, inglês e português, anotado de acordo com a anotação de Strong (*Strong's annotation*). Os exemplos são organizados em palavras, bigramas e trigramas.

Barcia *et al.* (1999) apresentaram a proposta da utilização de Raciocínio baseado em casos na recuperação de textos jurídicos e não apresentaram experimentos envolvendo bases de dados.

O trabalho apresentado em **Berber Sardinha (1999)** se desenvolve em

torno de estudos de caso considerados relevantes para a área. O corpus usado para os estudos de casos é constituído de mais de 32 milhões de palavras oriundas de jornais, conversas informais, artigos acadêmicos e outros.

Villavicencio (1999) não apresentou experimentos nem tão pouco simulações, e sim trechos da hierarquia proposta.

Jose Neto e Menezes (2000) afirmam que "a dificuldade central da anotação morfológica, em comparação com línguas tais como o inglês, reside no fato de que há a necessidade de um número bem maior de etiquetas para representar a maior riqueza morfológica da língua portuguesa" (p. 62). Os autores apresentam como exemplo o corpus *Penn Treebank*, com textos em inglês, que usa um conjunto de 36 etiquetas morfológicas, a menos de pontuações, enquanto que o corpus *Tycho Brahe*, com textos em português, usa 231 etiquetas. Mesmo assim, os autores afirmam que o método proposto neste trabalho não é afetado por esta dificuldade. Dois experimentos foram realizados: no primeiro experimento realizado utilizou-se um trecho que não faz mais parte do corpus *Tycho Brahe* (segundo os autores, foi usado o que era disponível na época da realização do experimento), composto de 1.812 palavras e dividido em duas partes: corpus de treinamento, contendo 1.684 itens lexicais (palavras e pontuações) e corpus de aplicação, com 128 itens lexicais. O segundo experimento, segundo os autores, é mais abrangente e confiável, sob o ponto de vista prático. Os três módulos foram treinados com o uso de um texto de António das Chagas, que faz parte do corpus *Tycho Brahe*, e que é composto de 57.425 palavras, divididas da seguinte forma: corpus de treinamento contendo 51.017 itens lexicais e corpus de aplicação com 6.408 itens lexicais (p. 62).

Berber Sardinha (2000) utilizou um corpus constituído por mais de 140 milhões de palavras com o intuito de contrastar algumas prosódias semânticas do inglês com as suas equivalentes do português.

Padilha e Viccari (2000) apresentaram um trabalho teórico sem a utilização de material empírico.

Larocca Neto *et al.* (2000) tem como objetivo propor um sistema treinável para sumarização de notícias. O sistema foi treinado e testado usando a base de documentos TIPSTER (HARMAN, 1994[125]), contendo textos em inglês de revistas sobre computação, hardware, software, etc. Dentre os documentos

[125] Harman, D. Data Preparation. In R. Merchant, editor, The Proceedings of the TIPSTER Text Program Phase I. Morgan Kaufmann Publishing Co. 1994.

disponíveis, 33.658 contém o sumário produzido pelo próprio autor do texto. Para o experimento realizado no trabalho foi usado um conjunto de 900 documentos, dividido em dois subconjuntos de 100 e 800 documentos.

Dias-da-Silva *et al.* (2000) não apresentaram experimentos nem simulações. Foram apresentadas algumas telas do editor de thesaurus construído ilustrando algumas entradas fornecidas.

Em **Rossi *et al.* (2001),** foi utilizado um corpus linguístico, para desenvolver os estudos de correferência nominal para o caso de descrições definidas, constituído por 15 textos e artigos do Jornal Correio do Povo, de Porto Alegre, editados no segundo semestre do ano de 1999. Do total de 248 sentenças dos 15 artigos do corpus, extraiu-se 1.879 sintagmas nominais, sendo que 880 destes (aproximadamente 50%) são descrições definidas. Segundo os autores, este processo de preparação do corpus está descrito em detalhes em outro trabalho (VIEIRA *et al.,* 2000[126]).

Gamallo, Agustini e Lopes (2001) testaram o o sistema com um corpus de textos em Português P.G.R. (*Portuguese General Attorney Opinions*), constituído por documentos de jurisprudência, do qual foram extraídas 1.643.579 ocorrências de palavras. Segundo os autores, o corpus foi, primeiro, marcado pelo etiquetador *(part-of-speech)* apresentado por Marques (2000[127]). Em seguida, sequências de blocos (*sequences of chucks*) foram analisadas por um parser parcial (ROCIO *et al.*, 2001[128]). Usando heurísticas de associação, essas porções foram unidas para criar dependências sintáticas binárias.

Em **Gonzalez e Strube de Lima (2001),** foi utilizado um corpus de teste com 7.095 palavras (excluindo-se as *stopwords*), constituído por 34 resumos de dissertações do Programa de Pós-Graduação em Ciência da Computação (PPGCC) da Faculdade de Informática da PUCRS. Em média, os documentos possuem, cada um, 208 palavras.

Souza, Pereira e Nunes (2001) utilizaram 12 exemplares de revistas científicas brasileiras da área de computação, formando um corpus de 58 artigos em português, objetivando um levantamento de padrões morfossintáticos das palavras-chave elaboradas pelos autores dos artigos: combinações de categorias gramaticais.

---

[126] Vieira, R. et al. . Extração de sintagmas nominais para o processamento de correferência. V Encontro para o processamento computacional da Língua Portuguesa escrita e falada - PROPOR, Atibaia SP, 19-22 Nov 2000

[127] Nuno Marques. Uma Metodologia para a Modelação Estatística da Subcategorização Verbal. PhD thesis, Universidade Nova de Lisboa, Lisboa, Portugal, 2000.

[128] V. Rocio, E. de la Clergerie, and J.G.P. Lopes. Tabulation for multi-purpose partial parsing. Journal of Grammars, 4(1), 2001.

A avaliação do sistema desenvolvido pelos autores foi feita utilizando-se dezoito artigos científicos de computação retirados da Revista Brasileira de Informática na Educação e dos anais do Simpósio Brasileiro de Informática na Educação - 1998.

Orengo e Huyck (2001) utilizaram um vocabulário de 32 mil palavras distintas obtidas a partir da versão para o português do algoritmo de Porter[129]. Deste conjunto de palavras, foram selecionadas aleatoriamente um conjunto de 2.800 palavras paras as quais foram atribuídos manualmente os radicais corretos.

Jose Neto e Moraes (2002) não apresentaram experimentos envolvendo coleções de documentos. Procurou-se ilustrar a construção de autômatos a partir de uma gramática que representa um subconjunto da língua portuguesa. Segundo os autores, o método proposto apresenta "uma aplicabilidade relativamente geral" (p. 4), e que portanto, pode ser devidamente estendido para levar em consideração os aspectos da linguagem natural não considerados na simplificação imposta. A gramática simplificada usada como base para o raciocínio não considera importantes aspectos de dependência de contexto, que certamente devem ser levados em conta em outras etapas do processamento da linguagem.

Em **Bidarra (2002)**, por ser um trabalho essencialmente teórico e descritivo, não foram realizados experimentos. O autor apresentou alguns exemplos de afasia, em português, para ilustrar o modelo proposto.

Em **Pardo e Rino (2002),** utilizou-se o Theses Corpus (PARDO, 2002[130]), contendo 10 introduções de teses e dissertações da área de computação, tendo, em média, 530 palavras cada introdução. Esse corpus foi escolhido pelo fato dos textos apresentarem a estrutura Problema-Solução e serem acompanhados por sumários autênticos, ou seja, aqueles produzidos pelos próprios autores dos textos.

Em **Schulz *et al*. (2002)** é apresentada uma metodologia de indexação e recuperação de textos médicos e como tal, não apresenta a realização de experimentos. Ao final do artigo, é apresentado o resultado da metodologia aplicada a dois exemplos de textos em português e inglês, de conteúdo idêntico.

Bonfante e Nunes (2002) não apresentaram resultados dos experimentos, apenas afirmaram que utilizou-se um conjunto de sentenças extraídas do corpus NILC. Como *treebank* alimentadora do processo, utilizou-se um conjunto

[129] Disponível em http://open.muscat.com.
[130] Pardo, T.A.S. (2002). DMSumm: Um Gerador Automático de Sumários. Dissertação de Mestrado. Departamento de Computação. Universidade Federal de São Carlos. São Carlos – SP.

de sentenças obtidas do corpus NILC[131] anotadas sintaticamente com o parser do Bick (2000[132]).

Zavaglia (2003) não apresentou experimentos, apenas dois exemplos da representação do item homônimo 'banco'.

Martins, Monard e Matsubara (2003) utilizam um corpus obtido do NILC com mais de 4.000 documentos em português dividido nos seguintes tópicos: didático, jornalístico, jurídico, literário e técnico. Foram selecionados 248 documentos jornalísticos classificados em quatro classes: informática, economia, esporte e política.

Em **Pardo, Rino e Nunes (2003),** utilizou-se, para treinar a rede neural do NeuralSumm, sentenças extraídas de um corpus de 10 textos científicos (introduções de teses e dissertações com aproximadamente 530 palavras e 19 sentenças cada) do domínio da Computação em Português do Brasil, chamado CorpusDT (FELTRIM *et al.*, 2001[133]). As sentenças dos textos foram classificadas em essencial, complementar ou supérflua, por 10 juízes linguistas computacionais e falantes nativos do Português do Brasil. Para cada sentença, foi extraído um conjunto de oito características (*features*), assumindo a classificação indicada pela maioria dos juízes.

Em **Gasperin e Strube de Lima (2003),** utilizou-se o corpus do NILC para gerar a lista de palavras e para avaliar os resultados obtidos com a recuperação com expansão de consultas. Este corpus contém 5.093 artigos em português publicados no jornal Folha de São Paulo no ano de 1994, sobre vários assuntos. Consultas foram realizadas e um especialista humano classificou os documentos como relevantes e não relevantes. Essa classificação manual foi usada para gerar os índices de revocação e precisão dos experimentos realizados.

Em **Alves e Chishman (2004),** quatro tradutores foram avaliados de acordo com sua capacidade de tradução de casos ambíguos utilizando como língua fonte, o Português e como língua alvo, o Inglês. Foram submetidas aos tradutores, 38 frases de fontes variadas, tais como os corpora eletrônicos CETENfolha e ZERO Hora (Brasil) e CETENpublico e COMPARA (Portugal).

---

[131] Disponível em www.nilc.icmc.sc.usp.br

[132] Bick, E. . The Parsing System. "Palavras" – Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus University Press, 2000.

[133] Feltrim, V.D.; Nunes, M.G.V.; Aluísio, S.M. (2001). Um corpus de textos científicos em Português para a análise da Estrutura Esquemática. Série de Relatórios do NILC. NILCTR-01-4. Available for download in www.nilc.icmc.usp.br/nilc/pessoas/valeria.htm

Oliveira, Garrao e Amaral (2003) destacam que os critérios definidos foram testados em um corpus em português brasileiro do Núcleo Interinstitucional de Linguística Computacional (NILC), descrito em (AIRES, 2001[134]), que contém cerca de 37 milhões de palavras, incluindo textos corretos (publicações como livros, jornais e revistas, ou seja, que foram submetidas a revisão de especialistas), incorretos (redações de alunos undergraduates e material de propaganda) e semi-corretos (extraídos de contratos, relatórios, dissertações de mestrado, etc).

Specia e Nunes (2004) não realizaram experimentos. Os verbos considerados problemáticos foram selecionados em um projeto anterior das autoras. Os idiomas escolhidos foram o inglês como língua fonte e o português como língua alvo.

Rino *et al.* (2004) compararam o desempenho de cinco sistemas de sumarização automática encontrados na literatura. Utilizou-se o corpus TeMário[135]), contendo 100 textos de jornais, construído com o propósito de sumarização automática. Estes textos foram obtidos da Folha de São Paulo (60 textos) e do Jornal do Brasil (40 textos). Os sumários apresentados foram produzidos manualmente por consultores da língua portuguesa.

Aluisio *et al.* (2004) autores apresentaram o Lácio-Web como sendo um repositório de recursos para o desenvolvimento de pesquisas da língua portuguesa do Brasil e de outras ferramentas linguísticas e computacionais.

Matsubara, Monard e Batista (2004) utilizaram duas bases de textos, news e lnai, para realizar os experimentos. A base news foi criada a partir da base mini-news[136], e contém 800 documentos classificados em duas classes, sci e talk, cada uma delas com 400 documentos. A base lnai contém títulos, resumos e referências de artigos sobre *Case-Based Reasoning (CBR)* e *Inductive Logic Programming (ILP)* retirados dos *Lecture Notes in Artificial Intelligence (LNAI)*, que contém 396 artigos, dos quais 277 (70%) são da classe CBR e 119 (30%) são da classe ILP.

Pardo, Marcu e Nunes (2005) utilizaram todas as sentenças extraídas dos dados da TREC'2002 (*Text REtrieval Conference*), com no máximo 10 palavras,

[134] Aires, R.V.X., Aluísio, S.M, Criação de um corpus com 1.000.000 de palavras etiquetado morfossintaticamente. Relatórios do NILC, NILC-TR-01-8, 2001.
[135] Pardo, T.A.S., Rino, L.H.M.: TeMário: A corpus for automatic text summarization (in Portuguese). NILC Tech. Report NILC-TR-03-09 (2003). Disponível em http://www.linguateca.pt/Repositorio/TeMario
[136] C.L. Blake and C.J. Merz. UCI Repository of Machine Learning Databases, 1998. http://www.ics.uci.edu/~mlearn/MLRepository.html.

que continham os 1.500 verbos mais frequentes do inglês. Segundo os autores, estes dados foram escolhidos, pelo fato de já estarem anotados por um reconhecedor de entidades mencionadas (REM): o BBN Identifinder (BIKEL *et al*., 1999[137]).

Caseli, Nunes e Forcada (2005), para testar o método proposto, utilizaram o corpus paralelo português-espanhol da Fapesp (CorpusFAPESP), composto por 1.292 artigos (646 em português e 646 em espanhol) da versão online da revista brasileira científica de Pesquisa da Fapesp[138].

Em **Specia, Nunes e Stevenson (2005),** utilizou-se sentenças em inglês contendo os verbos *to come, to get, to give, to go, to look, to make* e *to take,* extraídas do corpus **Compara** (FRANKENBERG-GARCIA; SANTOS, 2003[139]), contendo textos de livros de ficção. Cada sentença tem uma etiqueta de sentido que corresponde a tradução do verbo da sentença (feito automaticamente e revisado manualmente, em trabalho anterior). Juntamente com a etiqueta de sentido estão as etiquetas e lemas de todas as palavras e as relações sintáticas sujeito-objeto.

Silva, Vieira e Osorio (2005) utilizaram o corpus do Núcleo Interdisciplinar de Linguística Computacional (NILC) contendo 855 textos jornalísticos da Folha de São Paulo do ano de 1994, distribuídos nos assuntos informática, economia (*property*), esporte, política e turismo.

Em **Piltcher *et al.* (2005),** foram utilizadas três bases distintas presentes no Sistema de recomendação (LOH, 2004[140]): históricos de sessões, onde todas as mensagens enviadas ao chat foram gravadas. Assumiu-se que os termos que apareciam com frequência estavam grafados corretamente; documentos textuais, compostos pelos artigos científicos da biblioteca digital do sistema de recomendação; e a ontologia, que é considerada a fonte mais confiável, em relação as anteriores, porque foi criada de modo supervisionado por humanos.

Rino e Seno (2006) realizaram dois experimentos: num primeiro momento, foram utilizados 10 textos jornalísticos que já possuem seus sumários de referência, extraídos do corpus TeMário[141], com um total de 5.277 palavras,

[137] Bikel, D.M.; Schwartz, R.; Weischedel, R.M. (1999). An Algorithm that Learns What's in a Name. Machine Learning (Special Issue on NLP).
[138] A revista de pesquisa da FAPESP está disponível em http://revistapesquisa.fapesp.br com textos paralelos escritos em português do Brasil (original), e versões em Inglês e Espanhol.
[139] Frankenberg-Garcia, A. and Santos, D. (2003) "Introducing COMPARA: the Portuguese-English Parallel Corpus". Corpora in translator education, pp. 71-87.
[140] Loh, S. (2004). Investigação sobre a identificação de assuntos em mensagens de chat. Workshop de TI e Linguagem Humana - XXIV Congresso da Sociedade Brasileira de Computação, Salvador.
[141] Disponível em http://www.linguateca.pt/Repositorio/TeMario

aproximadamente uma página e meia para cada texto; num segundo momento, foram utilizados 20 textos jornalísticos, extraídos do corpus Rhetalho[142], anotados retoricamente por especialistas em RST, que também os anotaram com suas cadeiras de co-referência.

Caseli e Nunes (2006) citam que foi utilizado o corpus paralelo português-espanhol CorpusFAPESP composto por artigos da revista eletrônica de Pesquisa da Fapesp[143]. No entanto, apesar das autoras alegarem que experimentos já estão sendo feitos, não foram apresentados resultados de simulações ou exemplos de regras produzidas.

Balage Filho *et al.* (2006) utilizaram a coleção de documentos em português disponibilizada pelo CLEF que contém artigos jornalísticos de dois jornais: o brasileiro Folha de São Paulo e o português Público, dos anos de 1994 e 1995. Segundo os autores, o português foi escolhido por duas razões: por ser o idioma nativo dos autores e por ser a única linguagem suportada pelo GistSumm (sumarizador usado durante os experimentos).

Enembreck *et al.* (2006) utilizaram uma amostra de 52 projetos de mestrado escritos por membros do curso de Ciência da Computação da PUC-PR (curso de lotação dos próprios autores). Para cada um dos 22 professores do curso, foram selecionados 22 itens de conhecimento, e cada item de conhecimento foi representado por um vetor de 500 termos.

Leite e Rino (2006) utilizaram o corpus TeMário[144] durante os experimentos, dividido em três conjuntos de textos: 100 textos originais, com seus respectivos sumários (elaborados manualmente) e os extratos produzidos automaticamente.

Moraes e Strube de Lima (2007) utilizaram uma coleção de mais de 26 mil textos jornalísticos escritos em língua portuguesa do corpus PLN-BR CATEG[145], que reúne textos da Folha de São Paulo dos anos de 1994 a 2005. Para a realização dos experimentos de Langie (2004), foram utilizados os textos referentes a 1994 (Folha-Hierarq), e portanto não foram usados no presente trabalho. Segundo as autoras, a Folha-Hierarq é um subconjunto da Folha-RIcol[146].

---

[142] Disponível em http://nilc.icmc.usp.br/~thiago/rhetalho.html
[143] Disponível em http://revistapesquisa.fapesp.br
[144] Pardo, T.A.S., Rino, L.H.M.: TeMário: A corpus for automatic text summarization (in Portuguese). NILC Tech. Report NILC-TR-03-09 (2003)
[145] Coleção obtida através do projeto Recursos e Ferramentas para a Recuperação de Informação em Bases Textuais em Português do Brasil (PLN-BR).
[146] Disponível em http://www.linguateca.pt/Repositorio/Folha-RIcol/.

Kinoshita *et al*. (2007) apresentam o corretor gramatical para o Português CoGrOO, destacando seu funcionamento e componentes. Um importante recurso linguístico utilizado no CoGrOO é o corpus CETENFOLHA corpus[147], composto por textos jornalísticos com a notação morfosintática. Para comparar o desempenho do CoGrOO com o Regra, corretor gramatical do editor de texto Microsoft Word, os autores criaram um corpus a partir de informações contidas no site do Metrô-SP, contendo 16.536 palavra em 800 sentenças. Um especialista humano analisou o corpus e identificou 51 erros gramaticais.

Em **Specia, Stevenson e Nunes (2007),** dois experimentos foram realizados: no primeiro construiu-se um corpus Inglês-Português contendo 500 sentenças para cada um dos 10 verbos frequentes e considerados problemáticos, de acordo com trabalho anterior (SPECIA *et. al*., 2005[148]). No segundo experimento, adotou-se um cenário monolíngue, contendo sentenças em Inglês contendo 32 verbos do exemplo Senseval-3, usado em (MIHALCEA *et. al*. 2004).

Silva e Vieira (2007) utilizaram uma coleção de 855 documentos de um corpus composto por artigos do Jornal Folha de São Paulo do ano de 1994[149], classificados manualmente em 5 categorias, tais como: informática, imóveis, esporte, política e turismo. Segundo as autoras, este corpus foi cedido pelo NILC (Núcleo Interinstitucional de Linguística Computacional), ao grupo de pesquisa em PLN da Unisinos-RS. Em média cada documento da coleção possui 215 palavras e 124 palavras distintas por textos, totalizando 19.519 palavras distintas.

Milidiu, Duarte e Cavalcante (2007) utilizaram um corpus com 2.100 sentenças obtidas do SNR-CLIC[150], anotados manualmente com as etiquetas de discurso (*part-of-speech*).

Em **Caseli *et al.* (2008),** não foram realizados experimentos práticos. Os autores apresentaram a ferramenta gráfica VisualLIHLA desenvolvida baseada no alinhador lexical híbrido LIHLA, proposto anteriormente pelos próprios autores.

Aziz, Pardo e Paraboni (2008), com o objetivo de construir o modelo de tradução apresentado foram coletados 645 textos português-espanhol da revista eletrônica de Pesquisa da FAPESP. Apesar das bases conterem cerca de 450 mil

[147] Linguateca, CETENFolha, Brazilian-Portuguese annotated corpus (http://www.linguateca.pt/ Dec. 2006).
[148] Lucia Specia, Maria G.V. Nunes, and Mark Stevenson. 2005. Exploiting Parallel Texts to Produce a Multilingual Sense Tagged Corpus for Word Sense Disambiguation. Proceedings of the Conference on Recent Advances on Natural Language Processing (RANLP-2005), Borovets, pages 525-531.
[149] Disponível em http://www.inf.unisinos.br/~renata/laboratorio/mais_jornal_mt.htm
[150] Criado a partir do Mac-Morpho construído por sua vez a partir de textos do corpus do NILC.

palavras, os autores as consideraram bases pequenas, quando comparadas a outros experimentos realizados que utilizaram 200 milhões e até mesmo um bilhão de palavras. Utilizou-se um conjunto de treinamento composto por 17.000 pares de sentenças.

Em **Morais e Ambrosio (2008),** foram utilizados documentos de jurisprudência do Tribunal de Justiça do estado de Goiás – TJGO. Esses documentos foram manualmente categorizados por especialistas, permitindo avaliar os resultados obtidos pelo sistema. O repositório utilizado contém aproximadamente 150 mil documentos, que encontram-se disponíveis no site do TJGO para consultas. Para avaliar o sistema, dois conjuntos de 60 documentos foram usados, contendo cada um 30 documentos classificados como 'acidente de trânsito' e 30 de outra categoria.

Em **Caminada, Quental e Garrao (2008),** foram utilizados dois corpus: Corpus Jornalístico composto de textos de um jornal de grande circulação e possui material contemporâneo brasileiro, com não mais que uma década de idade, com mais de 32 milhões de *tokens* (palavras, sinais de pontuação, etc); e o Corpus Internet, construído a partir da ferramenta WebBootCat[151], que coleta textos na internet a partir de parâmetros definidos pelo usuário, como listas de palavras sementes.

Em **Seno e Nunes (2008),** utilizou-se um corpus composto por 20 coleções de notícias, todas contendo os mesmos tópicos, e coletado manualmente a partir de vários sites de agências de notícias, totalizando 1.153 sentenças em 71 documentos. Para criar um corpus de referência, cada sentença, de cada documento, foi manualmente classificada, ou seja, associada a um cluster, pela primeira autora desse trabalho.

Aziz, Pardo e Paraboni (2009) afirmam que todos os experimentos foram realizados para um corpus paralelo português-inglês da revista eletrônica de Pesquisa da PAFESP. Os dados de treinamento consiste num conjunto de 17.000 pares de sentenças. Para teste, foram usadas 649 pares desconhecidos.

Seno e Nunes (2009) construíram um modelo que foi induzido a partir do Corpus NILC[152], composto por 160 Mb de textos jornalísticos, usando o sistema

[151] KILGARRIFF, A., RYCHLY, P., SMRZ, P., TUGWELL, D., The Sketch Engine, Prodeedings from the Euralex 2004, França, p. 105-116. – 2004.
[152] Disponível em http://www.nilc.icmc.usp.br/~rh/corpus/

jNina (PEREIRA; PARABONI, 2007[153]).

Em **Salles *et al*. (2009),** para avaliar os classificadores propostos, foram realizados experimentos utilizando duas coleções de documentos provenientes de áreas de conhecimento distintas: a primeira é constituída por mais de 24 mil documentos coletados da Biblioteca Digital da ACM, contendo artigos da ciência da computação, criados no período entre 1980 e 2001, e classificados em 11 categorias; a segunda coleção é derivada da base de dados MedLine, constituída de mais de 800 mil documentos da área de Medicina, classificados em 7 classes distintas, criados entre 1970 e 1985.

Em **Braga, Monard e Matsubara (2009),** foram utilizadas cinco bases de documentos: uma contendo páginas da internet da base Courses[154], três contendo artigos de notícias de uma lista de discussão[155], e a última contendo dados de filmes[156].

Villavicencio, Caseli e Machado (2009) utilizaram um corpus paralelo português-inglês contendo 283 textos em português e sua versão em inglês, extraídos do jornal de Pediatria. Para avaliar as expressões multi-palavras candidatas, utilizou-se o glossário de pediatria, produzido pelo grupo TextQuim/TERMSUL[157] que contém ngramas extraídos do corpus com frequência superior a 5 e conferidos manualmente.

4.2.1.4. Resultados observados

Nesta seção são apresentados os resultados apresentados pelos autores dos trabalhos analisados, juntamente com as perpectivas de continuidade dos mesmos. Além disso, procurou-se identificar, para cada publicação avaliada, se os autores realizaram experimentos práticos, e se foram usados métodos automáticos de avaliação ou se o mesmo foi avaliado com base no julgamento humano. A síntese desses resultados é apresentada na TAB. 11.

---

[153] Pereira, D.B. and Paraboni, I. (2007). A Language Modelling Tool for Statistical NLP. In: Anais do V Workshop em Tecnologia da Informação e da Linguagem Humana – TIL, pp. 1679-1688.
[154] Blum, A., Mitchell, T.: Combining labeled and unlabeled data with Co-Training. In: COLT '98: Proceedings of the 11th Annual Conference on Computational Learning Theory. (1998) 92-100.
[155] http://people.csail.mit.edu/jrennie/20Newsgroups/
[156] Movie Review Data (http://www.cs.cornell.edu/People/pabo/movie-review-data/)
[157] http://www.ufrgs.br/textquim

TABELA 11
Análise de conteúdo das publicações: dimensão Resultados Observados

Autores	Realizou experimento?	Houve Avaliação?	Se "sim", manual ou automática
Semeghini-Siqueira, Costa e Cohn (1986)	Não, apenas exemplos	Não	Não
Ziviani e Albuquerque (1987)	Não, apenas considerações sobre o método desenvolvido	Não	Não
Ripoll e Mendes (1988)	Não, apenas algumas frases em português	Não	Não
Fusaro e Ziviani (1989)	Não, apenas os comandos que compõem a linguagem de consulta criada	Não	Não
Strube De Lima (1990)	Não.	Não	Não
Leffa (1991)	Sim, com usuários	Sim, usando medidas qualitativas (porcentagem) para avaliar "nível de compreensão fornecido ao leitor"	Manual
Rocha *et al.* (1992)	Não, apenas exemplos	Não	Não
Rocha, Rocha e Huff (1993)	Sim	Não apresentou resultados quantitativos (apenas "não apresentou diferença significativa")	Sim, por médicos
Robin (1994)	Sim	Não	Manual (Visual)
Julia, Seabra e Semeghini-Siqueira (1995)	Não.	Não	Não
Barros (1996)	Não, apenas um consulta.	Não	Manual (Visual)
Rosa (1997)	Sim	Sim	Manual: foram usadas sentenças ditas válidas e inválidas, que foram definidas manualmente.
Oliveira e Wazlawick (1998)	Sim	Não	Não apresentou índices de acerto/erro
Carvalho e Strube De Lima (1999)	Não foram apresentados nenhum experimento, nem mesmo exemplo de frase.	Não	Não
Kinoshita (1999)	Sim	Sem resultados estatísticos	Manual (Visual)
Barcia *et al.* (1999)	Não.	Não	Não
Berber Sardinha (1999)	Sim	Não	Não
Villavicencio (1999)	Não.	Não	Não
Jose Neto e Menezes (2000)	Sim	Sim	Automático

Berber Sardinha (2000)	Sim	Sim	Não
Padilha e Viccari (2000)	Não.	Não	Não
Larocca Neto *et al.* (2000)	Sim	Sim	Sim, usando validação cruzada, estatísticas (revocação/precisão)
Dias-Da-Silva *et al.* (2000)	Não.	Não	Não
Rossi *et al.* (2001)	Não.	Não	Sim, com três pesquisadores classificando manualmente as correferências nominais (humana)
Gamallo *et al.* (2001)	Sim	Não	Não
Gonzalez e Strube De Lima (2001)	Sim	Sim (precisão e revocação)	Automática
Orengo e Huyck (2001)	Sim	Sim	Automática
Souza, Pereira e Nunes (2001)	Sim	Sim	Manual
Jose Neto e Moraes (2002)	Não.	Não	Não
Bidarra (2002)	Não.	Não	Não
Pardo e Rino (2002)	Sim	Sim	sim.
Schulz *et al.* (2002)	Não.	Não.	Não
Bonfante, A. G. ; Nunes, M. G. V.	Não.	Não	Não
Zavaglia (2003)	Não.	Não	Não
Martins, Monard e Matsubara (2003)	Sim	Sim	Automático (%cross-validation error comparado a uma base de documentos previamente classificada por assunto)
Pardo, Rino e Nunes (2003)	Sim	Sim	Manual para classificar as sentenças em essencial, complementar e supérflua) e automática ao comparar o extrato gerado com o ideal (precisão e cobertura – redefinidos para este contexto)
Gasperin e Strube De Lima (2003)	Sim	Sim	Documentos relevantes foram classificados na mão e somente 7 consultas foram realizadas

Oliveira, Garrao e Amaral (2003)	Sim (listagem de locuções que casaram com o critério)	Não	Não
Alves e Chishman (2004)	Sim	Sim	Manual (Visual)
Specia e Nunes (2004)	Não.	Não	Não se aplica
Rino *et al.* (2004)	Sim	Sim	Automático (usando os sumários ideais)
Aluisio *et al.* (2004)	Não.	Não.	---
Matsubara, Monard e Batista (2004)	Sim	Sim	Automático (%10 fold-cross-validation)
Pardo, Marcu e Nunes (2005)	Sim	Sim	humana e automática (usando dados que foram anotados manualmente)
Caseli, Nunes e Forcada (2005)	Sim	Sim	automática (usando dados que foram anotados manualmente)
Specia, Nunes e Stevenson (2005)	Sim	Sim	automática (medidas objetivas) e manual (subjetiva)
Silva, Vieira e Osorio (2005)	Sim	Sim	Automático
Piltcher *et al.* (2005)	Sim	Sim	Automático (precisão e abrangência foram calculadas utilizando base corrigida manualmente)
Rino e Seno (2006)	Sim	Sim	Automático (informatividade) e manual (coerência)
Caseli e Nunes (2006)	Segundo as autoras SIM, mas não foram apresentados resultados	Não	Não
Balage Filho *et al.* (2006)	Sim	Sim	Manual (a partir do julgamento humano, que diz se está correta, errada, imprecisa, incompleta ou ausente)
Leite e Rino (2006)	Sim	Sim	Automática
Moraes e Strube De Lima (2007)	Sim	Sim	Automático (resultado melhor depende da avaliar manual)
Kinoshita *et al.* (2007)	Sim	Sim	Manual (correto foi definido por um humano)
Specia, Stevenson e Nunes (2007)	Sim	Sim	Automático
Milidiu, Duarte e Cavalcante (2007)	Sim	Sim	Manual e automática
Caseli *et al.* (2008)	Não foram realizados experimento	Não	Não

Aziz, Pardo e Paraboni (2008)	Sim	Sim	Automática (escore BLUE) e manual (erros sintáticos e lexicais e word error rates – WER)
Morais e Ambrosio (2008)	Sim	Sim	automática (usando documentos que foram classificados manualmente)
Caminada, Quental e Garrao (2008)	Sim	Sim	Confuso (classificação Teste-T)
Seno e Nunes (2008)	Sim	Sim	automática (usando sentenças que foram classificadas manualmente)
Aziz, Pardo e Paraboni (2009)	Sim	Sim	automática
Seno e Nunes (2009)	Sim	Sim	manual (dois humanos)
Salles *et al.* (2009)	Sim	Sim	automática
Braga, Monard e Matsubara (2009)	Sim	Sim	automática (acurácia e taxa de erro médio)

Semeghini-Siqueira, Costa e Cohn (1986) afirmam que o sistema implementado utilizando Prolog tem como finalidade facilitar a consulta a uma base de dados relacional. Como resultado, os autores apresentam vários "tipos" de perguntas em português que ele é capaz de "entender" (p. 124). Além disso, os autores concluem afirmando que "para se mudar a base de conhecimento, não é necessário modificar o mecanismo da gramática, basta trocar o vocabulário e estabelecer novas conexões sintático-semânticas" (p. 125).

Ziviane e Albuquerque (1987) apresentaram um novo método para armazenamento de grandes volumes de dados utilizando um índice de assinaturas. Como resultados, os autores apresentaram algumas observações sobre o método descrito relacionadas ao tempo de resposta em relação ao número de palavras (p. 181).

Ripoll e Mendes (1988) apresentaram como resultado algumas observações a cerca do sistema proposto. Segundo os autores, a análise da frase baseada em casos verbais resolve muitos casos de ambiguidade léxica, porém existem algumas situações nas quais a informação da estrutura sintática auxiliaria no tratamento da ambiguidade (p. 303). Além disso, os autores complementam que, do ponto de vista semântico, o auxílio de contexto mais global do que o da frase

resolve ambiguidade que os casos verbais não resolvem (p. 303). Sendo assim, os autores sugerem como trabalhos futuros acrescentar um nível sintático (seguindo as ideias apresentadas em Selman (1985) e Waltz (1985) e interligar os substantivos com uma rede semântica que forneceria o contexto mais global da frase.

Em **Fusaro e Ziviani (1989),** os autores apresentaram algumas observações sobre a linguagem proposta afirmando apenas que "uma versão simplificada da linguagem foi implementada no sistema PATPLUS com excelentes resultados" (p. 297).

Strube de Lima (1990) apresentou uma revisão de literatura sobre métodos e técnicas empregadas a correção ortográfica automática e portanto não apresentou nem experimentos, nem resultados práticos. Como discussão final, a autora conclui que "seguramente, nem todos os erros de sintaxe podem ser corrigidos sem levarmos em conta informações semânticas. Muitos deles nem poderão ser detectados sem um tratamento semânticos (...)" (pág 52). E complementa que a detecção e a correção de erros a nível semântico permanecem como um vasto campo de pesquisa do qual existem, segundo a autora, pouquíssimos resultados. Existe até o presente (na publicação do artigo), o tratamento semântico de subconjuntos da língua, inseridos em aplicações específicas, podendo-se visualizar, segundo a autora, para um futuro próximo, o aproveitamento de alguns conhecimentos semânticos buscando aprimorar os mecanismos de detecção e correção de erros (p. 53).

Leffa (1991) apresentou algumas estatísticas sobre cobertura do dicionário elaborado para diversas áreas de conhecimento, o que sugere que os 4.700 verbetes selecionados pelo critério de frequência, tinham condições de proporcionar um bom nível de compreensão ao leitor. Além disso, o autor apresentou dados qualitativos sobre a avaliação da utilização do dicionário pelos alunos selecionados (p. 195)

Rocha *et al.* (1992) concluíram que o conhecimento adquirido pelo sistema desenvolvido em bases de dados especialistas pode ser usado para construir sistemas especialistas na área médica.

Rocha, Rocha e Huff (1993) destacam que existe um esforço em traduzir os dados do dicionário *Iliad (OpenClinical AI Systems in clinical practice*) em termos do vocabulário alvo *UMLS Metathesaurus.* Sendo assim, procurou-se comparar o desempenho do sistema proposto a um método manual. O principal objetivo foi

verificar se o sistema foi pelo menos comparável à revisão manual realizada por um médico. Foram realizados experimentos com 150 termos distintos. Segundo os autores, a comparação entre o sistema proposto e a tradução manual não apresentou diferença significativa (estatisticamente). Os autores concluem ainda que, o esforço manual para se criar o léxico é certamente a fase mais trabalhosa do projeto, mas que ele é seguramente o centro de todo o sistema e que o sucesso depende dele (p. 693).

Robin (1994) apresenta vários exemplos de sentenças geradas sem e com a aplicação das operações de revisão proposta no modelo. Além disso, o autor discute a cerca da portabilidade deste modelo para outro domínio de conhecimento (diferente de basquete – esportes). Segundo o autor, os resultados sugerem que o modelo de geração, assim como os dados linguísticos usados neste trabalho, podem ser reutilizados em sistemas de sumarização de textos de qualquer assunto.

Julia, Seabra e Semeghini-Siqueira (1995) não apresentaram experimentos envolvendo exemplos do analisador desenvolvido, e sim inúmeros exemplos de regras de produção utilizadas (expressões *lambdas*) (p. 809 e 810). Como conclusão, os autores destacam que a principal contribuição do trabalho é o fato de que "a técnica utilizada neste analisador pode realmente fazer a tarefa do linguista de definir a gramática mais fácil, pois permite que o mesmo defina apenas as abstrações *lambda* que devem ser associadas ao verbo e às categorias da gramática" (p. 811).

Barros (1996) não apresentou experimentos envolvendo exemplos do modelo desenvolvido, e sim um exemplo detalhado de como uma consulta (*query*) seria processada.

Em **Rosa (1997),** a análise de desempenho do sistema considerou sentenças válidas, ou seja, que deveriam ser aceitas pelo sistema, e sentenças inválidas, ou seja, que deveriam ser rejeitadas. Das 6.000 sentenças válidas, o sistema rejeitou apenas 5, enquanto que das cerca de 3.000 frases inválidas, o sistema aceitou 26. Diante disso, o autor conclui que o desempenho do sistema é muito bom, mas faz uma ressalva: "para os tipos de frases para o qual foi treinado" (p. 243). Além disso, o autor destaca que a abordagem conexionista já provou ser eficaz no tratamento de um pequeno conjunto de construções lexicais em português (e cita um trabalho anterior de sua autoria datado de 1994).

Em **Oliveira e Wazlawick (1998),** duas abordagens conexionistas foram

apresentadas para a resolução de anáfora em segmentos de texto com mais de duas frases. Na primeira abordagem uma rede simples recorrente é treinada com um subconjunto de segmentos de textos gerados artificialmente. A rede aprende os exemplos que lhe foram apresentados com a mesma estrutura com a qual foi treinada. Na segunda abordagem, o modelo foi alterado para uma rede multi-camadas *feedfoward* que permitiu generalizar para sentenças de tamanho arbitrário (p. 1198). O sistema foi treinado com texto contendo poucas sentenças. Apesar de não terem apresentado índices de acertos, os autores afirmam que a abordagem proposta para resolução de anáfora "resolve eficientemente todos os exemplos apresentados que contêm a mesma estrutura dos que foram treinados" (p. 1.198).

Carvalho e Strube de Lima (1999) apresentaram como resultados algumas comparações entre os modelos construídos. Segundo elas, quanto à distribuição do conhecimento, no modelo léxico-categorial, os agentes são simples, parecendo-se com agentes reativos e contém basicamente conhecimento gramatical. Conhecimentos sintático e semântico estão contidos em um dicionário que deve estar completo em termos de conhecimento sobre as palavras da língua. Esse conhecimento é externo aos agentes. No modelo linguístico-cognitivo, o conhecimento está distribuído entre os agentes, que contém dicionários e analisadores relacionados a cada uma das fases do processamento. Além disso, as autoras complementam que o agente semântico contém um dicionário semântico baseado na semântica léxico-gerativa de Pustejovski (de 1995) e usa os mecanismos gerativos dessa teoria, já os agentes associados a fenômenos utilizam heurísticas do domínio.

Kinoshita (1999) conclui que apesar de existirem vários erros de tradução, é possível entender vários versos da Bíblia. E complementa que a tradução para o português apresentou resultados melhores que para o inglês, e justifica que a ordem das palavras no inglês é mais rígido que no português, e a correspondência das palavras entre o grego e o português é maior do que entre o grego e o inglês. Além disso, o autor destaca que a anotação de Strong pode ser usada para alinhamento de corpus.

Barcia *et al.* (1999) apresentaram a proposta da utilização de Raciocínio baseado em casos na recuperação de textos jurídicos e não apresentaram experimentos envolvendo bases de dados. Como conclusão, os autores destacam que o RBC oferece um potencial significativo para a recuperação inteligente de

documentos jurisprudenciais. Ainda segundo os autores, seus principais benefícios são "o enfoque no conhecimento em forma de episódios individuais, em lugar de conhecimento de domínio genérico e a recuperação baseada na similaridade" (p. 6). Além disso, os autores colocam como contunuidade do trabalho, a validação da aplicação da abordagem proposta.

Berber Sardinha (1999) apresentou alguns estudos de caso considerados relevantes para a área. O corpus usado para os estudos de casos é constituído de mais de 32 milhões de palavras oriundas de jornais, conversas informais, artigos acadêmicos e outros. Quatro estudos foram relatados e selecionados por serem potencialmente relevantes. Para cada estudo de caso, o autor, baseando-se nas medidas estatísticas definidas, discute a perfil semântico mais frequente. Segundo o autor, as três estatísticas mais empregadas no estudo de colocação são a razão entre observado e esperado, a informação mútua e o escore T. O autor conclui que "o emprego de um corpus e de ferramentas computacionais propicia maior consistência e abrangência na análise". Além disso, o autor ressalta que os resultados apresentados indicam que "a metodologia empregada para a descrição dos perfis semânticos é versátil, podendo ser aplicada para estudos exploratórios e contrastivos".

Villavicencio (1999) afirma que como resultado da utilização do padrão de herança para representar informações sobre subcategorização verbal é possível obter uma hierarquia altamente estruturada e sucinta. E conclui que, comparando o modelo proposto com outros encontrados na literatura, ele evita a necessidade de especificar e declarar tipos redundantes.

Jose Neto e Menezes (2000) destacam que no primeiro experimento, a taxa de acerto obtida (82,81%) é comparável ao relatado em outros trabalhos da época. Já no segundo experimento, os autores afirmam que obteve-se bom desempenho final, chegando-se aos 90%. No entanto, os autores ressaltam que outros autores argumentam que o método baseado em exemplos memorizados começa a produzir resultados satisfatórios a partir de um corpus com 300.000 palavras.

Berber Sardinha (2000) apresentou os resultados que indicam que as prosódias semânticas podem variar entre o português e o inglês. O presente estudo corrobora, portanto, outros estudos contrastivos, que identificaram discrepâncias entre itens equivalentes de línguas diferentes. Segundo o autor, para evitar erros, o

tradutor deveria ter acesso à informação sobre a prosódia semântica da língua-alvo. Além disso, segundo o autor, a informação sobre prosódia semântica, embora valiosa, não está documentada em materiais de referência. O dicionário consultado, por exemplo, embora de prestígio, não incluía informação conotacional. Diante disso, o autor conclui que em uma pesquisa futura poderia considerar a questão de como um sistema de tradução automática poderia beneficiar-se da prosódia semântica, além de como tornar automática a aquisição de informação sobre prosódia semântica. Outras questões que ficam para estudos posteriores, segundo o autor, seriam sobre até que ponto é factível incluir-se informação conotacional em todos os verbetes de um dicionário ou glossário e sobre como escolher e selecionar itens para inclusão.

Padilha e Viccari (2000) apresentaram um trabalho teórico sem a realização de experimentos. Segundo os autores, os transdutores são "sem dúvida adequados para o processamento morfológico" (p. 51). No entanto, os autores afirmam que duas limitações devem ser ressaltadas: sua construção generativa (não há algoritmos de aprendizado de novas transformações, a gramática deve ser alterada e o transdutor reconstruído); e a ausência de pesos para diferenciar mapeamentos ambíguos. Apesar disto, os autores concluem que "a abordagem apresentada para o problema da morfologia parece teoricamente mais adequada, abrangente e flexível do que a empregada em aplicações para resolver problemas específicos, como texto-para-fala somente" (p. 51).

Em **Larocca Neto *et al.* (2000),** o sistema foi treinado e testado usando uma base de documentos contendo textos em inglês de revistas sobre computação, hardware, software, etc. Para o experimento realizado no trabalho, a base utilizada foi dividida em dois subconjuntos de 100 e 800 documentos. Todos os experimentos realizados foram avaliados usando um procedimento de validação cruzada (*10-fold cross-validation*), e dois algoritmos de classificação: Naive-Bayes e o C4.5. Além disso, os autores destacam que, é usual avaliar os resultados, em recuperação da informação, utilizando-se as medidas revocação (*recall*) e precisão (*precision*). No entanto, no escopo deste trabalho, assim como em outros experimentos de sumarizadores treináveis, o número de sentenças retornadas é igual ao número de sentenças do sumário, e portanto revocação = precisão = taxa de acurácia, e sendo assim, somente esta última medida foi usada. Para os experimentos iniciais (usando 100 documentos), obteve-se cerca de 38% de acerto, usando todas as

características extraídas. Os autores procuraram, usando o algoritmo C4.5, quais as características mais relevantes. Usando as três características que apresentaram maior desempenho (ocorrência de nomes próprios, conectividade das sentenças e indicador de conceitos principais), a taxa de acerto neste primeiro experimento, alcançou cerca de 49%. Para o segundo subconjunto de 800 documentos, o sistema apresentou cerca de 37% de acerto, usando todas as características, e 50% usando somente as três características de maior desempenho. Como trabalho futuro, os autores afirmam que o sistema deveria ser avaliado por julgamento humano.

Dias-da-Silva *et al.* (2000) apresentaram algumas telas do editor de thesaurus construído ilustrando algumas entradas fornecidas. Os autores concluem afirmando que são inúmeros os ganhos acumulados durante o desenvolvimento do thesaurus eletrônico para o português. Complementam afirmando que "há que se ressaltar a salutar troca de experiências e o profícuo exercício de construção do diálogo necessário, cooperativo e colaborativo entre linguistas e informatas" (p. 9). Além disso, os autores finalizam afirmando que "os especialistas da computação tiveram a oportunidade de apreciar com maior profundidade os resistentes problemas postos pela linguagem humana, que parece resistir a qualquer tentativa de ser reduzida a um código de máquina" (p. 9).

Rossi *et al.* (2001) desenvolveram uma interface para permitir a resolução manual de correferência de descrições definidas, em corpus da língua portuguesa. Através dessa interface, as descrições definidas (sintagmas nominais que começam com artigos definidos – a, o, as, os) são classificadas em nova no discurso, anáfora direta, anáfora indireta, associativa ou não classificada. A partir dessa análise manual pôde-se estudar a concordância entre falantes nativos (três pesquisadores que realizaram simultaneamente a análise do corpus) em relação ao processo de interpretação de textos envolvendo correferência, confirmando índices de concordância obtidos anteriormente em estudos da Língua Inglesa. Os resultados encontrados por essa anotação manual também foram usados no desenvolvimento de um processo automático de resolução de correferência nominal em Prolog (em andamento). Em síntese, para a anotação manual, foram apresentados resultados relacionados com a distribuição das classificações obtidas, assim como algumas discussões quanto a concordância entre os três avaliadores humano. Quanto ao sistema desenvolvido, não foram apresentados resultados, pois o sistema encontra-se em desenvolvimento.

Gamallo, Agustini e Lopes (2001) apresentou uma estratégia não supervisionada para adquirir restrições de seleção baseada em hipóteses de contexto e de co-especificação. A estratégia baseia-se principalmente em dois pressupostos linguísticos: a hipótese da co-especificação, ou seja, duas expressões relacionadas por uma dependência binária apresentam restrições semânticas entre si; e a hipótese contextual, ou seja, dois contextos sintáticos compartilham as mesmas restrições semânticas se co-ocorrerem com as mesmas palavras. Os autores afirmaram que faz parte do trabalho atual deles (na época de publicação do artigo), medir a eficiência da estratégia de aprendizagem. Apesar disso, os autores apresentaram como resultados, alguns exemplos de agrupamentos gerados, destacando como o sentido (*sense*) de palavras polissêmicas é representado pela atribuição natural da palavra em vários agrupamentos. Os autores concluem que o trabalho se difere dos anteriores em duas questões específicas: tanto na maneira de extrair similaridade das palavras (usando hipótese contextual) como na forma de definição de contextos sintáticos (hipótese da co-especificação).

Gonzalez e Strube de Lima (2001) apresentaram os resultados comparando a recuperação com e sem expansão automática de consulta, com o intuito de avaliar os benefícios do thesaurus proposto. O objetivo central desta avaliação é verificar o ganho obtido pela expansão automática de consulta, em termos de precisão e resposta, num sistema de RI, utilizando-se um thesaurus com estrutura semântica fundamentada na TLG. Segundo os autores, os resultados obtidos na avaliação indicam que a expansão de consulta pode trazer benefícios à RI. Entretanto, esta expansão não pode ser feita indiscriminadamente, num enfoque quantitativo, sob pena de prejudicar os resultados do mecanismo de busca. As curvas de precisão/resposta foram calculadas de acordo com os procedimentos adotados pela comunidade internacional nas *Text Retrieval Conferences (TRECs)*. Como trabalhos futuros, os autores alertam que se faz necessária analisar a influência de cada papel da estrutura Qualia na expansão; avaliar o comportamento dos verbos como termos que viabilizam vínculos entre termos expandidos; e examinar o desempenho de termos composicionais e periféricos na expansão (p. 8).

Souza, Pereira e Nunes (2001) compararam os sumários resultantes com o texto original, com o sumário feito pelo próprio autor do artigo científico e com o sumário feito pela ferramenta AutoResumo do Word. Avaliou-se o percentual de sumarização, ou seja, o número de sentenças do sumário dividido pelo número total

de sentenças do texto-fonte, o percentual de erros de coesão e coerência, dado pelo número de sentenças problemáticas dividido pelo número total de sentenças do sumário, e se o sumário gerado manteve a idéia principal do texto-fonte. Segundo os autores, dos 18 textos sumarizados pelo AutoResumo do Word, 11 não preservaram a ideia central do texto, sendo que usando duas das estratégias avaliadas, somente em três sumários a idéia principal não foi preservada. Com relação aos erros de coesão e coerência, observou-se que os resultados variaram de texto pra texto. Os autores não deixam claro como as avaliações foram feitas, mas sugere-se que tenha sido feita manualmente.

Orengo e Huyck (2001) compararam a saída do algoritmo de remoção de sufixos implementado com a saída esperada, definida manualmente a partir de uma lista de palavras distintas. Utilizou-se o método de Paice, que determina o cálculo dos índices de *overstemming* (parte do radical é removida), e *understemming* (o sufixo não é removido totalmente). Ao comparar o algoritmo implementado com a versão para o português do algoritmo de Porter, os autores destacam que aquele produziram menos erros de *overstemming* e *understemming* que este.

Jose Neto e Moraes (2002) apresentam como exemplo ilustrativo, a construção de um autômato a partir de uma gramática que define uma aproximação livre de contexto de um pequeno subconjunto da língua portuguesa. Segundo os autores, para exemplificar, utilizou-se uma gramática, que "define grosseiramente alguns aspectos de uma linguagem natural (no caso, português), de uma forma puramente sintática, sem considerar os aspectos da dependência de contexto, que certamente deverão ser levados em conta em outras etapas do processamento" (p. 4). Os autores destacam que o método adotado para a construção de um autômato adaptativo a partir da gramática consiste em desenhar uma máquina de estados inicial que reconheça qualquer cadeia válida de símbolos representada pelo conjunto disponível de regras gramaticais. E complementam que não é usado nenhum método de construção de reconhecedores convencional a partir de gramáticas livres de contexto, mas explora-se a característica adaptativa do modelo de reconhecimento adotado. Finalmente, os autores afirmam que "é possível provar que o comportamento temporal desses autômatos é bastante adequado também em relação ao comprimento da cadeia de entrada, mesmo nos casos de ambiguidade e não-determinismo" (p. 6).

Bidarra (2002) apresenta alguns aspectos básicos para construção de

léxicos computacionais baseados em dados de parafasia semântica. O objetivo, segundo o autor, é discutir a pesquisa que o autor na sua tese de Doutorado, onde ele propôs um modelo de descrição lexical para dar suporte a questões relacionadas com patologias da linguagem e consequentemente modelos computacionais para este fim. Como resultados e conclusão, o autor afirma que ao longo do texto, ele tentou mostrar que o léxico desenvolvido é uma modelagem apropriada para a construção de sistemas de PLN, não apenas preocupado com a engenharia do produto mas, sobretudo, formalizado de acordo com os estudos realizados no âmbito de teorias (neuro) linguísticas envolvendo a linguagem humana. No entanto, ele reconhece que existem pendências e imprecisões, mas que o objetivo foi apresentar uma reflexão a respeito da importância de se elaborar sistemas (computacionais) que venham de algum modo contribuir para o avanço das pesquisas em linguagem. O autor conclui afirmando que o projeto encontra-se "atualmente" (na época da publicação desse artigo) em fase de especificação e implementação do léxico.

Em **Pardo e Rino (2002),** vários experimentos foram feitos para avaliar o DMSumm, enfocando, principalmente, as premissas básicas do sistema, isto é, a satisfação do objetivo comunicativo e a preservação da proposição central. Utilizou-se o Theses Corpus (PARDO, 2002[158]), contendo 10 introduções de teses e dissertações da área de computação, tendo, em média, 530 palavras cada introdução. Esse corpus foi escolhido pelo fato dos textos apresentarem a estrutura Problema-Solução e serem acompanhados por sumários autênticos, ou seja, aqueles produzidos pelos próprios autores dos textos. Foram considerados dois pontos de decisão principais (WHITE *et al.*, 2000[159]) para a avaliação dos sumários automáticos: textualidade e preservação da ideia principal. Os sumários automáticos e autênticos foram julgados por 10 juízes linguistas computacionais e falantes nativos do português do Brasil. Dentre os resultados do julgamento humano tem-se que 67% dos sumários automáticos mantiveram a textualidade, enquanto que 90% dos sumários autênticos mantiveram a textualidade; além disso, 61% dos sumários automáticos preservaram somente parcialmente a ideia principal e 31% a preservaram totalmente, enquanto todos os sumários autênticos preservaram

[158] Pardo, T.A.S. (2002). DMSumm: Um Gerador Automático de Sumários. Dissertação de Mestrado. Departamento de Computação. Universidade Federal de São Carlos. São Carlos – SP.
[159] White, J. S.; Doyon, J. B.; Talbott, S. W. (2000). Task Tolerance of MT Output in Integrated Text Processes. In ANLP/NAACL 2000: Embedded Machine Translation Systems, pp. 9-16. Seattle, WA.

totalmente a ideia principal. A medida Kappa foi calculada em 0.78, indicando a concordância entre os juízes humanos. Os autores avaliaram também a informatividade semântica dos sumários automáticos em relação aos autênticos, conforme sugerido por Mani (2001[160]), isto é, o quanto de informação foi reproduzida no sumário automático em relação a todo o conteúdo do texto-fonte. Segundo os autores, as medidas de precisão (*precision*), cobertura (*recall*[161]) e *f-measure* foram calculadas, sendo que no contexto da sumarização automática elas são definidas da seguinte forma: a precisão indica o quanto de informação do sumário autêntico, os sumários automáticos apresentaram em relação a tudo o que apresentaram, ou seja, indica o quão próximos os sumários automáticos estão dos autênticos; a cobertura indica o quanto de informação do sumário autêntico, os sumários automáticos apresentaram, ou seja, indica o grau de informatividade dos sumários automáticos em relação aos autênticos; a f-measure é uma distribuição da combinação da precisão e da cobertura, sendo, portanto, uma medida única de eficiência de um sistema. No caso da sumarização automática, ela indica o desempenho de um sistema em produzir sumários próximos dos ideais. Segundo os autores, o DMSumm produziu sumários com 44% de precisão e 54% de cobertura, com uma *f-measure* de 0,48.

Em **Schulz *et al*. (2002),** os autores destacam que a versão atual do MORPHOSAURUS tem aproximadamente 15 mil subwords, abrangendo a terminologia clínica em inglês, alemão e português, mas que a construção dos repositórios de nomes próprios e acrônimos ainda não havia sido abordada.

Bonfante e Nunes (2002) afirmam que não possuem ainda resultados concretos, pois os experimentos estão em andamento. No entanto, elas destacam que o fato de identificar previamente os sintagmas nominais de cada sentença facilita o processamento do *parser*, uma vez que evita gastar recursos computacionais avaliando as possíveis uniões improváveis.

Zavaglia (2003) não apresentou experimentos, apenas dois exemplos da representação do item homônimo 'banco'. Como considerações finais, a autora destaca que a estrutura Qualia do Léxico Gerativo serviu como estrutura representacional para expressar o significado lexical, e que a versão computacional do modelo desenvolvido (base de conhecimento lexical - BCL) encontra-se

[160] Mani, I. (2001). Automatic Summarization. John Benjamins Publishing Co., Amsterdam.
[161] Tradução adotada pelos próprios autores.

disponibilizada no NILC.

Martins, Monard e Matsubara (2003) utilizaram três algoritmos de aprendizado: dois algoritmos simbólicos de regras de associação; e um baseado em técnicas estatísticas de aprendizado – *Support Vector Machines (SVM)* para ilustrar o uso de redução de dimensionalidade do conjunto de treinamento, implementada no PreText. Os resultados obtidos demonstraram que, para a coleção usada, o algoritmo SVM apresentou desempenho melhor que os dois outros algoritmos usados e erro muito próximo de zero quando utilizou-se um número reduzido de atributos.

Em **Pardo, Rino e Nunes (2003),** os autores destacam que a avaliação do NeuralSumm foi realizada objetivando medir o desempenho da rede neural do tipo SOM em classificar sentenças corretamente como essenciais, complementares e supérfluas e, num segundo momento, verificar a proximidade dos extratos gerados automaticamente com seus sumários autênticos, isto é, aqueles produzidos pelos próprios autores dos textos-fonte. Para treinamento e teste, foram utilizadas as sentenças do CorpusDT: 10 teses e dissertações cujas as sentenças foram classificadas manualmente por juízes humanos. Para efeito de comparação, utilizou-se também os classificadores Naive-Bayes e C4.5. A rede neural do NeuralSumm obteve a menor taxa de erro (41%) em relação aos outros classificadores (51% a 57%). Segundo os autores, utilizou-se a validação cruzada (10-*fold cross-validation)*, pelo fato do corpus ser pequeno. No entanto, os autores ressaltam que a classificação humana de sentenças para compor um extrato é muito subjetiva, pois depende de vários aspectos relacionados, por exemplo, com o conhecimento prévio do leitor, com o tempo disponível para a leitura e com a intenção comunicativa percebida pelo leitor. Assim, para permitir uma maior flexibilidade na classificação, assumiu-se que as sentenças complementares também podem ser classificadas como essenciais e as supérfluas também podem ser classificadas como complementares. Para este novo experimento, as redes neurais alcançaram 27% de erro. Para verificar a proximidade dos extratos gerados automaticamente com seus sumários autênticos, os autores ressaltam que comparar os sumários autênticos, ou seja, os elaborados pelos próprios autores seria uma tarefa difícil, visto que os autores tendem a não preservar as sentenças dos textos-fonte. Assim, para resolver este problema, os autores destacam que tem-se adotado os sumários ideais *(gold-standards*), que consistem na versão extrativa dos sumários autênticos. Para se

produzir o sumário ideal[162], costuma-se utilizar a medida do co-seno (SALTON, 1989[163]): para cada sentença do sumário autêntico, procura-se a sentença correspondente no texto-fonte mais semelhante. Para essa avaliação, foram coletados outros 10 textos científicos (introduções de teses e dissertações) da computação (não contendo nenhum dos textos utilizados anteriormente) com seus respectivos sumários autênticos, para os quais foram gerados automaticamente os sumários ideais. A cobertura (*recall*) indica o número de sentenças do extrato que coincidem com as do sumário ideal; e a precisão indica a razão entre o número de sentenças coincidentes com as do sumário ideal e o total de sentenças do extrato. Os resultados obtidos para cobertura e precisão foram, respectivamente, de 32% e 41%, mas segundo os autores, apesar de aparentemente serem insatisfatórios, não indicam que os extratos produzidos pelo NeuralSumm sejam ruins. Os autores afirmam que os valores obtidos estão relativamente próximos dos valores que outros trabalhos obtiveram quando fizeram avaliação similar.

Em **Gasperin e Strube de Lima (2003),** a estratégia de teste realizada consistiu em para cada palavra inserida pelo usuário na consulta, percorrer a lista gerada automaticamente, procurando por novas palavras semanticamente relacionadas que poderiam ser incluídas na consulta; submeter ambas as consultas, original e expandida, a uma base de documentos; e verificar as medidas de precisão e revocação obtidas. Foram realizadas 7 consultas (nas versões original e expandida), e observou-se que com a consulta expandida, a revocação aumenta, enquanto que a precisão diminui, se comparado com a consulta original. As autoras finalizam afirmando que “o conhecimento semântico de uma lista gerada automaticamente pode melhorar a recuperação por meio da expansão de consulta” (p. 7).

Em **Alves e Chishman (2004),** como resultado da reorganização da nomenclatura usada no tratamento e abordagem da ambiguidade, as autoras propõem que na ambiguidade semântica lexical compreende os casos de ambiguidade que têm origem no léxico e podem ser polissemia, homonímia, vagueza ou vaguidade, e, uso conotativo da linguagem. Dentre os tipos de ambiguidades não lexicais estão a ambiguidade sintática ou estrutural e a

[162] Segundo os autores, o programa utilizado para gerar os sumários ideais encontra-se disponível para download em http://www.nilc.icmc.usp.br/~thiago/NeuralSumm.html

[163] Salton, G. (1989) Automatic Text Processing. The Transformation, Analysis and Retrieval of Information by Computer. Addison-Wesley.

pragmática ou aberta. Como resultado da análise do desempenho dos tradutores avaliados diante de casos de ambiguidades, as autoras afirmam que de uma maneira geral, os tradutores geralmente não percebem a tradução mais adequada para o contexto, e não indicam que pode haver outra possibilidade de tradução. Como exemplo, apresentaram os resultados obtidos para a palavra 'canto'. As autoras concluem que apesar da tradução automática ter sido a primeira aplicação não numérica da computação (datada de 1949), o desempenho desses sistemas ainda está muito longe do que seria ideal.

Oliveira, Garrao e Amaral (2003) apresentaram como resultado uma tabela listando as locuções preposicionais encontradas no corpus. As ocorrências foram organizadas de acordo com o critério: satisfaz totalmente, parcialmente e outras expressões frequentes co-ocorrentes. Segundo os autores, este trabalho realizou inicialmente uma pesquisa em corpus e que o critério, detalhado pelas ocorrências no corpus, deve ser implementado computacionalmente quando será formalizado. Os autores destacam que a grande dificuldade é encontrar os casos onde os fatores semânticos são determinando para identificar as locuções preposicionais.

Specia e Nunes (2004) apresentaram como justificativa para a escolha dos verbos problemáticos, a serem abordados no presente trabalho, os resultados de experimentos realizados em um projeto anterior (SPECIA; NUNES, 2004[164]). Não foram realizados experimentos, pois as autoras apresentaram a proposta de construção de um modelo que está em fase de especificação, que deverá ser implementado usando algum ambiente de programação lógica indutiva. As autoras apresentam como diferencial com relação aos trabalhos existentes, além da aplicação para o português, a utilização de um formalismo de representação do conhecimento e dos exemplos de desambiguação baseado na Lógica de Primeira Ordem, mais expressivo que o proposicional e segundo as autoras, ainda não explorado em quaisquer aplicações de desambiguação lexical de sentido, mesmo nas monolíngues.

Em **Rino *et al.* (2004),** os autores realizaram uma avaliação do tipo caixa-preta, ou seja, considerando somente as saídas dos sistemas. Além disso, a avaliação dos cinco sistemas analisados foi realizada comparando-se os sumários

[164] Specia, L. and Nunes, M.G.V. (2004) “A ambiguidade lexical de sentido na tradução do inglês para o português – um recorte de verbos problemáticos”, Série de Relatórios do NILC, NILC-TR-04-01, São Carlos, Março, 30p.

produzidos com os ideais (formados pelas sentenças do texto-fonte mais similares as sentenças do sumário feito pelo autor). Analisando-se as medidas precisão, revocação e f-measure, os sistemas SuPor e ClassSumm apresentaram os melhores resultados 42,8% e 42,4% de *f-mesasure*, respectivamente. O sistema NeuralSumm apresentou o pior resultado (31% de *f-measure*). Os autores destacam que a proximidade nos resultados do SuPor e do ClassSumm pode ser explicado pela relação existente entre algumas características usadas: no primeiro frequência de palavras e frases sinalizadas, e no segundo, o TF-ISF médio, o indicador de conceitos principais e a similaridade do título. Isto é justificado pois o TF_ISF médio é baseado na frequência das palavras, enquanto que os conceitos principais e o título podem apresentar sinalizadores das frases.

Aluisio *et al.* (2004) apresentam o Lácio-Web como sendo um repositório de recursos para o desenvolvimento de pesquisas da língua portuguesa do Brasil e de outras ferramentas linguísticas e computacionais. Como resultados, os autores apresentaram várias funcionalidades e requisitos do Lácio-Web, envolvendo tipos de buscas e recursos armazenados. Segundo os autores, a primeira versão do Lácio-Web, lançada em janeiro de 2004, disponibilizou o Lácio-Ref (com mais de 4 milhões de palavras distribuídas em textos de vários gêneros, tipos e domínios) e o Mac-Morpho (com mais de um milhão de palavras de textos jornalísticos da Folha de São Paulo). OS textos do Mac-Morpho foram automaticamente anotados pelo parser Palavras de Bick, e revisado manualmente.

Matsubara, Monard e Batista (2004) destacam que, ao analisar o número de exemplos rotulados errado e o erro dos classificadores induzidos, é possível concluir que o *co-training* atingiu excelentes resultados com o conjunto news (0,5%) e resultados muito bons com o conjunto lnai (8,7%). Além disso, os autores destacam que foi utilizada uma quantidade muito pequena de exemplos com seus rótulos originais, somente 5%, o que comprova que o algoritmo *co-training,* juntamente com a abordagem proposta, pode ser muito efetivo nos casos em que se possui apenas um pequeno conjunto de exemplos rotulados, e o custo para rotular mais exemplos é alto. Os autores propõem como trabalhos futuros, verificar o uso de diferentes indutores, como *Support Vector Machines*, na construção dos classificadores.

Em **Pardo, Marcu e Nunes (2005),** para avaliar se as estruturas argumentais aprendidas são plausíveis ou não, dois experimentos foram realizados.

No primeiro experimento, foram avaliadas as 20 estruturas argumentais mais prováveis aprendidas pelo modelo proposto, para três verbos escolhidos aleatoriamente. Estas estruturas foram apresentadas a três humanos linguistas computacionais para, independentemente, julgá-las em termos de sua corretude/plausibilidade. Cerca de 90% das estruturas foram julgadas corretas pelos juízes (estatística Kappa de 0.77, indicando concordância entre eles). No segundo experimento, para um conjunto de 20 verbos escolhidos aleatoriamente (incluindo os três anteriores), as estruturas argumentais foram comparadas com as estruturas previstas pelo PropBank (repositório construído manualmente para a especificação semântica dos verbos). As medidas precisão e cobertura foram redefinidas neste contexto, e utilizadas para avaliar o modelo. Como aspectos positivos do modelo, os autores afirmam que o modelo é capaz de aprender estruturas argumentais com grande precisão, sem esforço de anotação, usando ferramentas relativamente simples. No entanto, ele não é capaz de lidar apropriadamente com sintagmas verbais, advérbios e complementos verbais sentenciais complexos, e que estas limitações serão objeto de pesquisas futuras. Os autores finalizam apresentando o repositório de estruturas argumentais aprendidas para os 1.500 verbos mais frequentes do inglês (ArgBank[165]), e que como próximo passo deste trabalho, um repositório semelhante deve ser produzido para o português brasileiro, utilizando-se o Corpus NILC (PINHEIRO; ALUÍSIO, 2003[166]).

Caseli, Nunes e Forcada (2005) construíram manualmente um alinhamento de referência a partir de 20 pares de textos paralelos selecionados aleatoriamente do corpus português-espanhol da Fapesp (CorpusFAPESP). Os termos dde corpus de referência foram alinhados por dois anotadores bilíngues seguindo os princípios definidos em Caseli *et al.* (2005[167]). A maioria dos alinhamentos era do tipo 1:1 (83%), além dos casos de omissão (quase 7%) e sentenças (quase 10%). Esse alinhamento de referência foi usado para avaliar automaticamente o produzido pelo LIHLA, usando as medidas precisão, revocação e taxa de erro de alinhamento (AER). Experimentos similares foram feitos a partir de corpus paralelo português-inglês considerando 10 pares de textos selecionados aleatoriamente. Assim, o LIHLA obteve precisão entre 84-92% e revocação entre 76-

[165] Disponível em http://www.nilc.icmc.usp.br/~thiago/ArgBank/index.html
[166] Pinheiro, G.M. e Aluísio, S.M. (2003). Corpus NILC: Descrição e Análise Crítica com Vistas ao Projeto Lacio-Web. Série de Relatórios Técnicos do Instituto de Ciencias Matemáticas e de Computação – ICMC, Universidade de São Paulo, N. 190.
[167] Caseli, H. M., Scalco, M. A. G., and Nunes, M. G. V. (2005). Manual para anotação de alinhamentos lexicais. Série de Relatórios do ICMC 256 (NILC-TR-05-09), NILC, www.nilc.icmc.usp.br/nilc/download/NILC-TR-05-09.pdf.

88%, alcançando taxas de erro entre 9-19%. Os melhores resultados foram obtidos no primeiro experimento, quando utilizou-se corpus paralelo português-inglês. Os autores concluem afirmando que a grande contribuição do LIHLA é o fato de ser baseado em heurística independente de linguagem e por isto pode ser aplicado a outro par de linguagens sem qualquer modificação.

Em **Specia, Nunes e Stevenson (2005),** utilizou-se como medidas objetivas para avaliar a tarefa de desambiguação lexical de sentido: erro (versão negativa da precisão), cobertura (*coverage*), suporte (mesmo que revocação) e *novelty* (relação entre a premissa e a conclusão da regra), dentre as opções fornecidas pelo sistema Rulee. A avaliação foi dividida em dois passos: primeiro, essas medidas objetivas são aplicadas com o intuito de reduzir o número de regras, e então, as resultantes eram manualmente analisadas, com o intuito de selecionar aquelas que forem consideradas interessantes para o processo de desambiguação lexical de sentido. Ao analisar as regras individualmente, observou-se que os melhores resultados foram obtidos utilizando como fonte de conhecimento, os lemas da primeira e da segunda palavra, a esquerda e a direita do verbo, o primeiro nome, o primeiro adjetivo, o primeiro verbo a esquerda e o a direita do verbo, e a primeira preposição a direita do verbo, apresentou os melhores resultados. Os autores concluem que as regras obtidas foram analisadas por critérios objetivos e subjetivos, e que as regras de maior qualidade podem ser usadas como fonte de conhecimento em um sistema relacional de desambiguação lexical de sentido, o que segundo os autores, é inédito.

Silva, Vieira e Osorio (2005) realizaram inúmeros experimentos, alternando-se os parâmetros dos algoritmos de aprendizado de máquina utilizados, o número de termos usados, o método de pré-processamento avaliado, dentre outro. Segundo os autores, na tarefa de classificação, ao utilizar informações linguísticas (categorias gramaticais) para selecionar os termos indexadores dos documentos, os melhores resultados foram obtidos ao utilizar os nomes, sendo que a melhor taxa de acerto foi obtida quando utilizou-se nomes juntamente com os adjetivos (erro de 18%). No entanto, os autores destacam que o segundo melhor resultado (erro de 19,7%) foi observado quando utilizou-se o método tradicional (*stopwords* e *stemming*). Na tarefa de clusterização, a precisão aumentou, aproximadamente, de 50% (abordagem tradicional) para 63% (usando informação linguística).

Piltcher *et al.* (2005) realizaram três experimentos alternando-se o

dicionário de referência utilizado: histórico das sessões, artigos da biblioteca ou a ontologia. Para todos os experimentos, utilizou-se como entrada uma lista de 7.652 palavras, que foi analisada manualmente, e constatou-se que 2.976 delas precisavam de correção. Utilizando-se o histórico de sessões, observou-se que a precisão (razão entre termos corretos e corrigidos) ficou em 30% para o limiar de similaridade superior a 72%, e, aumentou para cerca de 80%, ao considerar o limiar em 84%. No entanto, em ambos os casos, a abrangência (razão entre termos corretos e correções esperadas) ficou muito baixa (3%), o que demostra, segundo os autores, que esta base não é confiável. Utilizando-se os documentos da biblioteca digital, observou-se que a precisão ficou em 30% (limiar de 72%) e 50% (limiar de 84%), mas que em ambos os casos, os resultados eram melhores quando mais documentos eram usados. Ao avaliar a ontologia, observou-se resultados melhores quando a ontologia havia passado por intervenção humana. Segundo os autores, a abrangência esta diretamente ligada à qualidade dos termos adotados para o dicionário, enquanto que a quantidade contribui para a melhoria da precisão. Analisando, em separado, as três métricas utilizadas, observou-se a técnica de Levenshtein apresentou precisão bem aquém das demais técnicas, apesar de ter apresentado a melhor abrangência. As demais técnicas, mesmo tendo uma boa precisão, obtiveram uma abrangência pouco significativa (entre 3% e 7%).

Rino e Seno (2006) afirmam que, em um trabalho anterior, o RHeSumaRST foi avaliado sob duas perspectivas sugeridas nas *Document Undestanding Conference (DUC)*: informatividade, que visa verificar se as heurísticas permitiam preservar as informações mais relevantes do texto-fonte; e coerência, que visa verificar se as heurísticas garantiam a inexistência de quebra de cadeias de co-referências. Para calcular a informatividade, foi usada a ferramenta ROUGE (LIN, 2004[168]; 2004[169]), que compara a informatividade de sumários gerados por sumarizadores diversos. Foram utilizados dois outros sumarizadores automáticos: o de Marcu (1997[170]) e um *baseline*, cujos sumários são construídos pela poda de todo satélite das estruturas RST. Utilizando-se os 10 textos extraídos do TeMario, o RHeSumaRST foi mais informativo que o *baseline*, porém menos que

[168] LIN, C. ROUGE: a Package for Automatic Evaluation of Summaries. In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), Barcelona, Spain, 2004.
[169] LIN, C. Looking for a Few Good Metrics: Automatic Summarization Evaluation - How Many Samples Are Enough?. In Proceedings of the NTCIR Workshop 4, Tokyo, Japan, 2004.
[170] MARCU, D. 1997. The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts. PhD Thesis, Department of Computer Science, University of Toronto.

o modelo de Marcu. A avaliação da coerência das estruturas RST, por sua vez, foi mais complicada, devido à necessidade de comparação manual: cada sumário produzido pelos mesmos três sistemas foi comparado com seu correspondente texto-fonte, anotado manualmente com as cadeias de co-referência. Desse modo, foi possível identificar as quebras de cadeias de co-referência nos sumários, para os casos em que a anáfora era comprovada no texto-fonte (isto é, quando não era uma referência nova). Segundo as autoras, o RheSumaRST apresentou o menor índice de quebra, quando comparado aos outros sistemas, mas que a diferença não justifica o esforço necessário de modelagem e processamento estrutural. As autoras concluem que à época, o estado da Sumarização Automática no Brasil, era caracterizado pela inexistência de recursos dicionarizados sofisticados (sobretudo ontológicos) para a adoção expressiva de métodos empíricos, e que por este motivo, a resolução profunda do RheSumaRST é promissora pela simplicidade de elaboração dos algoritmos de detecção das unidades elementares de discurso (orações demarcadas por sinais de pontuação) e de reconhecimento das veias (um conjunto de unidades do discurso que podem conter o antecedente de uma anáfora) de estruturas RST. Ainda segundo as autoras, o trabalho braçal do sistema (anotação manual das cadeiras de co-referência) e das estruturas retóricas pode ser superado com a associação ao DiZer (PARDO, 2005[171]), um analisador discursivo baseado no mesmo modelo de Marcu, mas voltado ao processamento de textos em português.

Caseli e Nunes (2006) alegaram que experimentos já estão sendo feitos, mas não apresentaram resultados de simulações ou exemplos de regras produzidas. Como conclusão, as autoras afirmam que “este artigo traz uma breve descrição do procedimento desenvolvido no projeto ReTraTos para indução de regras de transferência e dicionário bilíngue” (p. 9). Ainda segundo as autoras, a próxima etapa consistem em induzir regras para tipos de alinhamento e categorias (POS) separadamente, e estender o método para tradução entre português e inglês.

Balage Filho *et al.* (2006) afirmam que a principal métrica de avaliação usada no CLEF é a acurácia (*accuracy*), definida a partir do julgamento humano, que aponta se a resposta está correta, errada, imprecisa, incompleta ou ausente. Segundo os autores, algumas variações da acurácia também são usadas, tais como

[171] PARDO, T.A.S. 2005. Métodos para Análise Discursiva Automática. Tese de Doutorado. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos-SP, Junho, 211p.

*Confidence Weighted Score (*CWS), *Mean Reciprocal Rank Score (*MRRS) e a K1. Os autores afirmam que “os resultados obtidos para ambos os experimentos são muito pobres” (p. 374), pois a grande maioria das respostas foi sinalizada como incorretas. Diante disso, os autores apresentam inúmeras questões que deveriam ser investigadas no futuro, e concluem que “após os experimentos realizados no CLEF, acredita-se que técnicas de sumarização simples não são suficientes para a tarefa de responder perguntas, apesar de serem eficientes para o que eles se propõem” (p. 375).

Enembreck *et al.* (2006) compararam os membros sugeridos pelo sistema com os membros reais das dissertações de mestrado selecionadas. Segundo os autores, normalmente existem três avaliadores internos, o orientador e um membro externo. Assim, a avaliação foi feita usando duas medidas: comparando somente o primeiro candidato sugerido pelo sistema com o orientador; e comparando a equipe proposta pelo sistema com os três grupos de pesquisa. O sistema foi capaz de identificar cerca de 20% dos orientadores das dissertações, 75% dos grupos de pesquisas envolvidos no desenvolvimento da pesquisa. Analisando os casos para os quais o sistema não conseguiu identificar o grupo de pesquisa dos envolvidos, foi possível constatar que estes projetos caracterizavam temas novos, e consequentemente os membros ainda não apresentavam produção significativa. Descartando-se assim, estes casos, o sistema conseguiu identificar 22% dos orientadores e 88% dos grupos de pesquisa.

Leite e Rino (2006) utilizaram como medida de avaliação: a precisão, dada pelo número de sentenças relevantes divididas pelo tamanho do sumário (extrato); a revocação (*recall),* dada pelo número de sentenças relevantes dividido por tamanho do sumário ideal, e a medida F, obtida pela relação das duas anteriores. Segundo os autores, o classificador Naive-Bayes apresentou os melhores resultados, com medida F em torno de 45%. Ao comparar o sumarizador construído (Supor-v2) com outros sumarizadores, o mesmo apresentou resultados superiores aos demais. Os autores destacam que apesar da diferença ser de apenas 3% na medida F, o desempenho do sistema desenvolvido ainda sim é significante, considerando que conseguir tais melhorias é muito difícil. Além disso, os autores concluem que os três sumarizadores que apresentaram os melhores resultados (Supor-v2, Supor e o ClassSumm) utilizam como classificar o modelo Naive-Bayes, o que confirma a sua aplicabilidade.

Moraes e Strube de Lima (2007) destacam que, de uma maneira geral, os resultados obtidos foram muito ruim (precisão, revocação[172] e F1). Segundo elas, como os documentos utilizados não foram previamente rotulados, não há como distinguir certos documentos sem um processo manual que, dada a quantidade de textos, não foi realizado. As autoras concluem que "uma consequência imediata deste fato é a baixa precisão de classificação na maioria das categorias escolhidas"(p. 1665). Diante disso, as autoras finalizam afirmando que "os resultados apresentados ainda são preliminares. (…) É necessário também realizar uma avaliação manual dos resultados do categorizador, a fim de estudar sua eficácia principalmente em nível de subclasse: analisar, por exemplo, se os documentos classificados em Agricultura de fato pertencem a essa subcategoria"(p. 1666).

Kinoshita *et al*. (2007) destacam que o CoGrOO e o ReGra detectaram 7 erros em comum. CoGrOO detectou 8 erros que o Regra não detectou, por outro lado, o Regra detectou 7 que o CoGrOO não. Os autores concluem que a arquitetura do CoGrOO é híbrida e mescla o uso de regras simbólicas e estatísticas com aprendizado de máquina baseado em treinamento com corpus anotado (usando algoritmo de entropia máxima). Além disso, os autores destacam que a abordagem usada para identificar o sujeito e verbo das sentenças é inovadora, pois não foi encontrada na literatura, nenhuma metodologia similar.

Specia, Stevenson e Nunes (2007) destacam que na tarefa multilíngue a abordagem apresentou resultados superiores aos observados pelos outros algoritmos de aprendizado de máquina. Na tarefa monolíngue, os resultados foram comparáveis ao dos outros sistemas avaliados. Os autores concluem afirmando que os resultados confirmam a hipótese de que a programação lógica indutiva, para gerar regras expressivas, usada em conjunto com uma variedade de fonte de conhecimento, traz benefícios para sistemas de desambiguação lexical de sentido.

Silva e Vieira (2007) afirmam que os melhores resultados foram obtidos ao utilizar a combinação de substantivos com adjetivos, e substantivos, adjetivos e nomes próprios. Os experimentos mostraram que o algoritmo de árvore de decisão possui um desempenho melhor do que SVM para um número de termos reduzido, e estabiliza-se a partir de um certo ponto, enquanto que o SVM atinge melhores resultados consistentemente com o aumento do número de termos utilizados no aprendizado.

[172] As autoras mantiveram o termo recall sem traduzí-lo.

Milidiu, Duarte e Cavalcante (2007) afirmam que os três algoritmos analisados (cadeia de Markov, aprendizado baseado em transformações e *Support Vector Machine)* foram avaliados usam validação cruzada (*10 cross-validation*). Os autores definiram sete experimentos combinando estes métodos com a utilização de um sistema de referência (*baseline system)*, composto com nomes de localidades, personalidades e organizações extraídas da Web. Segundo os autores, as localidades foram mais facilmente reconhecidas com precisão média em torno de 93%, enquanto que as organizações são as mais difíceis de serem reconhecidas, com precisão média de 75%. Além disso, os autores destacam que os algoritmos SVM e o baseado em transformações mostraram-se excelentes alternativas para reconhecimento de nomes próprios quando for possível contar com um sistema de referência, como o usado neste trabalho. No entanto, destacam ainda que os resultados obtidos sem a utilização do sistema de referência representam uma "solução pobre". Os autores afirmam que os resultados obtidos (88% de medida F), utilizando-se SVM e o sistema de referência, foram melhores que os obtidos pelo conhecido PALAVRAS-NER[173] (que obteve 80.6%).

Caseli *et al.* (2008) citam os resultados obtidos em experimentos obtidos em trabalhos anteriores (CASELI *ET AL.,* 2005[174]; CASELI *ET AL.,* 2005[175]), e reafirmam que o método LIHLA alcançou precisão entre 84 e 92%, e revocação entre 76 e 91%, para Português-Inglês e Português-Espanhol, respectivamente. Como resultado do presente trabalho, os autores apresentam a interface gráfica da ferramenta VisualLIHLA[176], e descrevem o funcionamento da mesma.

Aziz, Pardo e Paraboni (2008) compararam os escores BLUE[177] do método estatístico proposto neste trabalho, com os obtidos pelo sistema baseado em regras Apertium[178], e observou-se que os resultados foram muito próximos. O escore BLUE representa o número de n-gramas compartilhadas entre a tradução automática e a referência usada e varia de 0 a 1. O sistema baseado em regras Apertium apresentou resultados sutilmente melhores que o método estatístico (0,6 e

[173] Bick, E. (2006). Functional aspects in portuguese ner. In Proc. of the 7th Intl. Workshop, PROPOR, Lecture Notes in Artificial Inteligence. Springer-Verlag, Heidelberg.
[174] H. M. Caseli, M. G. V. Nunes, and M. L. Forcada. Evaluating the LIHLA lexical aligner on Spanish, Brazilian Portuguese and Basque parallel texts. Procesamiento del Lenguaje Natural, (35):237–244, 2005.
[175] H. M. Caseli, M. G. V. Nunes, and M. L. Forcada. LIHLA: A lexical aligner based on language-independent heuristics. In Proceedings of ENIA 2005, pages 641–650, Sao Leopoldo, RS, Brazil, 2005.
[176] Disponível em http://www.nilc.icmc.usp.br/nilc/tools/visuallihla/lihla.htm.
[177] Papineni, K.; Roukos, S.; Ward, T. and Zhu, W. BLEU: a Method for Automatic Evaluation of Machine Translation. Proceedings of he 40th Annual Meeting of the Association for Computational Linguistics (2002) 311-318.
[178] Corbí-Bellot, A.M.; Forcada, M.L.; Ortiz-Rojas, S.; Pérez-Ortiz, J.A.; Ramírez-Sánchez, G.; Sánchez-Martínez, F.; Alegria, I.; Mayor, A.; Sarasola, K. An open-source shallow transfer machine translation engine for the romance languages of Spain. 10th Annual Conference of the European Association for Machine Translation (2005) 79-86.

0,58, respectivamente). Diante disso, os autores optaram por fazer uma avaliação qualitativa (manual) dos métodos. Mais especificamente, foram calculados os erros lexicais e sintáticos e uma medida mista *word error rates (WER*) que considera o número de inserções, deleções e trocas necessárias para transformar uma candidata a tradução a uma referência. Para essa segunda avaliação, uma amostra de 20 traduções (com 482 palavras) foi analisada no nível sintático e lexical. Os resultados apresentaram valores maiores de WER para a abordagem estatística, o que sugere que, se comparada ao Apertium, exige um esforço maior para transformar a saída do sistema na tradução correta (0,32 para o método estatístico e 0,26 para o baseado em regras).

Morais e Ambrosio (2008) utilizaram para avaliar o sistema as métricas precisão (*precision*), revocação (*recall*) e *fall-out*. Vários experimentos foram realizados com o intuito de identificar alguns parâmetros usados na configuração do modelo: limiar para um termo ser considerado relevante (*Term Weight Index*), limiar para que dois documentos sejam considerados similares (*Similarity Index*), método de cálculo da similaridade (*Jaccard* ou *Overlap*), dicionário usado nas técnicas de mineração de texto, dentre outros. Os melhores resultados foram obtidos utilizando-se *Term Weight Index* superior ou igual a 25%, *Similarity Index* maior ou igual a 41%, calculando-se a similaridade pelo coeficiente *Overlap* e utilizando no dicionário de referência as *stopword*, os nomes próprios e as negações. No segundo experimento, repetiu-se as simulações anteriores usando outro conjunto de documentos, o que mostrou que os resultados independe dos exemplos usados. Os autores concluem que foi possível observar que o uso de ontologias para categorização de documentos é eficiente, se a ontologia tiver o "mínimo de qualidade, ou seja, tenha representatividade dos conceitos, propriedades, relações, funções, restrições e instâncias" (p. 6). Além disso, os autores concluem que o uso de técnicas estatísticas são adequadas mas fortemente dependentes do dicionário usado. E complementam que a criação de um dicionário de nomes próprios, para o contexto da jurisprudência, pode ser um processo interminável, diante da dinamicidade da área.

Caminada, Quental e Garrao (2008) apresentam o diagrama de classes da ferramenta implementada *Linguistics Tools* e mostram os resultados obtidos em dois experimentos realizados: no primeiro cálculo, utilizou-se janela de tamanho 2 para identificar assim os bigramas, e posteriormente, janela de tamanho 3 para

identificar os trigramas. Segundo os autores, com isso é possível identificar quando um bigrama não é um multivocábulo e sim parte de uma expressão maior. Como resultados, os autores apresentam a classificação Teste-T para cada corpus.

Seno e Nunes (2008) destacam que medidas de qualidade, interna e externa, podem ser usadas para garantir a eficiência da clusterização. Medidas externas comparam os agrupamentos *(clusters)* gerados com os classificados manualmente, enquanto que as medidas internas não usam qualquer tipo de conhecimento externo, apenas a coesão (*cohesiveness*) dos agrupamentos gerados, ou seja, medir quanto similar são os elementos de cada um. Apesar disso, os autores argumentam que para medir a qualidade de uma solução e a eficiência do método de clusterização, as medidas externas são mais apropriadas. Assim, utilizou-se como medidas de avaliação a precisão (*precision*), a revocação (*recall*), a medida F (*F-measure*), a entropia (*entropy*) e a pureza (*purity*). Dentre os parâmetros ajustados no modelo, os autores destacam o tamanho do cluster e o limiar de similaridade. Como resultados, os autores apresentam inúmeras relações existentes entre as medidas de avaliação usadas, os parâmetros do modelo e as medidas de similaridade implementadas. Observa-se que os resultados são muito próximos: todos os métodos de similaridade apresentaram bons resultados (cerca de 86% de medida F) para determinadas configurações. Os autores concluem que, inicialmente, o SiSPI foi proposto para o português, mas como é independente de domínio, pode ser facilmente customizado para outras linguagens.

Aziz, Pardo e Paraboni (2009) analisaram quatro parâmetros de configuração dos métodos estatísticos de tradução automática: a heurística de alinhamento, o tamanho máximo da frase, o uso de pesos de importância lexical e *tuning*. Vários experimentos foram realizados e, de uma maneira geral, a diferença entre os resultados obtidos foi muito sutil, normalmente na terceira casa decimal (em torno de 0,3). Para avaliação, foram utilizadas as medidas BLEU[179] e NIST[180], que representam o número de n-gramas compartilhados entre a tradução da máquina e a humana (usada como referência).

Em **Seno e Nunes (2009),** a seleção de conteúdo foi avaliada comparando cada sentença gerada automaticamente com duas sentenças de

[179] Papineni, K.; S. Roukos; T. Ward and W. Zhu (2002) “BLEU: a Method for Automatic Evaluation of Machine Translation”. ACL-2002, pages 311-318.
[180] NIST (2002) “Automatic Evaluation of Machine Translation Quality using n-gram Cooccurrence Statistics”. http://www.nist.gov/speech/tests/mt/doc/ngram-study.pdf

referência produzidas por dois humanos. Para cada um dos 57 conjuntos (contendo de 2 a 4 sentenças cada), os humanos foram instruídos a produzir uma única sentença, preservando apenas as informações comuns entre elas. A concordância entre os humanos foi avaliada calculando a precisão, a cobertura e a medida F de cada sentença do primeiro humano em relação à sentença do segundo humano. As autoras concluem que o sistema obteve 91% de medida F no melhor caso, sendo próximo dos resultados reportados em tarefa similar para a língua inglesa (96%).

Em **Salles *et al*. (2009),** para quantificar a eficácia dos classificadores propostos, foram utilizadas as métricas precisão, revocação, macroF1 e acurária.Os resultados mostraram que a versão temporal dos algoritmos apresentaram um ganho na acurácia de 3% para o KNN e de 7% (documentos da computação) ou 11% (para os documentos da medicina) utilizando-se o algoritmo Rocchio.

Braga, Monard e Matsubara (2009) destacam que os resultados usando o *Self-Training* não apresentaram melhoria consistente ao utilizar unigramas e bigramas para representar os documentos. Além disso, os piores resultados foram obtidos utilizando-se somente bigramas, e que os resultados utilizando-se somente unigrama eram compartíveis aos observados quando utilizou-se a combinação dos dois. Segundo os autores, não houve diferença significativa entre os algoritmos avaliados: *Self-Training* e *Co-Training*.

Villavicencio, Caseli e Machado (2009) afirmam que para avaliar a eficácia das abordagens analisadas para identificação de expressões multi-palavras em um corpus de domínio específico, foi realizada uma comparação automática usando o padrão ouro construído (glossário de ngramas de pediatria). Foram utilizadas as medidas precisão (número de candidatas corretas dentre as opções retornadas), *recall* (número de candidatas corretas dentre todas as opções da lista de referência) e a medida F (combinação das duas anteriores). Os autores destacam que os melhores resultados foram obtidos utilizando-se a abordagem estatística (os ngramas que não contém pontuação ou números, com frequência superior a 5 que não comecem com determinantes, verbos auxiliares, pronomes, advérbios, conjunções, dentre outros) considerando-se somente os n melhores candidatos de acordo com as medidas *pointwise mutual information* (PMI) e informação mútua (MI): cerca de 56% de precisão.

4.2.2. Análise de Conteúdo das publicações: sistematização dos enunciados apresentados

Nesta seção são apresentados os resultados obtidos a partir da análise de conteúdo das 68 publicações sorteadas (de acordo com os critérios definidos e apresentados no capítulo anterior). Pretende-se nesta seção apresentar as publicações analisadas dentro de um sistema semântico definido a partir das categorias de análise utilizadas. Essa visão sistêmica foi elaborada a partir da análise de conteúdo desse material empírico.

Conforme mencionado anteriormente, algumas categorias de análise foram definidas tendo em vista que a presente tese tem como objetivo analisar a produção científica da área de PLN. Sendo assim, durante a análise de conteúdo das publicações sorteadas, procurou-se extrair principalmente a problemática discutida pelos autores, a metodologia adotada e os resultados alcançados. Para as categorias de análise 'problemática' e 'metodologia adotada', um mapa conceitual foi construído com o intuito de sintetizar as temáticas relevadas pelos autores das publicações analisadas. Todos os mapas conceituais apresentados nesta tese foram construídos usando a ferramenta CMap[181], e os critérios de formatação e *layout* adotados serão explicados juntamente com os diagramas. Vale adiantar que, durante a elaboração dos mapas procurou distribuir as publicações em ordem cronológica no sentido anti-horário (o que permitiu uma melhor distribuição e visualização dos blocos diagramáticos).

Analisando-se os trabalhos quanto a realização de experimentos é possível observar que ao longo dos anos a porcentagem de trabalhos que apresentaram experimentos práticos aumentou: das 4 publicações analisadas da década de 80, nenhuma apresentou experimentos, das 14 analisadas na década de 90, 7 apresentaram experimentos, enquanto que a partir dos anos 2.000, 74% das publicações analisadas apresentaram experimentos práticos. Isso sugere que as pesquisas na área de PLN, ao longo dos anos, têm apresentado um enfoque mais experimental, talvez como consequência da forte inserção da área da ciência da computação (TAB. 12).

[181] *CMapTools Knowledge kit* - versão 5.04, disponível em http://cmp.ihmc.us

TABELA 12
Publicações envolvendo experimentos práticos por década

Década	Total de publicações	Publicações envolvendo experimento	Percentual de publicações envolvendo experimento
1986-1989	4	0	0%
1990-1999	14	7	50%
2000-2009	43	32	74%

Neste sentido, procurou-se avaliar o desenvolvimento de experimentos práticos por área dos autores das publicações. Observou-se que 86% dos trabalhos que apresentaram experimentos, tinham pelo menos um autor da área da computação. Além disso, dos artigos que possuíam pelo menos um autor da linguística, a maioria (70%) não apresentou experimentos práticos, com exceção de três trabalhos: Leffa (1991) (dicionário), Rosa (1998), Oliveira *et al*. (2003). Além disso, observou-se que depois do ano de 2.000, todos os artigos que possuíam pelo menos um autor da área da ciência da computação, apresentaram experimentos práticos, com exceção de Caseli (2008).

Para cada artigo analisado, procurou-se avaliar também o tipo de avaliação adotada pelos autores. Analisando-se os artigos, foram encontrados basicamente dois tipos de avaliação: 35% adotaram avaliação automática, com medidas estatísticas de erros e acertos, 45% usaram validação manual, ou seja, envolveu a avaliação de um humano, e 20% dos trabalhos envolveu avaliação automática e manual. Um aspecto interessante foi observado quando analisou-se também se foi utilizado no trabalho algum corpus de documentos para teste e consequente avaliação. Correlacionando o tipo de avaliação com o corpus utilizado, observou-se que dos trabalhos que tiveram avaliação automática, 40% fez uso de corpora reutilizados, oque facilita que métricas estatísticas sejam adotadas.

Outro aspecto interessante observado foi o idioma para o qual o artigo foi desenvolvido: 65% dos artigos analisados foram desenvolvidos para o português, 20% para o inglês, enquanto que 15% dos trabalhos eram propostas de abordagens genéricas, ou seja, independente da linguagem natural foco.

A análise horizontal realizada com base nos títulos de todas as 621 publicações relevantes mostrou que 34% delas tinham o primeiro autor de área desconhecida, e que 66% havia sido desenvolvido de maneira multidisciplinar, ou

seja, envolvendo pesquisadores de várias áreas. Na análise profunda, realizada adentrando-se no conteúdo das publicações, foi possível ter acesso, a partir dos cabeçalhos dos artigos, à área dos autores que foram consideradas desconhecidas, de acordo com a coleta automática da Plataforma Lattes.

Apesar de a análise horizontal revelar que 66% das publicações foram desenvolvidas envolvendo pesquisadores de várias áreas, o mesmo não foi observado no recorte submetido à análise de conteúdo, onde a grande maioria (78%) foi escrito por autores da mesma área. Dentre os artigos sorteados para análise de conteúdo, nenhum deles foi escrito por pesquisadores da área da ciência da informação, 12% possuíam somente autores da linguística, 76% somente por autores da ciência da computação, 6% envolvendo pesquisadores das duas áreas (computação e linguística), e 7 % de outras áreas.

A primeira dimensão analisada foi a problemática abordada pelos autores dos artigos submetidos à análise de conteúdo. Na FIG. 15 é apresentado um mapa conceitual contendo todas as problemáticas abordadas nas 68 publicações analisadas. Pode-se observar que, a temática central é o PLN e foi colocado no centro do mapa em vermelho. A partir dele, identificou-se alguns problemas recorrentes, tais como sumarização, tradução, recuperação de informação, tratamento de ambiguidade e outros, que foram destacados em roxo. Em torno desses problemas, organizaram-se as problemáticas observadas nas publicações, que foram incluídas no mapa em negrito e vermelho. Os autores foram apresentados de maneira mais discreta (retângulos brancos) visto que o objetivo era destacar as problemáticas e não as personalidades.

Na TAB. 13 são apresentadas as problemáticas reveladas a partir da análise de conteúdo das publicações avaliadas, juntamente com o número de artigos relacionados. É possível observar que dos 68 artigos analisados, 18 foram sobre recuperação de informação, enquanto que as problemáticas de sumarização, tratamento de ambiguidade, analisadores (*parser*) e tradução foram igualmente enfatizadas (apesar dessa última ter tido um artigo a menos).

TABELA 13
Principais problemáticas reveladas a partir da análise de conteúdo

Problemática	**#Artigos relacionados**
Recuperação de informação	18
Sumarização	10
Tratamento de Ambiguidade	10
Analisadores (*parser*)	10
Tradução	9
Aplicações para a própria área	4
Exemplos de aplicações do PLN	4
Correção automática	3
Total	**68**

Com o intuito de facilitar a compreensão do mapa conceitual apresentado na FIG. 14, optou-se por apresentar também recortes de acordo com as principais problemáticas (FIG. 20 a 24). As discussões que segue são apresentadas de acordo com a evidência apontada pela quantidade de artigos relacionados com cada problemática (TAB. 13).

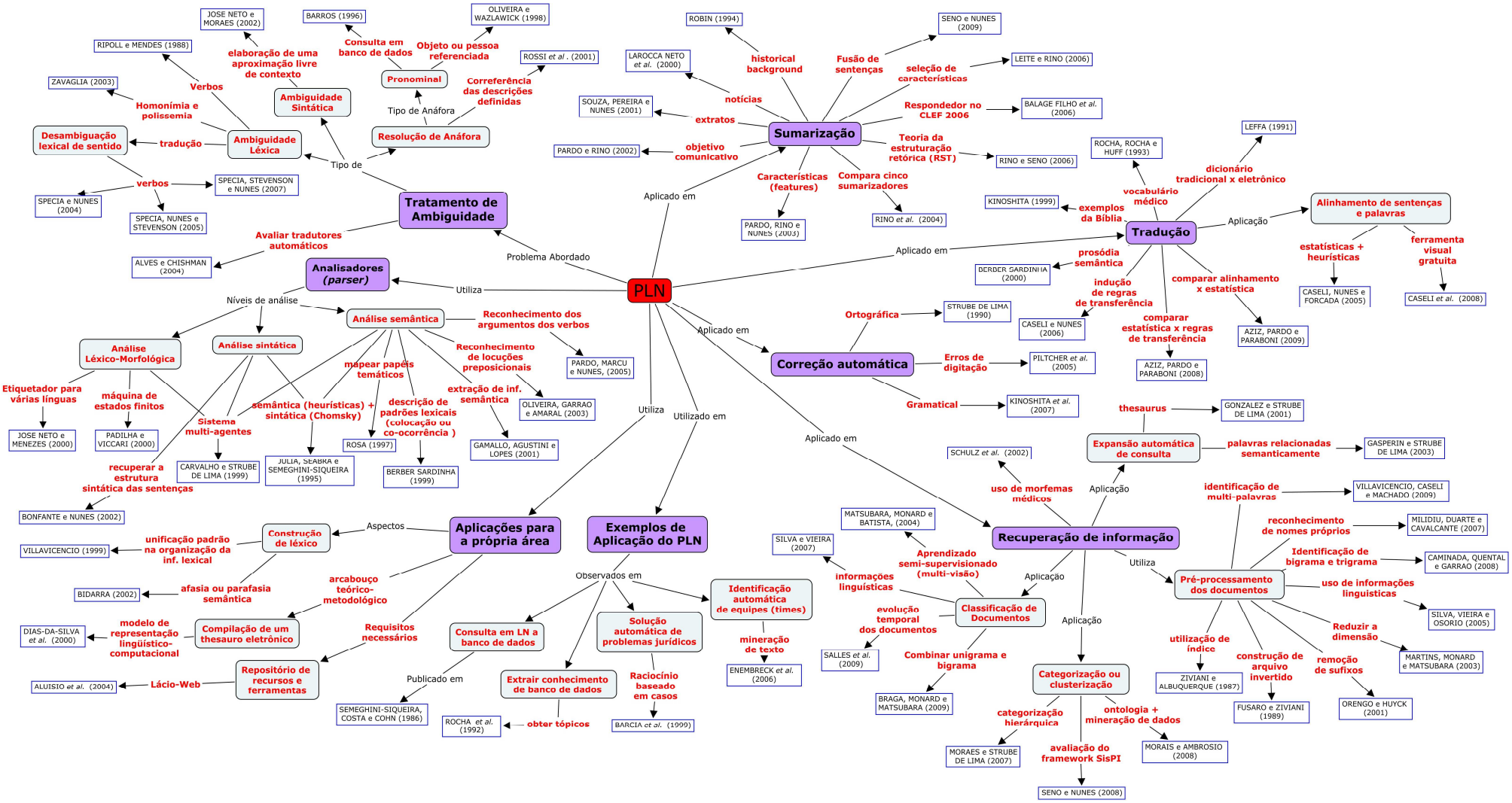

FIGURA 19 – Mapa conceitual contendo as problemáticas observadas nas publicações analisadas

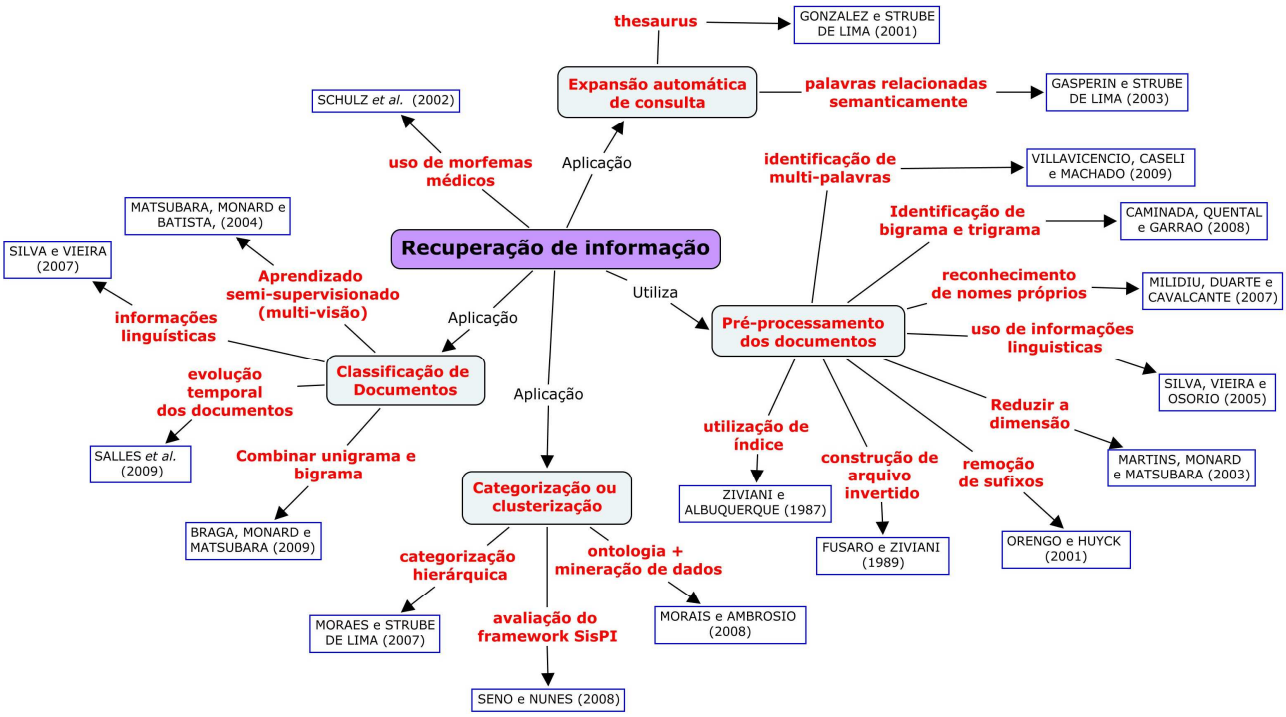

FIGURA 20 – Mapa conceitual apresentando as problemáticas observadas nas publicações analisadas: recorte RECUPERAÇÃO DE INFORMAÇÃO

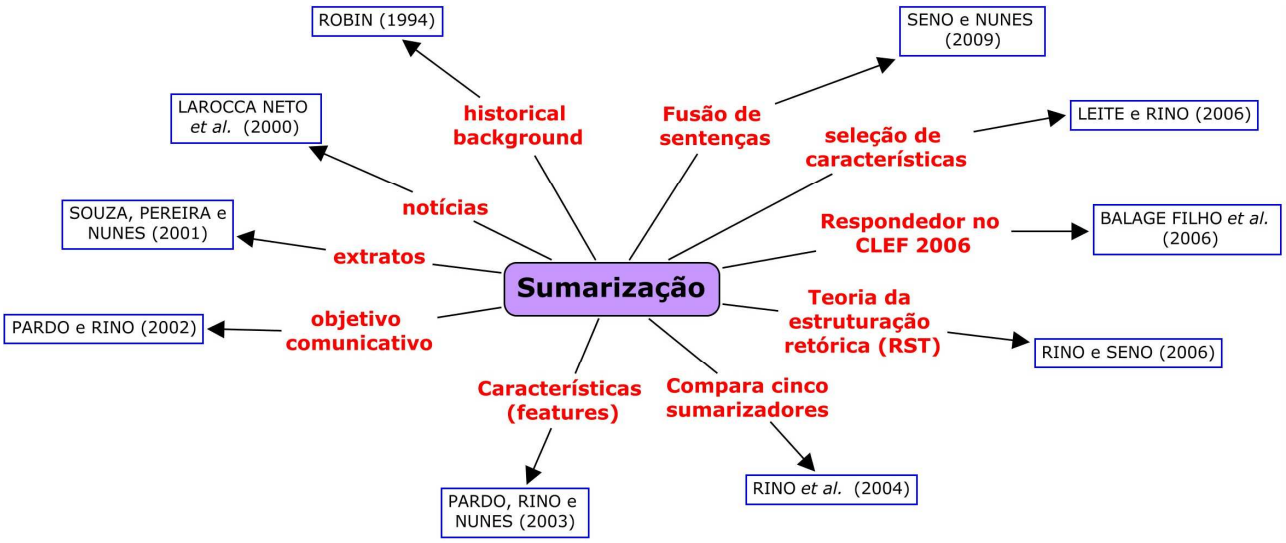

FIGURA 21 – Mapa conceitual apresentando as problemáticas observadas nas publicações analisadas: recorte SUMARIZAÇÃO

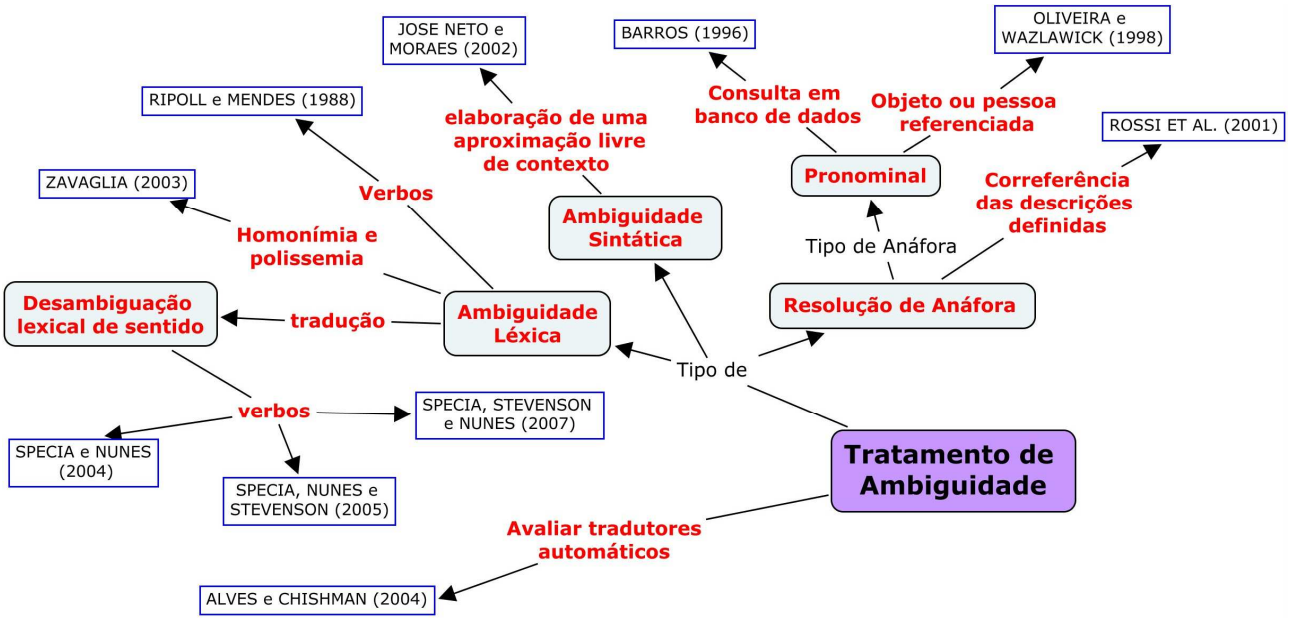

FIGURA 22 – Mapa conceitual apresentando as problemáticas observadas nas publicações analisadas: recorte TRATAMENTO DE AMBIGUIDADE

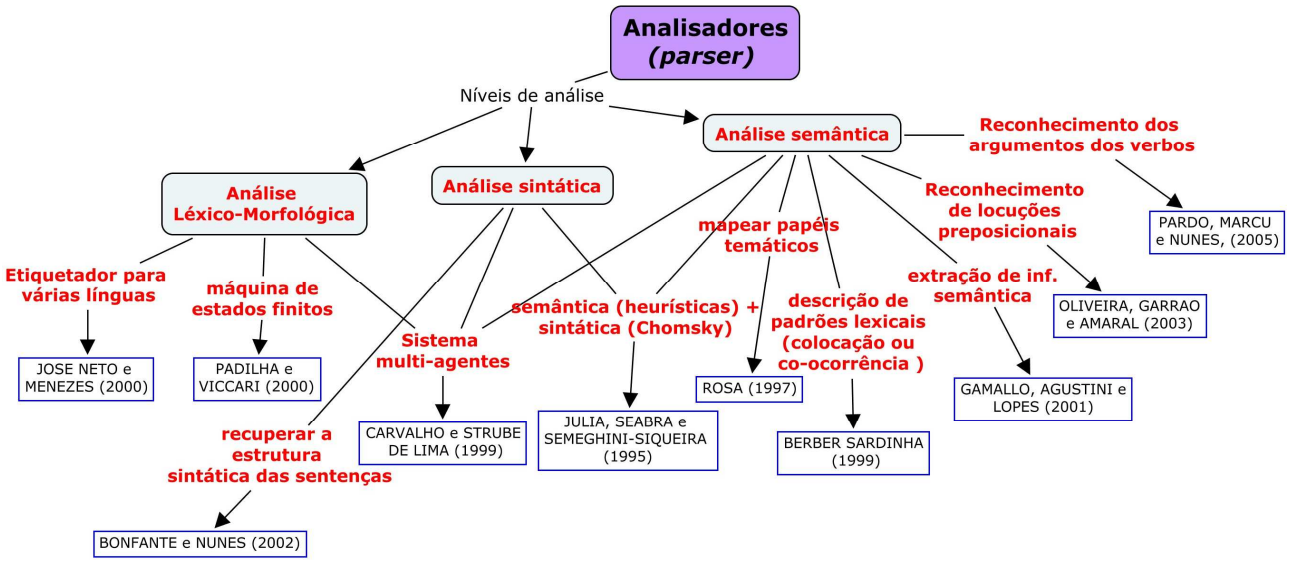

FIGURA 23 – Mapa conceitual apresentando as problemáticas observadas nas publicações analisadas: recorte ANALISADORES (*PARSER)*

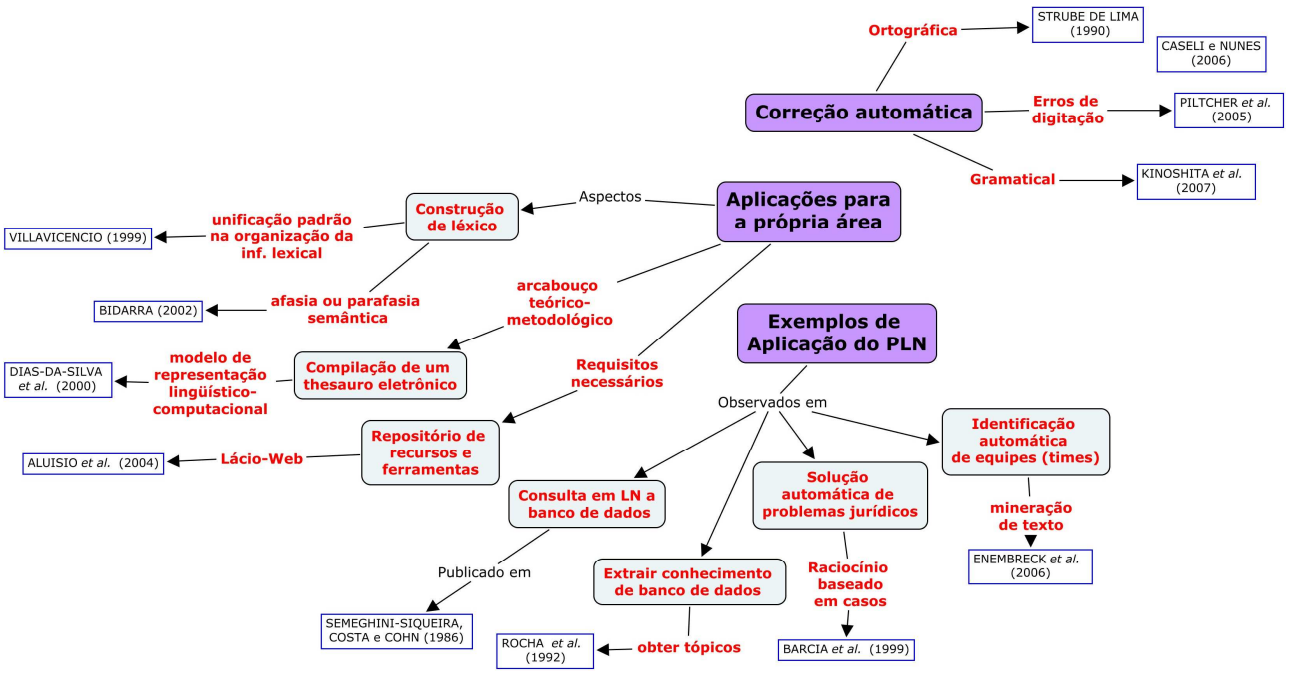

FIGURA 24 - Mapa conceitual apresentando as problemáticas observadas nas publicações analisadas: recorte OUTRAS

A próxima categoria de análise focalizada foi a metodologia adotada pelos autores dos artigos analisados. Procurou-se apresentar, para cada problemática observada, o substrato metodológico-conceitual que emergiu dos artigos analisados. No centro do mapa conceitual, manteve-se a problemática identificada (em vermelho), e a partir de cada artigo, a metodologia foi detalhada. Os recursos utilizados nos artigos analisados foram apresentados em amarelo.

A discussão que segue cada mapa conceitual construído traduz o rigor acadêmico adotado, tanto no tocante à análise, quanto nas correlações delineadas. As conclusões apresentadas foram direcionadas por algumas considerações e por alguns pressupostos. Por exemplo, no escopo desta tese, considerou-se classificação quando os documentos foram previamente atribuídos a alguma classe, e categorização como sendo sinônimo de clusterização (do inglês *clustering*), que ocorre quando os documentos são agrupados em função das suas similaridades, e não de conhecimento prévio das categorias. Outras considerações serão discutidas a medida que elas se tornarem necessárias.

4.2.2.1. Problemática RECUPERAÇÃO DE INFORMAÇÃO

Na FIG. 25 é apresentado o mapa conceitual construído a partir da metodologia adotada pelos artigos analisados sobre RECUPERAÇÃO DE INFORMAÇÃO.

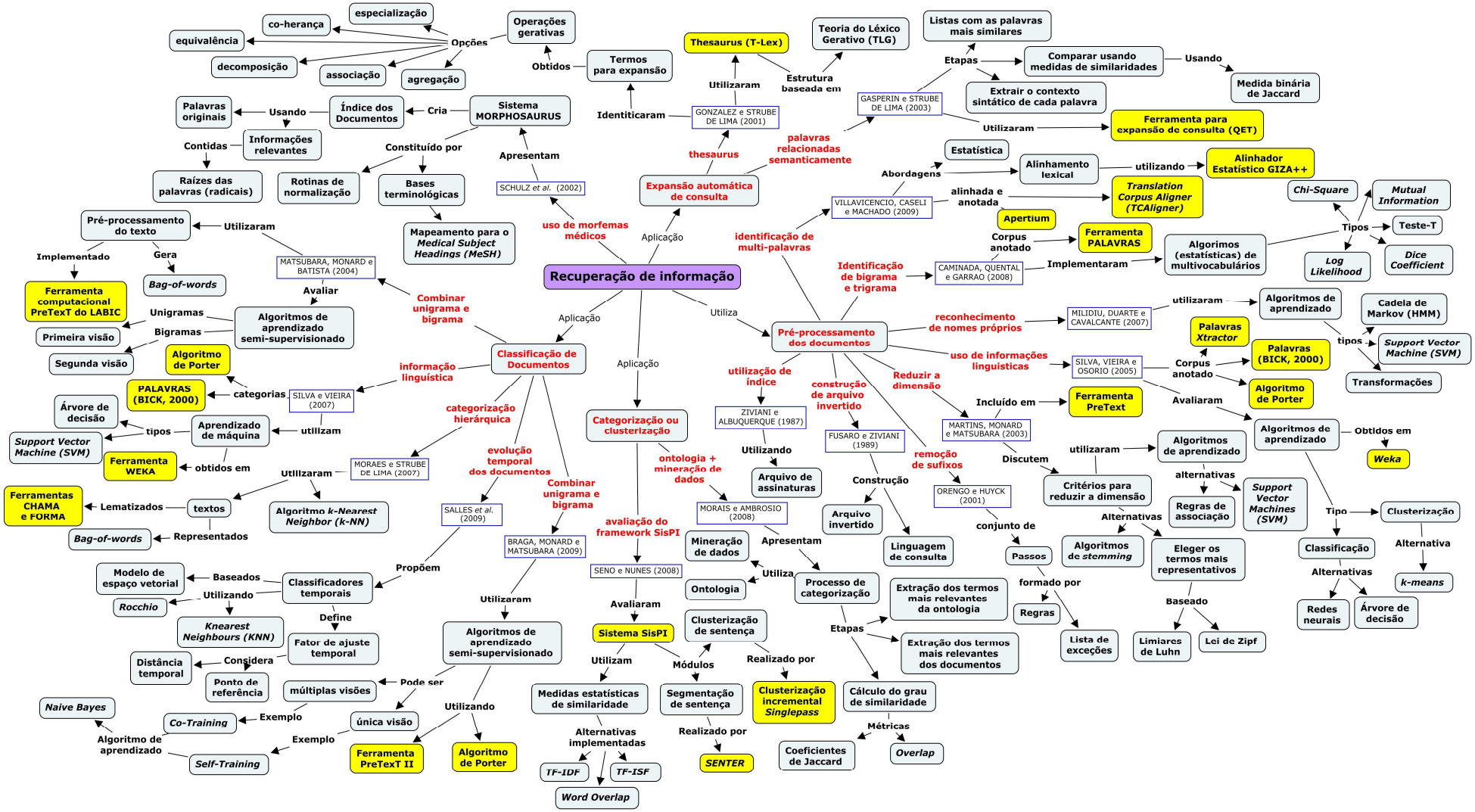

FIGURA 25 – Mapa conceitual apresentando as metodologias observadas nas publicações analisadas que tiveram como problemática RECUPERAÇÃO DE INFORMAÇÃO

Dentre as problemáticas reveladas pelos artigos analisados, a recuperação de informação foi a que sem dúvida teve maior destaque na produção científica nacional. Além disso, a grande maioria desses artigos são trabalhos recentes (nos anos 2.000), o que reflete a efervescência que a área vem passando. Além disso, dos 18 artigos analisados sobre recuperação de informação, oito são voltados para técnicas de pré-processamento de documentos, o que sugere que este tema ainda esteja em aberto.

Para facilitar a compreensão das observações que emergiram desta análise, optou-se por apresentar a problemática da recuperação de informação, de maneira segmentada, seguindo o mapa conceitual da FIG. 25, no sentido anti-horário.

Dos trabalhos analisados, dois tinham como problemática central a utilização de PLN na expansão automática de consultas em recuperação de informação, sendo que ambos apresentaram um autor em comum (Strube de Lima). O primeiro trabalho, Gonzalez e Strube de Lima (2001), utilizou um thesaurus, que possui uma estruturação semântica para implementar relacionamentos lexicais, considerando fundamentos da Teoria do Léxico Gerativo (TLG) de Pustejovsky. Segundo os autores, os resultados obtidos na avaliação indicam que a expansão de consulta pode trazer benefícios à RI, mas advertem que esta expansão não pode ser feita indiscriminadamente, sob pena de prejudicar os resultados do mecanismo de busca. O outro trabalho, Gasperin e Strube de Lima (2003), teve como objetivo avaliar a qualidade de uma lista de palavras semanticamente relacionadas, gerada automaticamente durante o mestrado da primeira autora. Foram realizadas sete consultas (nas versões original e expandida), e observou-se que, quando comparado à consulta original, a consulta expandida aumentou a revocação mas diminuiu a precisão.

De acordo com as definições adotadas no escopo desta tese, cinco artigos abordaram a classificação de documentos. Dentre os trabalhos analisados, observou-se que dois trabalhos apresentaram co-autorias coincidentes: Matsubara, Monard e Batista (2004) e Braga, Monard e Matsubara (2009). Ambos destacam a utilização do aprendizado de máquina semi-supervisionado. Segundo os autores, este modelo combina as facilidades do aprendizado supervisionado – no qual é fornecido um conjunto de exemplos de treinamento rotulado com a classe de cada exemplo, com as do aprendizado não-supervisionado – no qual a classe de cada exemplo não é conhecida. Segundo os autores, excelentes resultados foram alcançados, o que comprova a sua aplicabilidade nos casos em que se possui apenas um pequeno conjunto de exemplos rotulados, e o custo para rotular mais exemplos é alto. Além disso, Braga, Monard e Matsubara (2009) ressaltam que combinar unigramas e bigramas não tem apresentado melhorias

significativas na classificação supervisionada de textos. O trabalho mostrou que os piores resultados foram obtidos utilizando-se somente bigramas, e que os resultados utilizando-se somente unigrama foram compartíveis aos observados quando se utilizou a combinação dos dois. Apesar de esses trabalhos terem sido desenvolvidos em co-autoria de pesquisadores brasileiros, ambos utilizaram bases de documentos em inglês, o que sugere que experimentos análogos devessem ser realizados para o português.

O artigo de Silva e Vieira (2007) define categorização como sendo o processo de alocar os documentos em categorias pré-definidas, e por este motivo foi atribuído a subproblemática de classificação, diante da definição de classificação adotada no escopo desta tese.

Vale destacar que, este trabalho se assemelha com outro trabalho, Silva, Vieira e Osorio (2005), publicado em co-autoria pela mesma dupla de autoras, que foi alocado, no escopo desta tese, na subproblemática de pré-processamento. Isso se deve ao fato do primeiro (de 2005) ter dado ênfase à utilização de informações linguísticas (categorias gramaticais) como características, e por isso estar na categoria de pré-processamento, enquanto que o mais recente (de 2007) enfatizou a comparação entre os algoritmos de aprendizado de máquina avaliados, e, portanto foi atribuído à subproblemática dos trabalhos de classificação.

Os melhores resultados observados no artigo de Silva e Vieira (2007) foram obtidos quando utilizou-se nomes juntamente com os adjetivos como características descritivas dos documentos, com taxa de erro de 18%. No entanto, os autores destacam que o segundo melhor resultado (com erro de 19,7%) foi observado quando utilizou-se o método tradicional (*stopwords* e *stemming*), o que sugere que ambas abordagens apresentam desempenho similar.

Os autores do próximo artigo, Moraes e Strube de Lima (2007), usaram os termos classificação e categorização como sendo sinônimos, assim como classes e categorias. Segundo as autoras, o artigo considera a categorização sobre uma coleção de documentos não rotulados, que se encontram apenas organizados sob títulos de 29 seções da Folha de São Paulo. No entanto, essa organização foi usada como sendo a classe para efeito de avaliação do modelo. Segundo as autoras, os resultados obtidos foram muito ruins, e atribuem isso à incerteza da categorização por tópicos apresentadas na base de documentos usada. Segundo elas, como os documentos utilizados não foram previamente rotulados, não há como distinguir certos documentos sem um processo manual.

Ainda dentro da subproblemática de classificação de documentos, um trabalho

que merece destaque é o Salles *et al.* (2009) que propõe uma nova abordagem para o tratamento dos efeitos temporais em algoritmos de classificação já conhecidos, derivando assim classificadores robustos temporalmente. Os autores destacam que apesar do tempo ser uma dimensão importante para qualquer espaço informacional, a maioria das técnicas atuais de classificação automática de documentos não considera a evolução temporal dos documentos. Em outras palavras, ignora-se o fato de que a variação na definição dos termos e das classes ao longo do tempo tende a tornar o conjunto de treinamento muito confuso, impactando negativamente nos classificadores que negligenciam esta evolução. Essa preocupação não foi observada em nenhum artigo analisado dentre a produção científica nacional, nem mesmo no ARIST.

Dentro da subproblemática de categorização, dois artigos foram analisados: Seno e Nunes (2008) destacam que identificar sentenças ou trechos similares de textos tem desempenhado um importante papel em várias aplicações de PLN, tais como geração de parágrafo, sumarização automática, construção de ontologias, bibliotecas digitais, dentre outras. Sendo assim, o presente trabalho teve como objetivo comparar um método não supervisionado e incremental de clusterização, com métodos usando somente métricas estatísticas de similaridades. Para critério de avaliação, criou-se um corpus de referência, onde cada sentença, de cada documento, foi manualmente classificada, ou seja, associada a um cluster. Os resultados obtidos foram muito próximos: todos os métodos de similaridade apresentaram bons resultados (cerca de 86% de medida F).

O outro artigo relacionado à categorização de documentos foi o de Morais e Ambrosio (2008) que avaliou o uso de ontologia de domínio na tarefa de clusterização, usando documentos de jurisprudência. Este trabalho poderia ter sido alocado, no escopo desta tese, em exemplos de aplicações, no entanto, a análise de conteúdo revelou a ênfase dada pelos autores no método de categorização, enquanto que o domínio jurisprudência foi usado exclusivamente com o propósito de ilustrar o experimento. Os autores destacaram que, para identificar o contexto semântico dos documentos, tem-se usado técnicas de mineração de texto *(text-mining)* ou ontologias. No entanto, os autores destacam que não foi encontrada na literatura alguma pesquisa que combine ambas as estratégias para desenvolver mecanismos de categorização automática de documentos. Assim, este trabalho teve como objetivo analisar automaticamente se um documento é relevante, dado o domínio representado por uma ontologia. Mesmo sendo um trabalho de categorização, os autores utilizaram dois conjuntos de classificados para avaliar o modelo construído. Os autores concluíram que o uso de ontologias para categorização de documentos é eficiente, se a ontologia tiver o “mínimo de qualidade, ou seja, tenha

representatividade dos conceitos, propriedades, relações, funções, restrições e instâncias" (p. 6). Além disso, os autores concluem que o uso de técnicas estatísticas são adequadas, mas fortemente dependentes do dicionário usado (lista de *stopwords,* nomes próprios, adjetivos, dentre outros).

Os dois primeiros artigos analisados envolvendo pré-processamento são voltados para a construção de índices e arquivos invertidos para representação dos documentos no processo de recuperação de informação. O artigo seguinte Martins, Monard e Matsubara (2003) apresentou a ferramenta PreText, desenvolvida com o objetivo de realizar automaticamente a tarefa de pré-processamento de uma coleção de documentos. Segundo os autores, na fase de pré-processamento, os documentos podem ser transformados em um vetor de termos (*bag-of-words*) que ocorrem no documento. Os termos que compõem este vetor podem ser palavras simples ou compostas (2, 3, …, n-gram). Neste trabalho, os autores apresentam como alternativas para redução da dimensão desse vetor, utilizar o radical dos termos, utilizando-se um algoritmo de remoção de sufixos *(stemming),* ou eleger os termos mais significativos, usando a lei de Zipf e o limiar de Luhn. Os autores destacam ainda que representações mais elaboradas têm sido avaliadas, mas apresentado resultados piores.

Apesar disso, o próximo trabalho, Silva, Vieira e Osorio (2005), propõem uma nova técnica de pré-processamento utilizando informações linguísticas, selecionando combinações de categorias (nomes, adjetivos, nomes próprios e verbos) nas tarefas de classificação e clusterização de documentos. Os melhores resultados foram obtidos ao utilizar os nomes, sendo que a melhor taxa de acerto foi obtida quando utilizou-se nomes juntamente com os adjetivos. O método tradicional (*stopwords* e *stemming*) apresentou o segundo melhor resultado, confirmando o que discutido pelos autores do artigo anterior.

O próximo trabalho, Milidiu, Duarte e Cavalcante (2007), envolve o reconhecimento automático de nomes próprios. Os autores criaram manualmente um sistema de referência composto com nomes de localidades, personalidades e organizações extraídas da Web. Utilizou-se uma janela de tamanho igual a 5, incluíndo a palavra corrente, as duas anteriores e as duas posteriores. Utilizando-se o sistema de referência como treinamento, o modelo apresentou 88% de medida F. No entanto, sem o sistema de referência, os resultados obtidos são, segundo os autores, muito pobres.

O trabalho seguinte, Caminada, Quental e Garrao (2008), tem como objetivo apresentar uma abordagem estatística para a busca e identificação de bigramas e trigramas multivocabulares da língua portuguesa, baseados em padrões gramaticais definidos por um processo de anotação. O próximo trabalho, Villavicencio, Caseli e

Machado (2009), apresentam como abordagens para identificação de expressões multi-palavras a utilização de estatísticas e de alinhamento lexical. Segundo os autores, os melhores resultados foram obtidos utilizando-se a abordagem estatística.

As bases de documentos, assim como as ferramentas utilizadas pelos trabalhos analisados e discutidos nesta problemática, serão apresentados posteriormente em momento oportuno.

4.2.2.2. Problemática SUMARIZAÇÃO

Na FIG. 26 é apresentado o mapa conceitual construído a partir da metodologia adotada pelos artigos analisados sobre SUMARIZAÇÃO.

A análise de conteúdo apontou para a existência de duas abordagens principais de sumarização automática de textos: abordagem empírica e fundamental. A abordagem empírica é também chamada de superficial ou pobre de conhecimento, uma vez que é baseado em informações estatísticas ou empíricas. Nesta abordagem, as sentenças normalmente são representadas por tabelas de atributo-valor que representam as características *(features)* extraídas do texto. A partir da abordagem empírica ou superficial, é possível gerar extratos, ou seja, selecionar as sentenças relevantes do texto original e por meio de justaposição, obter uma síntese. Já a abordagem fundamental é também chamada de profunda ou rica em conhecimento, visto que é baseada em informações linguísticas. A abordagem profunda tende a produzir sumários textuais, ou resumos da designação em português, reformulando o conteúdo do texto original e gerando novas sentenças. Dos dez trabalhos analisados dentro da problemática sumarização, somente dois (PARDO; RINO, 2002; RINO; SENO, 2006) usaram a abordagem profunda e produziram sumários (resumos). Assim, pode-se afirmar que os trabalhos analisados sugerem que a maioria das pesquisas em sumarização automática tem privilegiado a abordagem empírica utilizando diferentes características extraídas dos textos.

O trabalho de Pardo e Rino (2002) aplicou a abordagem fundamental para avaliar se o objetivo comunicativo foi mantido no sumário gerado. Segundo os autores, 31% dos sumários automáticos gerados preservaram totalmente a ideia central do texto original, enquanto que 61% deles mantiveram parcialmente. Além disso, os autores destacam que todos os resumos elaborados pelos próprios autores (chamado de sumários autênticos) preservaram totalmente a ideia do texto original. O trabalho de Rino e Seno (2006) voltou a utilizar a abordagem fundamental ou profunda no sistema

implementado RHeSumRST, para avaliar a importância do tratamento co-referencial na sumarização automática. O sistema desenvolvido foi comparado com dois outros sistemas, e segundo as autoras, o sistema foi mais informativo que um deles, e menos informativo que o outro. Além disso, as autoras afirmaram que o sistema apresentou menos quebra (índice de 3%) de co-referência nos sumários, quando comparado com os outros sistemas (que apresentaram 5% e 15% de quebra). As próprias autoras concluem que o RheSumaRST apresentou o menor índice de quebra, quando comparado aos outros sistemas, mas que a diferença não justifica o esforço necessário de modelagem e processamento estrutural.

Os autores que utilizaram a abordagem empírica destacaram que verificar a proximidade dos extratos gerados automaticamente com seus sumários autênticos (elaborados pelo próprio autor) seria uma tarefa difícil, visto que os autores tendem a não preservar as sentenças do texto original. Assim, os autores têm adotado uma versão extrativa dos sumários autênticos, chamada de sumários ideais.

Dos artigos analisados, somente o terceiro, Souza, Pereira e Nunes (2001), utilizou o português como idioma fonte. Os autores usaram a abordagem empírica e afirmaram que apesar de ser um método simples, não havia notícia de trabalhos anteriores voltados para o português.

O trabalho seguinte, Pardo, Rino e Nunes (2003), utilizou a abordagem empírica (estatística) utilizando as redes neurais artificiais *SelfOrganizing (SOM)* para a geração de extratos. Segundo os autores, este foi o primeiro trabalho de sumarização automática para o português utilizando RNAs. A rede neural utilizada apresentou resultado melhor (41% de erro) quando comparada com outros algoritmos de classificação: *Naive-Bayes* (57% de erro) e árvore de decisão C4.5 (51% de erro). Além disso, os autores destacam que estes resultados foram obtidos a partir de sentenças classificadas manualmente por juízes humanos. Assim, ao assumir que é possível flexibilizar essa classificação manual, as redes neurais apresentaram apenas 27% de erro. Os próprios autores assumem que os resultados alcançados: 32% de cobertura (*recall*) e 41% de precisão são aparentemente insatisfatórios. No entanto, isso não significa que os extratos produzidos sejam ruins. Os autores afirmam que os valores obtidos são próximos aos apresentados por outros trabalhos semelhantes.

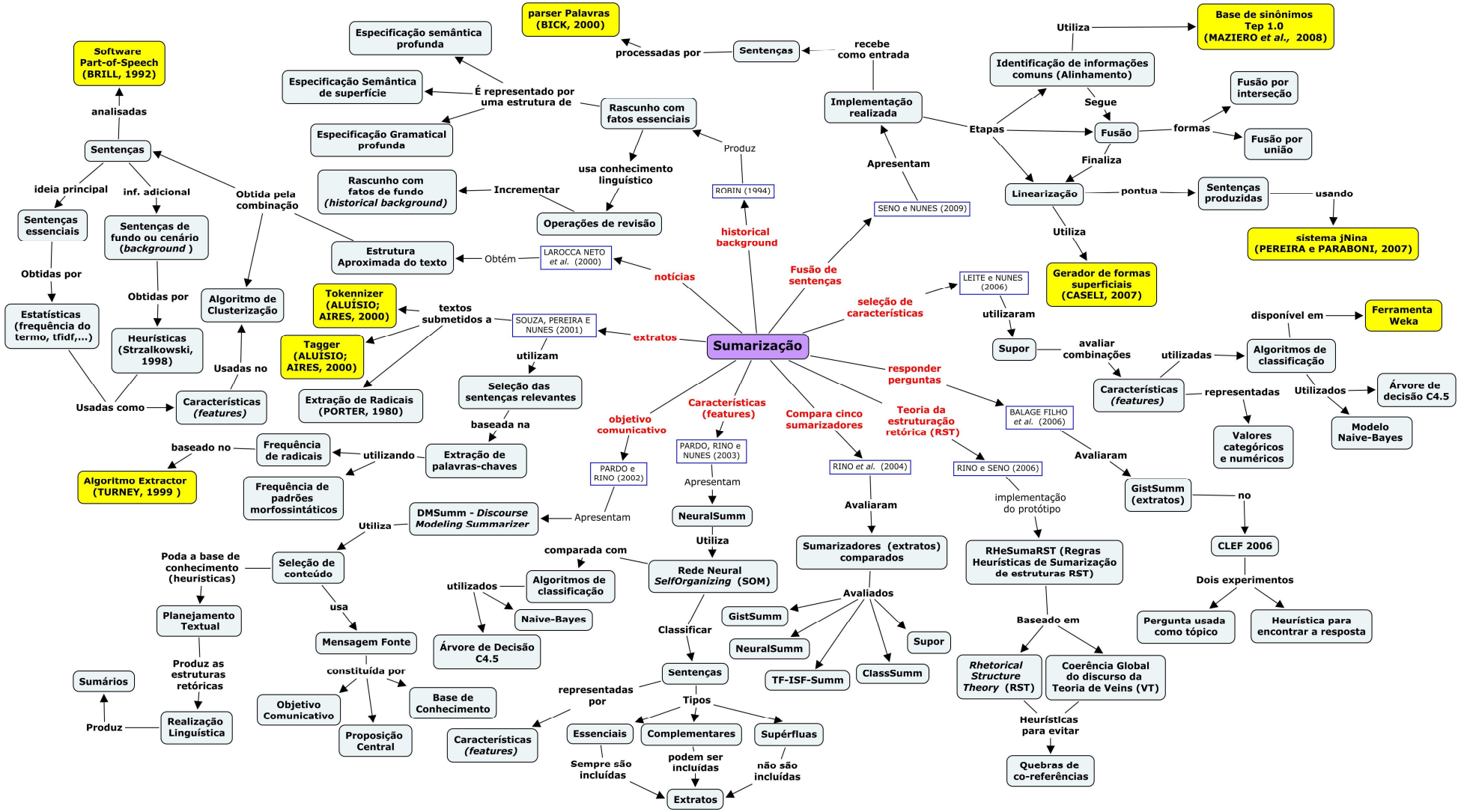

FIGURA 26 - Mapa conceitual apresentando as metodologias observadas nas publicações analisadas que tiveram como problemática SUMARIZAÇÃO

O trabalho de Rino *et al.* (2004) comparou cinco sistemas sumarizadores que utilizam a abordagem empírica (estatística), comparando os sumários gerados e os sumários ideais (construídos a partir dos sumários autênticos). Os sistemas Supor e ClassSum apresentaram os melhores resultados (42,8% e 42,4% de medida F, respectivamente), enquanto que o NeuralSum, apresentou os piores resultados (31% de medida f). Segundo os autores, a similaridade entre os sistemas Supor e ClassSum deve-se ao fato deles utilizarem características semelhantes.

O trabalho de Balage *et al.* (2006) aplicou o sumarizador GistSumm em um sistema respondedor automático, mas obteve resultados considerados pelos próprios autores como sendo “muito pobres”, pois a grande maioria das respostas foi sinalizada como sendo incorretas. E concluíram que “técnicas de sumarização simples não são suficientes para a tarefa de responder perguntas, apesar de serem eficientes para o que elas se propõem” (p. 375).

O trabalho de Leite e Rino (2006) utiliza características, mas propõe que sejam usados valores categóricos e numéricos ao invés de valores binários. Esta sugestão gerou uma melhoria de apenas 3% na medida f, quando comparado com outros sistemas.

Vale destacar que, as características utilizadas foram evoluindo ao longo dos anos. Inicialmente usava-se frequência de palavras, posição da sentença no texto, frequência de palavras temáticas, ocorrência de palavras-chaves, do título, dentre outras. Já o trabalho de Leite e Rino (2006) utiliza como características ocorrência de cadeia lexical (*lexical chains method*), que verifica a existência de palavras relacionadas (por exemplo, sinônimos/antônimos ou hipônimos/hiperônimos); o método do mapa de relacionamento textual (*text relationship map method*), semelhante ao método anterior, mas considera parágrafos ao invés de sentenças, e constrói um grafo chamado de mapa de relacionamento do texto fonte que representa o seu grau de coesão; o método da importância dos tópicos (*importance of topics method*), que tem como objetivo identifica os principais tópicos do texto fonte, que deverão orientar a seleção de sentenças, dentre outras.

Quanto a realização de experimentos, vale destacar que praticamente todos os trabalhos analisados realizaram experimentos práticos envolvendo corpus de documentos. Apesar de as bases de documentos serem apresentadas posteriormente, vale destacar a utilização recorrente do corpus TeMário, criado no âmbito do NILC sob a coordenação da profa. Lucia H. Machado Rino.

4.2.2.3. Problemática TRATAMENTO DE AMBIGUIDADE

Na FIG. 27 é apresentado o mapa conceitual construído a partir da metodologia adotada pelos artigos analisados sobre TRATAMENTO DE AMBIGUIDADE.

Analisando-se os trabalhos que tiveram como problemática observada o tratamento de ambiguidade, foi possível observar que ela tem sido um tema de pesquisa desde os primórdios do período analisado, com o artigo de Ripoll e Mendes (1988) até os dias atuais, com o artigo de Specia, Stevenson e Nunes (2007).

Dentre os problemas abordados no tratamento de ambiguidade, observou-se uma concentração em basicamente dois problemas: desambiguação lexical, principalmente de verbos (com cinco trabalhos) e resolução de anáforas (com três trabalhos). Os outros dois trabalhos eram voltados para simplificação da linguagem natural para uma gramática livre de contexto, e avaliação de tradutores automáticos quanto ao tratamento de ambiguidades.

O primeiro trabalho analisado foi o de Ripoll e Mendes (1988), que propõe utilizar um modelo conexionista e uma gramática de casos para tratar a ambiguidade léxica de um subconjunto de verbos no português. Os autores utilizam algumas características para representar os verbos, tais como se o verbo é causal, se existe um agente, qual a natureza da mudança, dentre outras. O sistema proposto resolve ambiguidade léxica de verbos em três níveis: léxico, de significado e dos casos verbais (utilizando-se os casos de Fillmore). Segundo os autores, a análise da frase baseada em casos verbais resolve muitos casos de ambiguidade léxica, porém existem algumas situações nas quais a informação da estrutura sintática auxiliaria no tratamento da ambiguidade.

O próximo trabalho analisado, que abordou tratamento de ambiguidade lexical, foi o de Zavaglia (2003). A autora destaca que, o problema da homonímia gramatical é resolvido facilmente por sistemas computacionais, mas o mesmo não acontece com outros problemas da ambiguidades, tais como homonímia semântica e a polissemia. Segundo a autora, isso se deve ao fato da máquina não ser capaz de relacionar semanticamente itens lexicais em meio a construções sintáticas ou inseridos no contexto. Assim, este trabalho teve como objetivo propor o tratamento de itens lexicais homônimos da língua portuguesa do Brasil, por meio da construção de uma base de dados conceitual, ou seja, uma base de conhecimento lexical. Segundo a autora, tal base irá suprir as necessidades de um analisador sintático, assim, a homonímia poderá ser tratada, uma vez que será fornecido à máquina, subsídios linguísticos tais como relações semânticas

de itens lexicais em redes de significação. Essa base é constituída de vários componentes: informação ontológica (esta ontologia foi construída na tese de Doutorado da própria autora), informação Qualia (baseada na Teoria do Léxico Gerativo de Pustejovsky), informação morfossintática, informação definicional (extraída de um dicionário base), e informação pragmática (exemplos do uso do item homônimo extraídos de um corpus de 11 milhões de palavras do Laboratório de Estudos Lexico-gráficos da UNESP de Araraquara). Para este trabalho, a autora não apresentou experimentos, apenas dois exemplos da representação de um item homônimo.

Os próximos três trabalhos, que foram publicados em co-autorias das mesmas pesquisadoras, abordaram a desambiguação lexical de sentido de verbos. Specia e Nunes (2004) alertaram para o problema da ambiguidade lexical que ocorre quando são identificadas apenas variações de significado (de sentido) nas opções de tradução, ou seja, todas as opções são da mesma categoria gramatical (o que é chamada de ambiguidade lexical de sentido). O projeto propõe a construção de um modelo híbrido linguístico-computacional, ou seja, baseado em conhecimento linguístico (dicionários) e em algoritmos de aprendizado de máquina (corpus de exemplos). As autoras procuraram identificar os casos mais problemáticos de ambiguidade utilizando três sistemas de tradução automática inglês-português. As traduções foram, então, manualmente analisadas para verificar a ocorrência da ambiguidade, seus efeitos na tradução das sentenças e o comportamento dos sistemas diante desse fenômeno. Esse trabalho (apresentado em 2004) encontrava-se em especificação e voltou a ser discutido em dois outros trabalhos em co-autoria em 2005 e 2007. O próximo trabalho analisado foi Specia, Nunes e Stevenson (2005) que teve como objetivo extrair regras do modelo predito que possam ser usadas como fonte de conhecimento no processo de aprendizado de máquina. Ao analisar as regras individualmente, observou-se que os melhores resultados foram obtidos utilizando como fonte de conhecimento, os lemas da primeira e da segunda palavra, a esquerda e a direita do verbo, o primeiro nome, o primeiro adjetivo, o primeiro verbo a esquerda e o a direita do verbo, e a primeira preposição a direita do verbo.

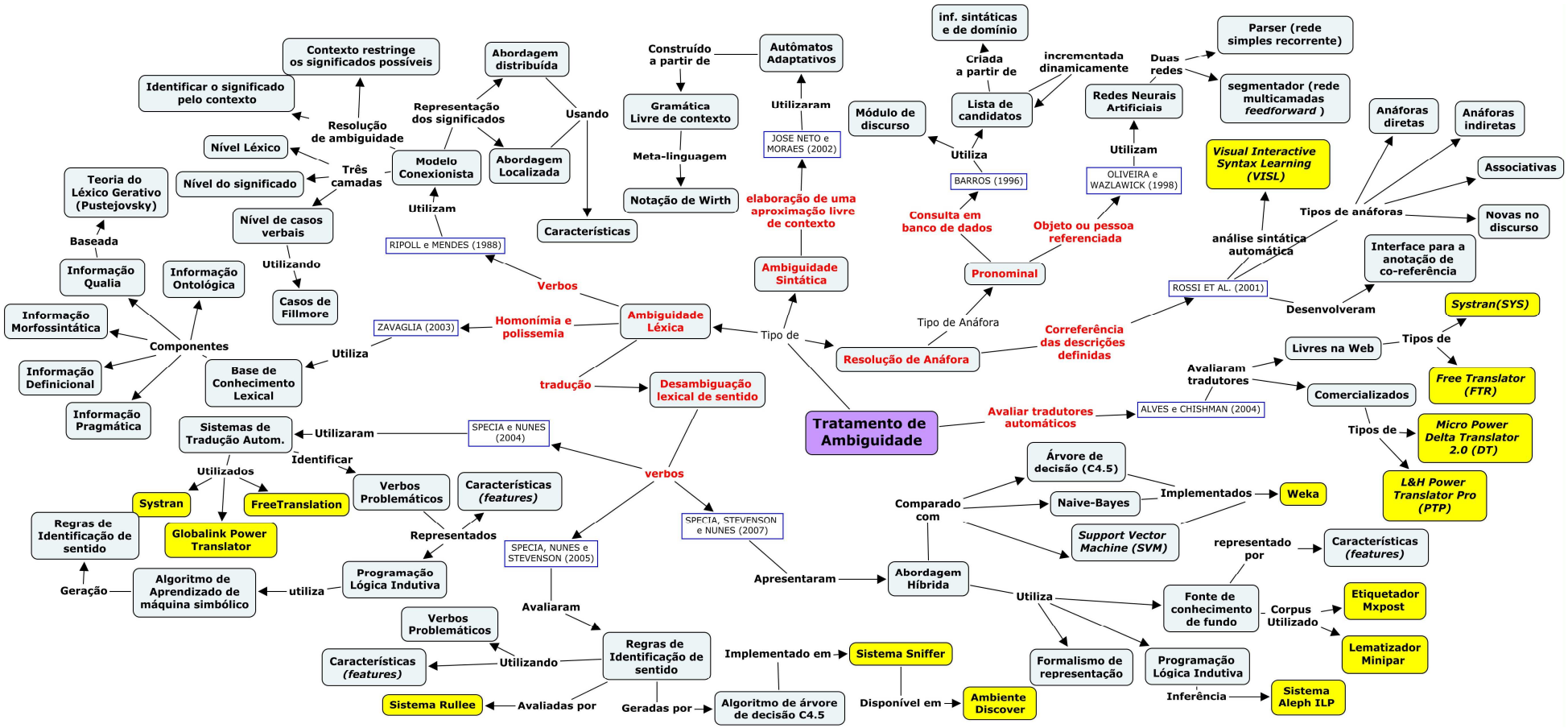

FIGURA 27 – Mapa conceitual apresentando as metodologias observadas nas publicações analisadas que tiveram como problemática TRATAMENTO DE AMBIGUIDADE

O trabalho de Specia, Stevenson e Nunes (2007) teve como objetivo apresentar uma abordagem de desambiguação lexical de sentido baseada em corpus com conhecimento de fundo (*background knowledge*), ou seja, considera o contexto próximo à palavra ambígua fazendo-se uso de uma lista de palavras. Neste trabalho será dada ênfase a tradução de dez verbos ambíguos do inglês para o português. Como fonte de conhecimento de fundo usada nos algoritmos de aprendizado, os autores experimentaram 12 alternativas diferentes de características, contendo *bag-of-words*, bigramas, palavras a direita e a esquerda do verbo, dentre outras.

A outra subproblemática observada dentro do tratamento de ambiguidade foi a resolução de anáforas que englobou três trabalhos, que serão discutidos a seguir.

O primeiro trabalho analisado sobre resolução de anáfora foi Barros (1996). Este artigo descreve um mecanismo para resolução de anáfora pronominal sem a utilização de modelo do mundo (*world models*), para garantir a portabilidade e ainda oferecer uma interface para consultas em linguagem natural em banco de dados. Assim, quando uma anáfora é encontrada numa consulta, os candidatos são selecionados tendo como base informações sintáticas e de domínio. Cabe ao usuário, escolher um dentre as opções ou rejeitar todas. Segundo a autora, este modelo provê um processo semi-automático de resolução de anáforas independente do domínio, mas não apresentou experimentos.

O próximo trabalho de Oliveira e Wazlawick (1998) discute o problema da ambiguidade diante da resolução de anáforas presente nos pronomes "ele" e "ela". Os autores propõem a utilização de redes neurais artificiais, usando como dados de treinamento, padrões tais como "sujeito verbo objeto. Ele/ela verbo objeto". O modelo é composto por duas redes neurais artificiais: o *parser* (rede simples recorrente) e o segmentador (rede multicamadas *feedforward*), cada uma com função específica. Os autores não apresentaram índices de acertos, mas afirmaram que a abordagem proposta resolveu eficientemente todos os exemplos apresentados que contêm a mesma estrutura dos que foram treinados.

O terceiro e último artigo analisado sobre resolução de anáforas foi o de Rossi *et al*. (2001), que teve como objetivo identificar as sequências de expressões em um texto que se referem a uma mesma entidade. Mais especificamente, investiga-se a correferência das descrições definidas (sintagmas nominais iniciados por artigo definido). Segundo os autores, quatro tipos de anáforas foram definidas: anáforas diretas, indiretas, associativas e novas no discurso. Os autores desenvolveram uma interface para a anotação manual de correferência em corpus da língua portuguesa, classificação e

contabilização dos tipos usados. Com esta classificação manual, os autores apresentaram um sistema (em Prolog) para o tratamento automático de correferência nominal, baseado no estudo feito manualmente. Os autores apresentaram resultados oriundos da anotação feita manualmente, mas não apresentaram resultados do sistema, pois o mesmo encontra-se em desenvolvimento.

Finalmente, o trabalho de Jose Neto e Moraes (2002) propõe efetuar uma redução inicial da complexidade da linguagem que se deseja definir, através da elaboração de uma aproximação livre de contexto da mesma. Os autores complementam que eliminando-se aspectos mais complexos da linguagem, tais como ambiguidades, pode-se obter uma boa aproximação da linguagem natural. A gramática simplificada (no formato de um autômato adaptativo) usada como base para o raciocínio não considera importantes aspectos de dependência de contexto, que certamente devem ser levados em conta em outras etapas do processamento da linguagem.

O trabalho de Alves e Chishman (2004) é um trabalho teórico e tem como objetivo mostrar como alguns tradutores automáticos tratam o complexo fenômeno linguístico da ambiguidade. Como resultado da análise do desempenho dos tradutores avaliados, as autoras afirmam que de uma maneira geral, os tradutores geralmente não percebem a tradução mais adequada para o contexto, e não indicam que pode haver outra possibilidade de tradução. As autoras concluem que apesar da tradução automática ter sido a primeira aplicação não numérica da computação (datada de 1949), o desempenho desses sistemas ainda está muito longe do que seria ideal.

Nenhum dos trabalhos analisadores que abordaram tratamento de ambiguidade realizou experimentos práticos, com exceção dos artigos de Specia, Nunes e Stevenson de 2005 e 2007.

4.2.2.4. Problemática ANALISADORES *(PARSER)*

Na FIG. 28 é apresentado o mapa conceitual construído a partir da metodologia adotada pelos artigos analisados sobre ANALISADORES *(PARSER)*.

A partir da análise de conteúdo realizada, foi possível identificar três níveis de análise da linguagem natural: análise léxico-morfológica, análise sintática e análise semântica. Além disso, observou-se que alguns trabalhos abordaram dois ou até mesmo três níveis de análise, o que justifica as linhas interceptadas no mapa conceitual apresentado na FIG. 28.

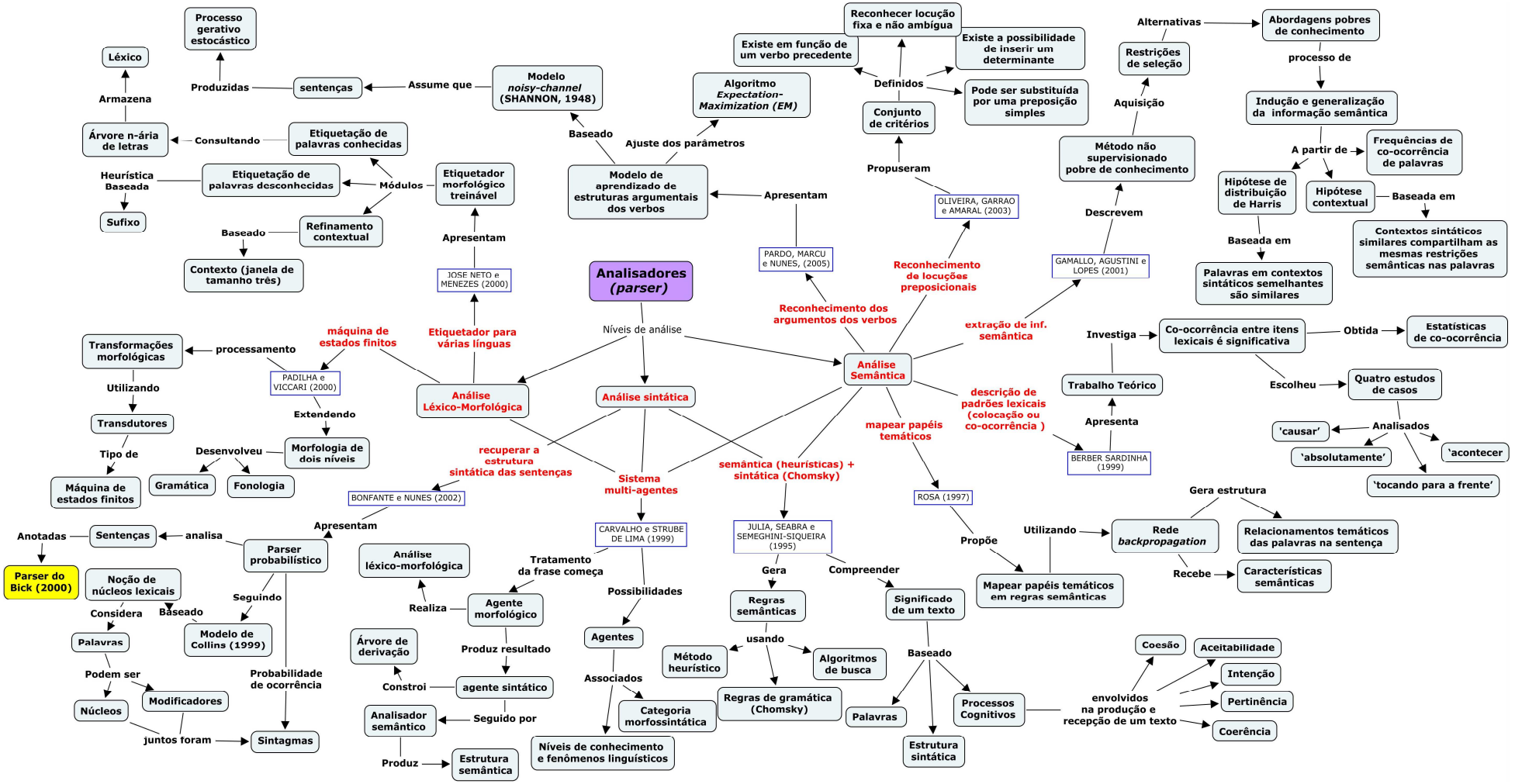

FIGURA 28 – Mapa conceitual apresentando as metodologias observadas nas publicações analisadas que tiveram como problemática ANALISADORES *(PARSER)*

Dentre os trabalhos que propõem análise léxico-morfológica estão Jose Neto e Menezes (2000) que propõe um método para a construção de um etiquetador morfológico, que possa ser usado em várias línguas. Segundo os autores, existem, basicamente, quatro paradigmas ou métodos de etiquetagem morfológica de textos em linguagem natural: o estatístico; o que se utiliza de regras escritas manualmente; o baseado em regras inferidas automaticamente; e o com base em exemplos memorizados. Segundo os autores, todos utilizam três fontes de informação linguística, extraídas de um corpus de treinamento: os sufixos de palavras, como parte do processo de inferência da etiqueta morfológica de palavras desconhecidas; uma lista de palavras associadas a categorias morfológicas (léxico), para fornecer informações sobre palavras conhecidas; e o contexto próximo ao item lexical que se quer etiquetar (2 ou 3 etiquetas ao redor), para refinar a escolha de sua etiqueta. Assim, o método proposto etiqueta primeiro as palavras conhecidas, depois as desconhecidas usando heurística de acordo com o sufixo, e finalmente faz um refinamento, de acordo com o contexto. Os experimentos revelaram uma taxa de acerto comparável a de outros trabalhos da época. No entanto, os autores argumentaram que o método baseado em exemplos memorizados começa a produzir resultados satisfatórios somente a partir de um corpus com 300.000 palavras. O outro trabalho sobre análise léxico-morfológica foi o de Padilha e Viccari (2000) que propõe o desenvolvimento de processadores para a morfologia do português utilizando máquinas de estados finitos (transdutores). Os autores apresentaram um trabalho teórico sem a realização de experimentos. Apesar disso, os autores alegam que os transdutores são adequados para o processamento morfológico da língua portuguesa, mas ressaltam como limitações a sua construção generativa (não há algoritmos de aprendizado de novas transformações, a gramática deve ser alterada e o transdutor reconstruído); e a ausência de pesos para diferenciar mapeamentos ambíguos.

Dentre os trabalhos que abordaram a análise sintática está o trabalho de Bonfante e Nunes (2002) que destacaram a importância de se recuperar a estrutura sintática das sentenças. Este trabalho apresenta parte da tese de doutorado da primeira autora (em desenvolvimento na época), que visa investigar o comportamento de analisadores sintáticos usando a abordagem empírica. O modelo proposto baseia-se na noção de núcleos lexicais, onde, para cada regra observada no conjunto de treinamento, as palavras que não são núcleo são chamadas de modificadores, exercendo influência sobre ele. Na época da publicação deste trabalho não haviam ainda resultados concretos, pois os experimentos estavam em andamento.

Já o próximo trabalho, abordou todos os níveis de análise usando um sistema

multi-agentes. Carvalho e Strube de Lima (1999) afirmam que vários trabalhos têm utilizado a abordagem sequencial, com processamentos associados aos diferentes níveis linguísticos. No entanto, os sistemas distribuídos apresentam-se como uma alternativa viável para o processamento da língua natural, uma vez que módulos autônomos, especializados e distribuídos podem se cooperar para resolver o problema. Para testar essas abordagens, as autoras propõem o desenvolvimento de dois sistemas: no primeiro, os agentes foram associados à categoria morfossintática das palavras e no segundo, os agentes foram associados a níveis de conhecimento e a fenômenos linguísticos. Sob o ponto de vista linguístico, o sistema faz análise léxico-morfológica, sintática e semântica. Segundo as autoras, o tratamento da frase começa com uma análise léxico-morfológica, através do agente morfológico, que envia seus resultados para o agente sintático, para que este possa construir a árvore de derivação; o agente sintático, por sua vez, envia seus resultados para o analisador semântico para a construção da estrutura semântica. As autoras não apresentaram resultados de experimentos realizados.

O trabalho de Julia, Seabra e Semeghini-Siqueira (1995) propõem um parser que realiza a análise sintática e semântica. O analisador proposto corresponde a uma estrutura, que gera automaticamente regras semânticas durante a análise, baseado em heurística. A parte sintática da gramática é expressa por meio de regras (gramática de Chomsky). Os autores não apresentaram experimentos envolvendo exemplos do analisador desenvolvido.

Os próximos trabalhos analisados abordam análise semântica da linguagem natural. Rosa (1997) teve como objetivo representar as palavras usando um conjunto de características semânticas que possuem um significado associado, e construir uma arquitetura capaz de analisar e aprender a atribuição correta dos relacionamentos temáticos das palavras nas sentenças. Assim, o autor propõe mapear papéis temáticos em regras semânticas usando vetores de características organizados com base nas relações temáticas entre o verbo e as outras palavras de uma frase. O autor conclui que a abordagem conexionista já provou ser eficaz no tratamento de construções lexicais, mas que o sistema apresenta bons resultados para os tipos de frases para o qual foi treinado. O trabalho seguinte, Berber Sardinha (1999), apresentou um trabalho teórico com relatos de um estudo cujo foco é a descrição de padrões lexicais e colocações do português. Os relatos apresentados visaram fornecer uma descrição dos perfis semânticos de várias palavras da língua portuguesa. Segundo o autor, a co-ocorrência entre os itens pesquisados é significativa, dependendo da estatística de co-ocorrência obtida (razão entre observado e esperado, a informação mútua e o escore T). O autor apresentou

quatro estudos de caso e, baseando-se nas medidas estatísticas definidas, discutiu a perfil semântico mais frequente.

O próximo trabalho analisado que aborda análise semântica foi o de Gamallo, Agustini e Lopes (2001) que tem como objetivo descrever um método baseado em corpus para a extração de informação semântica. Segundo os autores, o intuito é utilizar informações sintáticas para extrair as restrições de seleção e preferências semânticas ao invés de combinação de palavras. Em outras palavras, é apresentado um método não supervisionado "pobre de conhecimento" (usando apenas a co-ocorrência das palavras) para adquirir restrições de seleção baseado em hipóteses de contexto (para extrair similaridade das palavras) e de co-especificação (para definir os contextos sintáticos). O principal objetivo é calcular a frequência da co-ocorrência dentro de construções sintáticas, ou sequências de n-gramas, com o objetivo de extrair informações semânticas, tais como restrições de seleção e ontologias de palavras. Como resultados, os autores apresentaram alguns exemplos de agrupamentos gerados, destacando como o sentido de palavras polissêmicas é representado pela atribuição natural da palavra em vários agrupamentos.

O outro trabalho analisado que abordou análise semântica foi o de Oliveira, Garrao e Amaral (2003) que propuseram um conjunto de critérios aplicados às expressões para a detecção de locuções preposicionais. Como resultado, os autores apresentaram uma tabela listando as locuções preposicionais encontradas no corpus. Os autores concluíam que este trabalho realizou inicialmente uma pesquisa em corpus e que o critério deve ser implementado computacionalmente.

Finalmente, o último trabalho analisado que abordou análise semântica foi o de Pardo, Marcu e Nunes (2005) que apresentou uma abordagem não supervisionada, completamente automática, para o aprendizado das estruturas argumentais de verbos, utilizando-se um modelo estatístico gerativo baseado no modelo *noisy-channel* de Shannon (1948) e treinado por meio do algoritmo *Expectation-Maximization.* Para avaliar se as estruturas argumentais aprendidas são plausíveis ou não, dois experimentos foram realizados. Os resultados foram obtidos comparando algumas estruturas argumentais aprendidas pelo modelo proposto, com estruturas julgadas por humanos linguistas computacionais. Os índices médios de precisão e cobertura foram de 76% e 86%, respectivamente. Como aspectos positivos do modelo, os autores afirmam que o modelo é capaz de aprender estruturas argumentais com grande precisão, sem esforço de anotação, usando ferramentas relativamente simples. No entanto, ele não é capaz de lidar apropriadamente com estruturas complexas (sintagmas, por exemplo).

Em síntese, vale destacar que das problemáticas observadas, a que parece ter perdido espaço foi a de desenvolvimento de analisadores (o que sugere o número de trabalhos relacionados e as datas). Além disso, a maioria dos trabalhos apresenta modelos sem a realização de experimentos que comprovem a sua real aplicabilidade.

4.2.2.4. Problemática TRADUÇÃO

Na FIG. 29 é apresentado o mapa conceitual construído a partir da metodologia adotada pelos artigos analisados sobre TRADUÇÃO.

A partir da análise de conteúdo realizada nos artigos relacionados com a problemática tradução automática (*machine translation*), foi possível observar que existem algumas abordagens clássicas que têm sido usadas pelos pesquisadores: utilização de conhecimento linguístico para extrair regras de transferência (tradução), de métodos estatísticos e de alinhamento.

A tradução automática pode ser considerada uma tarefa difícil, principalmente por precisar de conhecimento linguístico profundo de várias linguagens. Os métodos estatísticos, por outro lado, também podem ser complicados por necessitarem de grandes *corpora* paralelos e alinhados (CASELI; NUNES, 2006). Apesar disso, os trabalhos analisados revelaram que as duas abordagens têm sido avaliadas.

Observou-se também que vários métodos têm sido desenvolvidos com o objetivo de encontrar automaticamente correspondências estruturais, sintáticas ou lexicais a partir de textos paralelos. Esses textos paralelos são etiquetados de tal maneira que alinhamentos sejam possíveis. Tais correspondências são usadas para construir gramáticas de tradução, no formato de regras de transferência, e para obter a probabilidade de um alinhamento ocorrer.

O primeiro trabalho que propõe o desenvolvimento de tradução completamente automática foi o de Caseli e Nunes (2006) que sugere a utilização de textos paralelos para induzir regras de transferências. No entanto, o trabalho encontra-se em andamento e não apresentaram resultados.

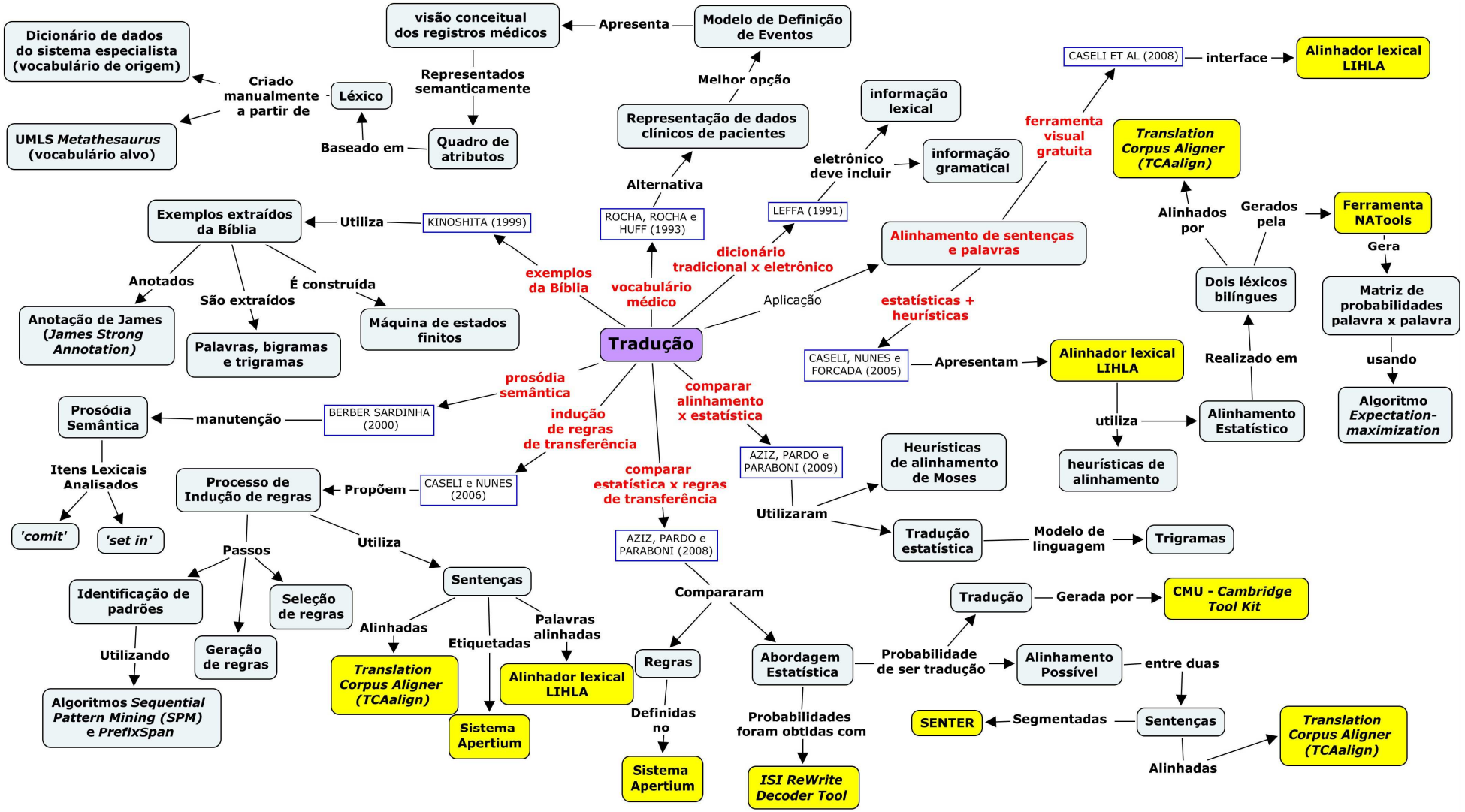

FIGURA 29 – Mapa conceitual apresentando as metodologias observadas nas publicações analisadas que tiveram como problemática TRADUÇÃO

Os próximos dois trabalhos foram publicados pelos mesmos autores. Aziz, Pardo e Paraboni (2008) afirmam que os métodos estatísticos têm sido amplamente utilizados e propõem compará-los com as regras de transferências, o que apresentou resultados muito próximos. Ao realizar uma avaliação manual dos métodos, os autores concluíram que a abordagem estatística exigiu um esforço maior para transformar a saída do sistema na tradução correta. Em 2009, os mesmos autores analisaram quatro parâmetros de configuração dos métodos estatísticos de tradução automática, e, de uma maneira geral, a diferença entre os resultados obtidos foi muito sutil, normalmente na terceira casa decimal. Ambos usaram o escore BLUE para avaliar os resultados. O escore BLUE representa o número de n-gramas compartilhadas entre a tradução automática e a referência usada e varia de 0 a 1. O sistema baseado em regras apresentou resultados sutilmente melhores que o método estatístico (0,6 e 0,58, respectivamente). O que sugere que os resultados obtidos foram ruins.

Os dois últimos trabalhos analisados abordaram o desenvolvimento de um alinhador lexical (LIHLA), que inclusive foi usado nos outros trabalhos e será apresentado no quadro de recursos.

Vale destacar que todos os trabalhos, que abordaram métodos automáticos de tradução envolvendo o português, utilizaram o corpus da revista de pesquisa da FAPESP com textos paralelos escritos em português do Brasil (original), e versões em inglês e espanhol.

4.2.2.5. Outras Problemáticas

As aplicações para a própria área, os exemplos de aplicação do PLN e correção automática foram incluídas numa mesma figura (FIG. 30), por terem apresentado menos trabalhos analisados, e serão discutidas seguindo o sentido anti-horário do mapa conceitual apresentado na FIG. 30. Na problemática das aplicações para a própria área foram incluídos os trabalhos que objetivaram desenvolver repositórios de recursos e ferramentas para o desenvolvimento de pesquisas na área de PLN. Vale ressaltar que nas demais problemáticas também foram relatadas pesquisas que produziram ferramentas e/ou recursos que poderiam e podem ser reutilizadas. No entanto, aqueles artigos tinham como objetivo discutir a(s) técnica(s) e método(s) utilizado(s) na construção do recurso, sendo este último, consequência da pesquisa. Já os trabalhos alocados dentro dessa problemática, a construção do recurso foi apresentada como sendo o foco na pesquisa e não coadjuvante.

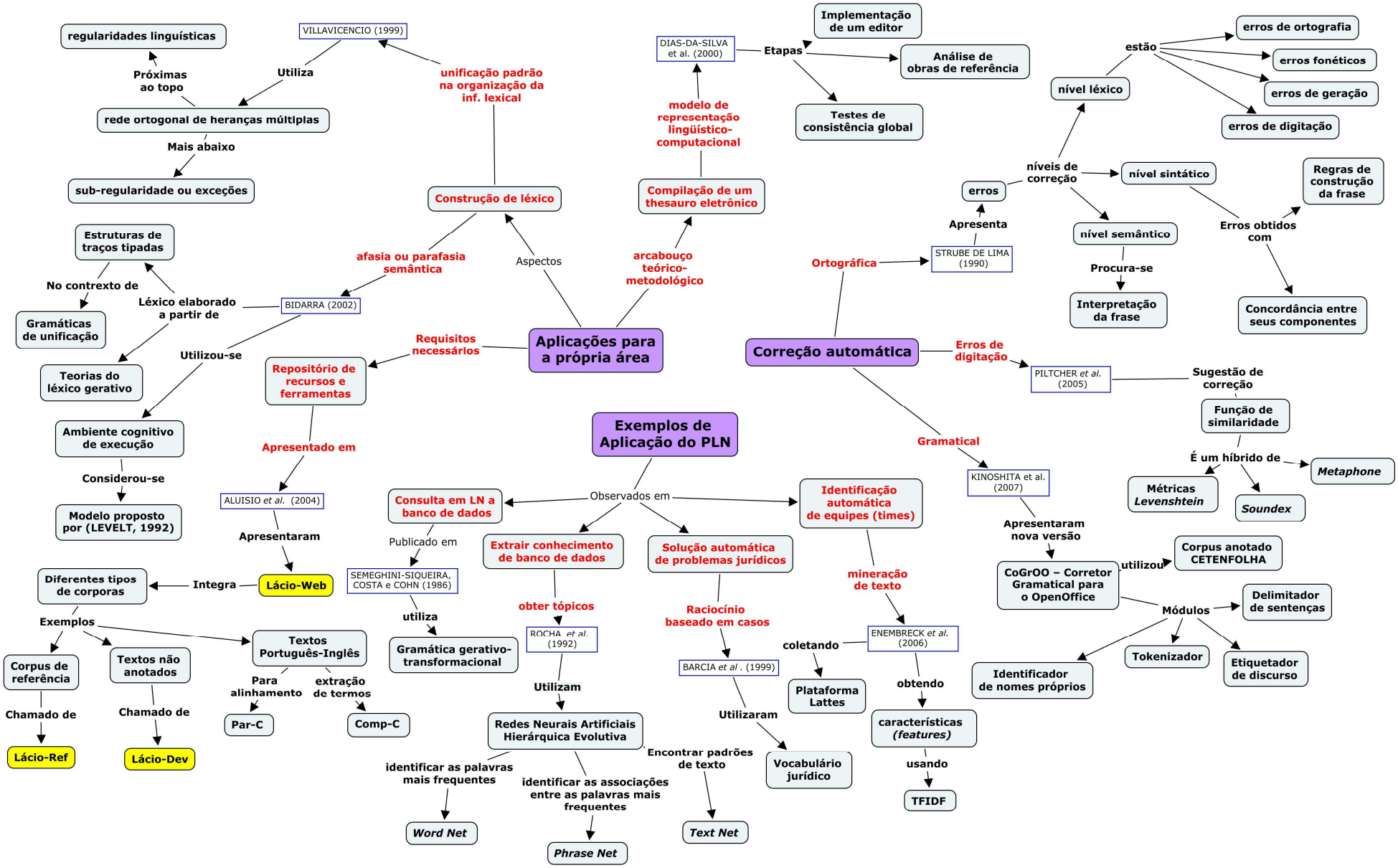

FIGURA 30 – Mapa conceitual apresentando as metodologias observadas nas publicações analisadas que tiveram como problemática OUTRAS

Dois trabalhos analisados tiveram como objetivo a construção de um léxico: Villavicencio (1999) demonstrou como o uso de unificação padrão na organização da informação lexical pode fornecer descrição não redundante de tipos lexical. Segundo a autora, padrões foram utilizados para estruturar o léxico, concentrando-se na descrição das informações de categorização verbal. A autora não apresentou experimentos, mas considerou que da utilização do padrão de herança, para representar informações sobre subcategorização verbal, é possível obter uma hierarquia altamente estruturada e sucinta.

O outro trabalho que descreve a construção de léxicos foi o de Bidarra (2002) que considerou alguns aspectos básicos para a construção de léxicos para o PLN, relacionados à afasia ou parafasia semântica. Por ser um trabalho essencialmente teórico e descritivo, não foram realizados experimentos, mas o autor apresentou alguns exemplos de afasia, em português, para ilustrar o modelo proposto.

O trabalho de Dias-da-Silva *et al*. (2000) abordou inúmeras questões envolvidas no processo de compilação de um Thesaurus Eletrônico Básico para o Português do Brasil (TeP). Foram apresentadas algumas telas do editor de thesaurus construído, ilustrando algumas entradas fornecidas.

O último trabalho analisado foi o de Aluisio *et al*. (2004) que tem como objetivo discutir os requisitos necessários para se construir um grande repositório de recursos e ferramentas, e apresentar o corpora Lácio-Web, projeto em desenvolvimento desde 2002, na universidade de São Paulo (NILC, IME e FFLCH). Segundo os autores, o Lácio-Web foi projetado tanto para pesquisadores linguísticos teóricos como práticos, e para o desenvolvimento tanto de ferramentas linguísticas computacionais como de aplicações, tais como etiquetadores *(tagger),* analisadores *(parsers),* corretores gramaticais *(grammar checkers),* métodos de recuperação de informação e sumarização automática.

Dentro da problemática dos exemplos de aplicação do PLN, foram incluídos os trabalhos que tiveram o PLN como atividade meio e não como atividade fim da pesquisa. Assim, serão apresentados trabalhos que usaram métodos e técnicas do PLN para resolver problemas diversos, tais como consultas em banco de dados ou solução automática de problemas jurídicos.

O primeiro trabalho analisado incluído na problemática de exemplos de aplicação do PLN foi também o primeiro artigo analisado: Semeghini-Siqueira, Costa

e Cohn (1986). Os autores apresentam um discurso muito próximo do que foi observado na leitura do ARIST: o processamento de linguagem natural está engatinhando em relação a outros campos pesquisa, tais como inteligência artificial, e sofre forte influência das teorias de Chomsky, e de gramáticas formais (gerativa e transformacional). Apesar dos autores passarem a proposta conexionista no título e citar a inteligência artificial no início do artigo (sugeriu a utilização de redes neurais), eles descrevem um sistema em Prolog, com a finalidade de facilitar a consulta a uma base de dados relacional usando linguagem natural. Como resultado, os autores apresentam várias perguntas em português que o sistema foi capaz de compreender.

O próximo trabalho analisado foi o de Rocha *et al*. (1992) que teve como objetivo apresentar um sistema com rede neural artificial hierárquica, capaz de compreender o conteúdo de textos e produzir listas de tópicos a partir de registros de banco de dados. Os autores concluíram que o conhecimento adquirido pelo sistema desenvolvido em bases de dados especialistas pode ser usado para construir sistemas especialistas na área médica.

O terceiro e último trabalho analisado que apresentou exemplos de aplicação do PLN foi o de Barcia *et al*. (1999) que propôs a utilização da técnica de Raciocínio baseado em Casos (RBC) para solução de problemas jurídicos. Os casos jurídicos foram representados na forma de um caso, que consiste no texto do documento original e um conjunto de índices na forma de pares atributo-valor. Os atributos dos documentos textuais, usados como índices para a recuperação devem indicar a utilidade das informações do caso na situação presente. Para reforçar esta forma de representação, o conhecimento do domínio foi incluído na forma de um vocabulário jurídico controlado e um dicionário de termos.

Dentro da problemática de correção automática, o primeiro trabalho analisado foi o de Strube de Lima (1990) que apresentou uma revisão de literatura, que teve como objetivo dar ao leitor uma visão panorâmica no que se refere ao tema correção ortográfica automatizada, apresentando as técnicas e os métodos empregados na época no tratamento da língua natural, abordando vantagens e deficiências.

O próximo trabalho analisado dentre dessa problemática foi o de Piltcher *et al*. (2005) que tratam a correção de palavras dentro de um ambiente de chat (salas de bate-papo). A abordagem utilizada é probabilística (estatística) e não

requer a utilização de analisadores sintáticos (*parsers*). Para calcular a similaridade entre as palavras, utilizou-se as métricas Levenshtein, Metaphone e Soundex para correção automática de erros de digitação. Analisando as três métricas utilizadas, observou-se a técnica de Levenshtein apresentou precisão bem aquém das demais técnicas, apesar de ter apresentado a melhor abrangência. As demais técnicas, mesmo tendo uma boa precisão, obtiveram uma abrangência pouco significativa (entre 3% e 7%).

O terceiro e último artigo analisado dentro da problemática de correção automática foi o Kinoshita *et al*. (2007) que apresentaram o CoGrOO – Corretor Gramatical para o OpenOffice. Os autores compararam o CoGrOO com o ReGra, corretor gramatical do editor de texto Microsoft Word, detectaram 7 erros em comum. CoGrOO detectou 8 erros que o Regra não detectou, por outro lado, o Regra detectou 7 que o CoGrOO não. Os autores concluem que a arquitetura do CoGrOO é híbrida e mescla o uso de regras simbólicas e estatísticas com aprendizado de máquina baseado em treinamento com corpus anotado.

4.2.2.6. Ferramentas utilizadas e corpora

A partir dessa análise foi possível identificar as ferramentas computacionais utilizadas pelos autores dos artigos analisados durante a realização dos experimentos. Uma síntese desses recursos é apresentada na TAB. 14.

TABELA 14
Ferramentas utilizadas pelos artigos submetidos à análise de conteúdo

Ferramenta	Descrição	Citado em	Problemática
PreText		Matsubara, Monard e Batista (2004)	Recuperação de Informação
		Braga, Monard e Matsubara (2009)	Recuperação de Informação
		Martins, Monard e Matsubara (2003)	Recuperação de Informação
PALAVRAS (BICK, 2000)		Silva e Vieira (2007)	Recuperação de Informação
		Silva, Vieira e Osorio (2005)	Recuperação de Informação
		Caminada, Quental e Garrao (2008)	Recuperação de Informação
		Seno e Nunes (2009)	Sumarização
		Bonfante e Nunes (2002)	Analisadores *(parser)*
Algoritmo de Porter		Silva e Vieira (2007)	Recuperação de Informação
		Silva, Vieira e Osorio (2005)	Recuperação de Informação

	Braga, Monard e Matsubara (2009)	Recuperação de Informação
Ferramenta WEKA	Silva e Vieira (2007)	Recuperação de Informação
	Silva, Vieira e Osorio (2005)	Recuperação de Informação
	Leite e Nunes (2006)	Sumarização
	Specia, Stevenson e Nunes (2007)	Tratamento de Ambiguidade
Ferramentas CHAMA e FORMA	Moraes e Strube De Lima (2007)	Recuperação de Informação
SENTER	Seno e Nunes (2008)	Recuperação de Informação
	Aziz, Pardo e Paraboni (2008)	Tradução
Clusterização incremental Singlepass	Seno e Nunes (2008)	Recuperação de Informação
Palavras Xtractor	Silva, Vieira e Osorio (2005)	Recuperação de Informação
Apertium	Villavicencio, Caseli e Machado (2009)	Recuperação de Informação
	Caseli e Nunes (2006)	Tradução
	Aziz, Pardo e Paraboni (2008)	Tradução
Translation Corpus Aligner (TCAligner)	Villavicencio, Caseli e Machado (2009)	Recuperação de Informação
	Caseli e Nunes (2006)	Tradução
	Aziz, Pardo e Paraboni (2008)	Tradução
	Caseli, Nunes e Forcada (2005)	Tradução
Base de sinônimos Tep 1.0 (MAZIERO *et al*., 2008)	Seno e Nunes (2009)	Sumarização
Gerador de formas superficiais (CASELI, 2007)	Seno e Nunes (2009)	Sumarização
sistema jNina (PEREIRA e PARABONI, 2007)	Seno e Nunes (2009)	Sumarização
Alinhador lexical LIHLA	Caseli e Nunes (2006)	Tradução
	Caseli, Nunes e Forcada (2005)	Tradução
	Caseli *et al* (2008)	Tradução
Visual Interactive Syntax Learning (VISL)	Rossi *et al*. (2001)	Tratamento de Ambiguidade
Tradutor Systran(SYS)	Alves e Chishman (2004)	Tratamento de Ambiguidade
	Specia e Nunes (2004)	Tratamento de Ambiguidade
Tradutor Free Translator (FTR)	Alves e Chishman (2004)	Tratamento de Ambiguidade
	Specia e Nunes (2004)	Tratamento de Ambiguidade
Tradutor Micro Power Delta Translator 2.0 (DT)	Alves e Chishman (2004)	Tratamento de Ambiguidade
Tradutor L&H Power Translator Pro (PTP)	Alves e Chishman (2004)	Tratamento de Ambiguidade
Software Part-of-Speech (BRILL, 1992)	Larocca Neto *et al*. (2000)	Sumarização
Tokennizer (ALUÍSIO; AIRES, 2000)	Souza, Pereira e Nunes (2001)	Sumarização
Tagger (ALUÍSIO; AIRES, 2000)	Souza, Pereira e Nunes (2001)	Sumarização
Algoritmo Extractor (TURNEY, 1999)	Souza, Pereira e Nunes (2001)	Sumarização
Sistema Rullee	Specia, Nunes e Stevenson (2005)	Tratamento de Ambiguidade
Sistema Sniffer	Specia, Nunes e Stevenson (2005)	Tratamento de Ambiguidade
Ambiente Discover	Specia, Nunes e Stevenson (2005)	Tratamento de Ambiguidade

Sistema Aleph ILP	Specia, Stevenson e Nunes (2007)	Tratamento de Ambiguidade
Lematizador Minipar	Specia, Stevenson e Nunes (2007)	Tratamento de Ambiguidade
Etiquetador Mxpost	Specia, Stevenson e Nunes (2007)	Tratamento de Ambiguidade
Ferramenta NATools	Caseli, Nunes e Forcada (2005)	Tradução
ISI ReWrite Decoder Tool	Aziz, Pardo e Paraboni (2008)	Tradução
CMU - Cambridge Tool Kit	Aziz, Pardo e Paraboni (2008)	Tradução
Lácio-Web	Aluisio *et al.* (2004)	Aplicações para a própria área
Lácio-Ref	Aluisio *et al.* (2004)	Aplicações para a própria área
Lácio-Dev	Aluisio *et al.* (2004)	Aplicações para a própria área

Analisando o material empírico usado nos artigos submetidos à análise de conteúdo, foi possível observar que no desenvolvimento de pesquisas na área de PLN, alguns autores optaram por construir bases de documentos especificamente para os experimentos realizados no próprio trabalho, enquanto que outros optaram por reutilizar alguma base de pesquisas anteriores.

Observou-se que 60% dos trabalhos que utilizaram um corpus de documentos, optaram por reutilizar algum construído em trabalhos anteriores, enquanto que 40% foram construídos especificamente para o trabalho analisado. Além disso, constatou-se que à medida que a área foi se desenvolvendo, os corpora passaram a ser reutilizados com maior frequência: a primeira ocorrência de reutilização de corpus foi observada em trabalhos publicados no ano de 2.000, e a partir do ano 2.002, 80% dos trabalhos que fizeram uso de corpus, o fizeram de algum outro conhecido.

Na TAB. 15, é apresentado o catálogo de recursos disponíveis para o desenvolvimento de pesquisas na área de PLN, que foram citados nos artigos analisados. Para cada corpus, são apresentados, além do nome, uma pequena descrição, o idioma e as referências das publicações que o utilizaram. Os corpora foram listados obedecendo a ordem de citação: do mais antigo para o mais recente.

TABELA 15
Corpora de documentos utilizados pelos artigos submetidos à análise de conteúdo

Corpus	Idioma	Citado em
Penn Treebank	inglês	Jose Neto e Menezes (2000)
Tycho Brahe	português	Jose Neto e Menezes (2000)
Base de documentos TIPSTER	inglês	Larocca Neto *et al*. (2000)
P.G.R. - *Portuguese General Attorney Opinions*	português	Gamallo, Agustini e Lopes (2001)
Theses Corpus (PARDO, 2002)	português	Pardo e Rino (2002)
corpus do NILC – Núcleo Interdisciplinar de Linguística Computacional	português	Bonfante e Nunes (2002) Martins, Monard e Matsubara (2003) Gasperin e Strube de Lima (2003) Oliveira, Garrao e Amaral (2003) Silva, Vieira e Osorio (2005) Seno e Nunes (2009)
CETENfolha	português	Alves e Chishman (2004)
ZERO Hora (Brasil)	português	Alves e Chishman (2004)
CETENpublico	português	Alves e Chishman (2004)
COMPARA (Portugal)	português	Alves e Chishman (2004) Specia, Nunes e Stevenson (2005)
TeMário	português	Rino et al. (2004) Rino e Seno (2006)
Rhetalho	português	Rino e Seno (2006)
TREC'2002 (*Text REtrieval Conference*)	inglês	Pardo, Marcu e Nunes (2005)
CorpusFAPESP	português – inglês – espanhol	Caseli, Nunes e Forcada (2005) Caseli e Nunes (2006) Aziz, Pardo e Paraboni (2009)
Documentos da CLEF - *Cross Language Evaluation Forum*	português	Balage Filho *et al*. (2006)
PLN-BR CATEG	português	Moraes e Strube de Lima (2007)
Senseval-3	inglês	Specia, Stevenson e Nunes (2007)
Biblioteca Digital da ACM	inglês	Salles *et al*. (2009)
Base de dados MedLine	inglês	Salles *et al*. (2009)

5. Conclusão

A Plataforma Lattes do CNPq permitiu ter uma visão panorâmica da produção científica nacional na área de processamento de linguagem natural. No entanto, fez-se necessário discutir as dificuldades encontradas durante essa coleta. A maioria dos problemas encontrados está relacionada à inconsistência dos dados fornecidos pelos próprios pesquisadores. O pesquisador muitas vezes não imagina que a Plataforma Lattes pode ser usada como fonte de pesquisa para caracterizar uma área, ou mesmo uma instituição. Algumas inconsistências foram identificadas por processo automático, sendo que a correção na maioria das vezes foi feita manualmente. Apesar desses obstáculos, não se imagina uma fonte alternativa que seja melhor, ou pelo menos, similar a Plataforma Lattes para fornecer um retrato da pesquisa nacional.

Diante do volume de publicações obtidas a partir dos currículos cadastrados na plataforma tornou-se fundamental construir um parâmetro conceitual que permitisse identificar a produção nacional sobre processamento de linguagem natural. A construção desse instrumento de seleção automática permitiu minimizar as interferências intrínsecas em um processo manual de indexação. Assim procurou-se voltar o olhar para pesquisas sabidamente reconhecidas, e usar esse reconhecimento como garantia literária do critério. Neste sentido, a análise de assunto realizada em onze capítulos de revisão do ARIST propiciou a elaboração de um parâmetro conceitual de atinência para a área de processamento de linguagem natural. Tendo em vista que este processo é impregnado de subjetividade, técnicas estatísticas foram aplicadas com o intuito de avaliar a qualidade do instrumento criado.

As publicações consideradas atinentes foram submetidas a uma análise horizontal, baseando-se apenas nas suas características descritivas, obtidas na Plataforma Lattes. No entanto, para que os objetivos desta tese fossem alcançados, era imprescindível adentrar no conteúdo dessas publicações, para caracterizar a produção nacional sobre processamento de linguagem natural. Assim, por meio de critérios estatísticos, definiu-se uma amostra representativa de documentos que foi submetida à análise de conteúdo. Essa análise permitiu aprofundar em temáticas que as características descritivas das publicações, aquelas coletadas dos currículos

da Plataforma Lattes, não permitiriam revelar. Essa amostra de publicações foi obtida considerando-se todo o período produtivo dos pesquisadores cadastrados na Plataforma Lattes, sem descartar as mais antigas, mas priorizando a atualidade (anos 2.000), por incluir a maioria das publicações.

Apesar da importância inegável de se aprofundar nas publicações selecionadas, a análise horizontal apresentou alguns fatos que confirmaram algumas hipóteses, ou apresentaram constatações desafiadoras.

Ao analisar todas as publicações atinentes para a área de PLN, pode-se observar que a área passou por um “boom” no início dos anos 2.000, sendo que a grande maioria (70%) da produção científica foi publicada depois deste marco. Analisando a área de vinculação dos autores, foi possível observar que 64% das publicações envolveram pesquisadores de várias áreas.

A participação da ciência da informação na área de PLN é muito modesta, sendo que a ciência da computação e a linguística justas foram responsáveis por quase 85% da produção nacional. Além disso, na década de oitenta a ciência da computação foi o campo disciplinar mais produtivo na área de PLN, enquanto que a década de noventa foi a mais produtiva para a linguística. A ciência da informação, na década de noventa, recuou a sua contribuição para a área tentando recompor nos anos 2.000.

Analisando as personalidades nacionais que mais publicaram na área de PLN, pode-se observar que doze pesquisadores foram responsáveis por mais de 20% de toda a produção nacional, sendo que dentre eles, nove são da ciência da computação, dois da linguística, e um é da engenharia elétrica. Ou seja, dentre a elite de pesquisadores na área de PLN não se encontra nenhum pesquisador declaradamente da ciência da informação. Além disso, vale destacar que dentre esses doze pesquisadores, sete fazem parte do grupo de pesquisa NILC formado por cientistas da computação e da linguística da USP, UFScar e UNESP.

Dentre as problemáticas mais abordadas, foi possível observar que: a tradução foi intensamente abordada na década de 90; os estudos com indexação diminuíram consideravelmente a partir da década de 80; e que as pesquisas sobre classificação passaram por um período de dormência na década de 90; e que existe uma tendência clara na área de PLN de desenvolvimento de pesquisas em sumarização automática.

Outro aspecto que a pesquisa revelou foi que a ciência da informação tem priorizado as pesquisas em indexação automática, seguido da análise de conteúdo. Já a ciência da computação tem priorizado as pesquisas em tradução e sumarização, enquanto que a linguística não tem priorizado o desenvolvimento de aplicações e sim de estudo relacionados ao léxico, o que sugere que os trabalhos da linguística tenham um cunho mais teórico.

Ao adentrar nos trabalhos selecionados foi possível constatar uma consequência natural do processo de amadurecimento no qual as pesquisas tendem a passar: as bases de documentos utilizadas em experimentos práticos começaram a ser reutilizadas com maior frequência. Após o ano 2.002, 80% dos trabalhos que fizeram uso de corpus, fizeram-no reutilizando-o de algum outro trabalho. Além disso, depois dos anos 2.000, a maioria das publicações analisadas (74%) apresentou experimentos práticos, o que sugere que as pesquisas na área de PLN, ao longo dos anos, têm apresentado um enfoque mais experimental, refletindo a forte inserção da área da ciência da computação. Observou-se também que a grande maioria (86%) dos trabalhos que apresentaram experimentos, tinha pelo menos um autor da computação, enquanto que, dos artigos que possuíam pelo menos um autor da linguística, a maioria (70%) não apresentou experimentos práticos. Outro aspecto interessante que a análise vertical permitiu avaliar foi os métodos de avaliação adotados nos trabalhos: observou-se que uma pequena parcela (35%) dos trabalhos fez uso de métodos automáticos de avaliação, enquanto que a maioria (65%) apresentou algum processo de validação manual, com verificação humana dos resultados obtidos. Como já era de se esperar, a maioria dos trabalhos analisados (65%) foi desenvolvida para o português, enquanto que 20% eram para o inglês, e 15% para abordagens genéricas (ou seja, independente da linguagem natural foco).

Dentre os artigos sorteados para análise de conteúdo, nenhum deles foi escrito por pesquisadores da ciência da informação, 12% possuíam somente autores da linguística, 75% somente por autores da ciência da computação, 6% envolvendo pesquisadores das duas áreas (computação e linguística), e 7 % de outras áreas.

A análise de conteúdo realizada nas 68 publicações selecionadas permitiu construir um substrato metodológico para a área de processamento de linguagem natural para cada problemática revelada. Dentre os artigos analisados, observou-se

que a **recuperação de informação** foi a problemática que sem dúvida teve maior destaque na produção científica nacional. A grande maioria desses artigos são trabalhos recentes, o que sugere que a área possa estar sendo impulsionada pelo próprio desenvolvimento da web. Além disso, a maioria dos artigos analisados sobre recuperação de informação está voltada para técnicas de pré-processamento de documentos, o que sugere que este tema ainda esteja em aberto.

Dentre as técnicas de pré-processamento dos documentos, observou-se que existe uma tendência forte em representar os documentos por meio de uma estrutura atributo-valor, que sintetiza as características extraídas do documento. Observou-se que os critérios usados para construir esta representação tem se modificado ao longo dos anos, apesar de não estarem alcançando melhorias significativas. Inicialmente usava-se frequência de palavras, posição da sentença no texto, frequência de palavras temáticas, ocorrência de palavras-chaves, das palavras do título, dentre outras. Trabalhos mais recentes têm avaliado o uso de categorias gramaticais (nome, verbo, e outras), palavras relacionadas (sinônimos, hipônimos, dentre outras), relacionamento textual (dado pelo grau de coesão), dentre outras alternativas. Além disso, a análise dos artigos selecionados revelou que combinar unigramas e bigramas não tem apresentado melhorias significativas na classificação supervisionada, nem mesmo em algoritmos semi-supervisionados, quando comparado as estratégias tradicionais (*stopwords* e *stemming*).

Um aspecto interessante observado durante a análise foi a possibilidade da relevância de um conceito mudar ao longo do tempo, visto que este último é uma dimensão importante para qualquer espaço informacional (SALLES *et al,,* 2009). Esse aspecto foi tratado incorporando, no classificador, características temporais do documento e da consulta do usuário. Essa preocupação não foi observada em nenhum outro artigo analisado dentre a produção científica nacional, nem mesmo no ARIST.

Ao analisar os trabalhos sobre recuperação de informação em documentos textuais, foi possível observar a utilização de vários algoritmos de aprendizado de máquina supervisionado (tais como redes neurais, árvore de decisão, regras de associação e SVM), não supervisionado (tais como *k-means*, KNN), além de algoritmos de aprendizado semi-supervisionado (tais como *co-training* e *self-training*). Vale destacar também a utilização recorrente de unigramas e

bigramas para representar os documentos, juntamente com medidas estatísticas e categóricas (substantivos, adjetivos, advérbios, nomes próprios, etc). Para o cálculo do grau de similaridade, observou-se a utilização das medidas de *Word overlap*, medidas de Jaccard, TFIDF e TFISF.

Quanto a realização de experimentos, vale destacar que a grande maioria dos trabalhos que abordaram recuperação de informação apresentaram experimentos práticos e utilizaram como métricas de avaliação pelos menos as medidas precisão e revocação. Merece destaque o fato dos resultados obtidos em alguns trabalhos terem sido muito ruins, em outras palavras, não apresentarem melhorias significativas (MORAES; STRUBE DE LIMA, 2007; SALLES *et al.*, 2009; BRAGA; MONARD; MATSUBARA, 2009; SENO; NUNES, 2008).

Dentro da problemática **sumarização**, dos trabalhos analisados, somente dois usaram a abordagem profunda e produziram sumários, o que sugere que a maioria das pesquisas em sumarização automática tem privilegiado a abordagem empírica (para gerar extratos), alternando-se diferentes características extraídas dos textos. Mais uma vez, merece destaque o fato de todos os artigos analisados apresentarem experimentos práticos, mas alguns com resultados pouco expressivos (RINO; NUNES, 2003; RINO *et al.,* 2004; RINO; SENO, 2008; AZIZ; PARDO; PARABONI, 2009).

O **tratamento de ambiguidade**, por sua vez, foi alvo de pesquisa por todo o período analisado, o que talvez sinalize para uma impossibilidade de a linguagem natural ser interpretada completamente por métodos automáticos. Os próprios autores analisados destacaram que, o problema da homonímia gramatical é resolvido facilmente por sistemas computacionais, mas o mesmo não acontece com outros problemas da ambiguidades, tais como homonímia semântica e a polissemia. Os trabalhos analisados que abordaram o tratamento da ambiguidade não apresentaram experimentos práticos com exceção de dois trabalhos de Specia, Nunes e Stevenson de 2005 e 2007.

Das problemáticas observadas, a que parece ter perdido espaço foi a de desenvolvimento de **analisadores *(parsers)*,** tendo em vista o número de trabalhos analisados e as datas de publicação. A maioria dos trabalhos apresenta modelos teóricos, sem a realização de experimentos que comprovem a sua real aplicabilidade.

A pesquisa em **tradução automática** continua em evidência, e dentre as abordagens clássicas usadas pelos pesquisadores estão: a utilização de conhecimento linguístico para extrair regras de transferência (tradução), de métodos estatísticos e de alinhamento. Os métodos estatísticos têm sido amplamente utilizados, mas ao compará-los com as regras de transferências, apresentaram-se resultados muito próximos. No entanto, a abordagem estatística exigiu um esforço maior para transformar a saída do sistema na tradução correta. Ao comparar parâmetros de configuração dos métodos estatísticos de tradução automática, a diferença entre os resultados obtidos foi muito sutil, normalmente na terceira casa decimal. Isso sugere que, apesar da pesquisa ter sido priorizada nos últimos anos, os resultados apresentados ainda são pouco expressivos.

Vale destacar que todos os trabalhos que abordaram métodos automáticos de tradução envolvendo o português, utilizaram o corpus da revista de pesquisa da FAPESP com textos paralelos escritos em português do Brasil (original), e versões em inglês e espanhol. As outras bases de documentos, assim como as ferramentas utilizadas pelos trabalhos analisados e discutidos nesta problemática, foram apresentadas para compor um *framework* da pesquisa em processamento de linguagem natural. Assim, espera-se que esta tese seja usada como ponto de partida para aqueles que almejarem aventurar-se pela área de PLN.

Diante dos resultados que emergiram da análise de conteúdo das publicações analisadas, alguns resgates no ARIST se fazem oportunos. Walker no capítulo de 1973 afirmou que inúmeros linguistas estavam questionando não somente a relevância dos resultados das pesquisas, como também a existência de tais resultados. Alguns trabalhos analisados apontam nesta direção: dificuldade de se obter resultados que reflitam melhorias significativas (MORAES; STRUBE DE LIMA, 2007; PARDO; RINO; NUNES, 2003; AZIZ; PARDO; PARABONI, 2008).

Simmons no capítulo de 1966 afirmou que o principal problema ainda não resolvido das pesquisas de classificação era utilizar a abordagem estatística para classificar não apenas 500 e sim 50 mil ou 150 mil linhas. Os resultados apresentados sugerem que essa dificuldade já foi contornada diante dos vários trabalhos que apresentam testes envolvendo milhares de documentos e milhões de palavras (CAMINADA; QUENTAL; GARRAO, 2008; SALLES *et al., 2009;* SENO; NUNES, 2009).

Damerau em seu capítulo de 1976 considerou o analisador *(parser)* como sendo um dos principais componentes de um sistema de processamento automático de linguagem e que continuaria sendo um tema substancial de pesquisas. Os artigos analisados mostraram que o tema foi priorizado no final dos anos 90 e início dos anos 2.000, mas não apresentaram experimentos práticos.

Warner em seu capítulo de revisão 1987 questionou se seria possível construir um sistema de linguagem natural completamente robusto, e em que medida este eventual sucesso dependeria de boas práticas de engenharia, e do conhecimento de processos cognitivos humanos. Diante dos resultados desta tese, cabe incluir outro questionamento: em que medida tal sucesso dependeria da própria linguagem. Em outras palavras, será que a linguagem natural, com suas especificidades, permite que métodos completamente automáticos sejam robustos no processo de reconhecimento de, por exemplo, regionalismos e gírias?

Simmons no primeiro capítulo de revisão do ARIST analisado (de 1966) afirmou que tradução automática era um objetivo distante de ser alcançado. Chowdhury, autor do último capítulo de revisão do ARIST sobre PLN (de 2003) afirmou que o projeto de sistemas de tradução automática é um trabalho duro, e que tal problema estaria longe de ser resolvido. A partir dos trabalhos analisados, observa-se que é inegável o desenvolvimento alcançado pelas pesquisas em tradução automática, o que pode ser comprovado pela existência de vários tradutores publicados e reutilizados (ALVES; CHISHMAN, 2004), apesar dos resultados alcançados ainda serem merecedores de novas investidas.

Simmons, em 1966, afirmou que apesar de um progresso significativo ter sido feito em pesquisas em respondedores automáticos, sistemas de perguntas e respostas em linguagem natural completamente automáticos era um objetivo muito distante, além de depender da realização de toda a área de processamento de linguagem. Dentre os trabalhos analisados, nenhum abordou diretamente o desenvolvimento de respondedores automáticos, apenas avaliou a utilização de um sumarizador (gerador de extratos) na tarefa de responder perguntas (BALAGE FILHO *et al.,* 2006).

Embora as pesquisas estejam ocorrendo em campos diferentes da CI, verificou-se que os estudos feitos devem ser conhecidos, pois podem contribuir com a área. Dentre os trabalhos analisados, merecem ser destacados: sobre

categorização (SILVA; VIEIRA, 2007), sobre mineração de texto (SILVA; VIEIRA; OSORIO, 2005), sintagmas nominais (ROSSI *et al.*, 2001; OLIVEIRA; GARRAO; AMARAL, 2003), índices para termos indexadores (ZIVIANE; ALBUQUERQUE, 1987), identificação de autoridades (MILIDIU, DUARTE; CAVALCANTE, 2007) e textos similares (SENO; NUNES, 2008). Novas explorações devem ser feitas pela CI no sentido de adequação dos sistemas de recuperação de informação ao contexto da web semântica considerando a presença, cada vez mais expressiva, de acesso a textos completos.

Além disso, a ciência da informação pode e deve se beneficiar das ferramentas computacionais desenvolvidas no âmbito das pesquisas em PLN, aplicando-as nos processos clássicos de catalogação e posterior recuperação nos centros informacionais, assim como na concretização de modelos abstratos de representação de informação inerentes ao campo. Por outro lado, os resultados obtidos no escopo desta tese, deixa no ar uma questão inerente a exiguidade de processos completamente automáticos envolvendo linguagem natural, visto que cada vez mais, bases previamente analisadas, principalmente manualmente, têm sido utilizadas em etapas de aprendizado, o que torna o desenvolvimento engessado e dependente de esforço manual.

Finalmente, vale destacar que no escopo desta tese não se esgotaram as possibilidades de discussão e ainda há muito que ser explorado. A partir do material empírico focalizado, outras dimensões de análise podem ser definidas e exploradas futuramente. Como trabalho futuro, o instrumento de seleção elaborado pode ser reformulado, retirando-se a faceta computacional para investigar como a ciência da informação tem abordado a linguagem natural e a representação documentária.

REFERÊNCIAS

ALUISIO, S. M. ; PINHEIRO, Gisele Montilha ; MANFRIN, Aline P M ; OLIVEIRA, Leandro H M de ; GENOVES JR, Luiz C ; TAGNIN, Stella E O . The Lacio-Web: Corpora and Tools to Advance Brazilian Portuguese Language Investigations and Computational Linguistic Tools. IN: 4th International Conference on Language Resources and Evaluation (LREC 2004), 2004, Lisboa. Proceedings of The 4th International Conference on Language Resources and Evaluation, 2004. v. 5. p. 1779-1782.

ALVES, Isa Mara da Rosa ; CHISHMAN, R. L. O. . A Ambiguidade e a Tradução Automática: uma análise do desempenho. IN: III Colóquio Anual de Lusofonia, 2004, Bragança. Anais do II Colóquio Anual de Lusofonia, 2004. v. 1.

AZIZ, W. F. ; PARDO, T. A. S. ; PARABONI, I. . An Experiment in Spanish-Portuguese Statistical Machine Translation. In: 19th Brazilian Symposium on Artificial Intelligence (SBIA-2008), 2008, Salvador, Brasil. Springer Lecture Notes in Artificial Intelligence. v. 5249. p. 248-257.

AZIZ, W. F. ; PARDO, T. A. S. ; PARABONI, I. . Fine-tuning in Portuguese-English Statistical Machine Translation. In: 7th Brazilian Symposium in Information and Human Language Technology (STIL-2009), 2009, São Carlos, Brasil. Proceedings of STIL-2009, 2009.

BAEZA-YATES, R.; RIBEIRO-NETO, B., Modern Information Retrieval. Addison-Wesley, 1999.

BALAGE FILHO, P. P ; UZEDA, V. R. ; PARDO, Thiago Alexandre Salgueiro ; NUNES, Maria das Graças Volpe . Experiments on applying a text summarization system for question answering. In: Cross Language Evaluation Forum 2006 Workshop, 2007, Alicante. Lecture Notes in Computer Science. Berlin Heidelberg : Springer-verlag, 2006. v. 4730. p. 372-376.

BARCIA, R. M. ; HOESCHL, HUGO ; MATTOS, EDUARDO DA SILVA ; BUENO, TANIA CRISTINA D'AGOSTINI ; GRESSE VON WANGENHEIM, C. . Uso da Teoria Jurídica para Recuperação em amplas bases de textos jurídicos. In: Encontro Nacional de Inteligência Artificial, 1999, Rio de Janeiro. Anais do XIX Congresso Nacional da Sociedade Brasileira de Computação. Rio de Janeiro : Edições Entrelugar, 1999. v. 4. p. 107-120.

BARDIN, Laurence. Análise de conteúdo. Lisboa: Edições 70; 1977.

BARROS, F. A. . Semi-automatic Anaphora Resolution in Portable Natural Language Interfaces. In: XIII Brazilian Symposium on Artificial Intelligence (sbia'96), 1996, curitiba. Lecture Notes in Artificial Intelligence. berlim : springer, 1996. v. 1159. p. 121-130.

BECKER, David, Automated Language Processing, Annual Review of Information Science and Technology, Vol. 16, Pág. 113-138, 1981.

BERBER SARDINHA, TONY . Estudo baseado em Corpus da Padronização Lexical no Português Brasileiro: colocações e perfis semânticos. In: PROPOR'99. IV Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada, Evora, 1999, p. 269-287.

BERBER SARDINHA, TONY . Prosódia Semântica na Tradução do Português e Inglês: um estudo baseado em corpus. In: V PROPOR - Encontro para o Processamento Computacional da Língua Portuguesa Falada e Escrita, 2000, Atibaia, São Paulo. PROPOR 2000. São Carlos, SP : ICMC/USP, 2000. p. 93-104.

BIDARRA, J. . Notas para a Especificação de um léxico computacional, baseadas em dados de Parafasia Semântica. In: Congresso Brasileiro de Computação, II Workshop de Informática na

Saúde, 2002, Itajaí - SC. Congresso Brasileiro de Computação - II Workshop de Informática na Saúde, 2002.

BOBROW, D.G.; FRASER, J.B.; QUILLIAN, M.R., Automated Language Processing, Annual Review of Information Science and Technology, Vol. 2, Pag. 161-186, 1967.

BONFANTE, A. G. ; NUNES, M. G. V. . Parsing Probabilístico para o Português do Brasil. In: I Workshop de Teses e Dissertações em Inteligência Artificial (I WTDIA), 2002, Porto de Galinhas - Recife. I Workshop de Teses e Dissertações em Inteligência Artificial (I WTDIA), 2002.

BRAGA, I. A. ; MONARD, M. C. ; MATSUBARA, E. T. . Combining Unigrams and bigrams in Semi-supervised Text Classification. In: Portuguese Conference on Artificial Intelligence, 2009, Aveiro. Lecture Notes in Artificial Intelligence, 2009.

BUSH, Vannevar. As we may think. Atlantic Monthly, v. 176, n. 1, p. 101-108. 1945

CAMINADA, Nuno ; QUENTAL, V. S. D. B. ; GARRAO, Milena Uzeda . Linguistic Tools – Uma Plataforma Expansível de Funções de Consulta a Corpus. In: VI Workshop em Tecnologias da Informação e Linguagem Humana, 2008, Vila Velha. Anais do VI Workshop em Tecnologias da Informação e Linguagem Humana. Vila Velha : VI Workshop em Tecnologias da Informação e Linguagem Humana, 2008.

CAMPOS LEAL, I., Análise de citações da produção científica de uma comunidade: a construção de uma ferramenta e sua aplicação em um acervo de teses e dissertações do PPGCI-UFMG, Dissertação de Mestrado. Programa de Pós-graduação de Ciencia da Informação da UFMG, 2005.

CARVALHO, A. M. B. R. ; STRUBE DE LIMA, V. L. . Processamento de Língua Natural: duas experiências com sistemas multi-agentes. In: IX Intercâmbio de Pesquisas em Linguística Aplicada, 1999, São Paulo. Anais do 9o. INPLA, 1999.

CASELI, H. M. ; NUNES, M. G. V. . Automatic transfer rule induction from parallel corpora. IN: 3rd Workshop on Msc dissertations and PhD thesis in Artificial Intelligence (WTDIA'2006), Ribeirão Preto, Brazil, October 23–28, Proceedings of the International Joint Conference IBERAMIA/SBIA/SBRN 2006, 2006. p. 1-10.

CASELI, H. M. ; NUNES, M. G. V. ; FORCADA, M. L. . LIHLA: A Lexical Aligner Based on Language-Independent Heuristics. IN: V Encontro Nacional de Inteligência Artificial (ENIA 2005), 2005, São Leopoldo – RS. Proceedings of the V Encontro Nacional de Inteligência Artificial , 2005. p. 641-650.

CASELI, H. M. ; PARDO, T. A. S. ; GOMES, F. T. ; NUNES, M. G. V. . VisualLIHLA: the visual online tool for lexical alignment. In: VI Workshop em Tecnologia da Informação e da Linguagem Humana, 2008, Vila Velha - ES. Proceedings of the VI Workshop em Tecnologia da Informação e da Linguagem Humana, 2008.

CHOWDHURY, Gobinda C, Natural Language Processing, Annual Review of Information Science and Technology, Vol. 37, Pág. 51-89, 2003.

CORNELIUS, Ian. Theorizing Information for Information Science, Annual Review of Information Science and Technology – ARIST, v. 36, 2002. .

DAMERAU, Fred J., Automated Language Processing, Annual Review of Information Science and Technology, Vol. 11, Pág. 107-161, 1976.

DIAS, E. W.; NAVES, M. M. L. Análise de assunto: teoria e prática. São Paulo: Thesaurus, 2007.

DIAS-DA-SILVA, Bento Carlos ; MORAES, Helio Roberto de; OLIVEIRA, Mirna Fernanda de; HASEGAWA, Ricardo; AMORIM, Daniela Angelucci de; PACHOALINO, Christie . Construção de um Thesaurus Eletrônico para o Português do Brasil. Processamento Computacional do Português Escrito e Falado, Atibaia, v. 4, p. 1-10, 2000.

FERNEDA, E. Recuperação de Informação: análise sobre a contribuição da Ciência da Computação para a Ciência de Informação. 2003. 137 f. Tese (Doutorado em Ciência da Informação) – Escola de Comunicações e Artes, Universidade de São Paulo, São Paulo, 2003.

FUJITA, M. S. L.. A identificação de conceitos no processo de análise de assunto para indexação. Rev. Digital de Biblioteconomia e Ciência da Informação, Campinas, v. 1, n. 1, p.60-90, dez. 2003. Disponível em: <http://dici.ibict.br/archive/00000239/>. Acesso em: 15 ago. 2009.

FUSARO, P. S. ; ZIVIANI, N. . Uma linguagem de consulta para um sistema de recuperação de informação em texto completo. In: IX Congresso da Sociedade Brasileira de Computação, 1989, Uberlândia. Anais do IX Congresso da Sociedade Brasileira de Computação, 1989. p. 284-288.

GAMALLO, Pablo ; AGUSTINI, Alexandre ; LOPES, Jose Gabriel Pereira . Selection Restrictions Acquisition from Corpora. In: 10th Portuguese Conference on Artificial Intelligence (EPIA), 2001, Porto. Lecture Notes in Artificial Inteligence, LNAI. Berlin : Springer Verlag, 2001. v. 2258. p. 30-43.

GASPERIN, Caroline Varaschin ; STRUBE DE LIMA, V. L. . Evaluating Automatically Computed Word Similarity. In: PROPOR 2003, 2003, Faro - Portugal. Computational Processing of the Portuguese Language (Lecture Notes in Artificial Intelligence). Berlin : Springer-Verlag, 2003. v. 2721. p. 243-250.

GONZÁLEZ DE GÓMEZ, Maria Nélida. Metodologia de pesquisa no campo da ciência da informação. DataGramaZero - Revista de Ciência da Informação, v. 1, n. 6, dez. 2000.

GONZALEZ, M. A. I. ; STRUBE DE LIMA, V. L. . Recuperação de Informação e Expansão Automática de consulta com thesaurus: uma avaliação. In: XXVII Conferência Latinoamericana de Informática (CLEI'2001), 2001, Ciudad de Merida. CD-ROM, 2001.

GUEDES, Emanuel Guedson Ferreira. O conceito *aboutness* na Organização e Representação do Conhecimento, Dissertação de Mestrado em Ciência da Informação, Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, 90 p., 2009, disponível em < http://www.marilia.unesp.br/Home/Pos-Graduacao/CienciadaInformacao/Dissertacoes/ guedes_egf_me_mar.pdf>. Acesso em: 15 de fev. 2010.

GUINCHAT, Claire; MENOU, Michel. Introdução geral às ciências e técnicas da informação e da documentação. Brasília: MCT: CNPq: Ibict, 1994.

HAAS, Stephanie W, Natural Language Processing: Toward large-scale, robust systems, Annual Review of Information Science and Technology, Vol. 31, Pág. 83-119, 1996.

HULLEY et al., Delineando A Pesquisa Clinica 3ªed. 2006.

JOSE NETO, J. ; MENEZES, C. E. D. . Um Método para a Construção de Etiquetadores Morfológicos Aplicado a Língua Portuguesa, baseado em Autômatos Adaptativos. In: PROPOR 2000 – V Encontro para o Processamento Computacional da Língua Portuguesa, 2000, Atibaia. Anais do V Encontro para o Processamento Computacional da Língua Portuguesa. São Carlos : ICMS-USP, 2000. p. 53-64.

JOSE NETO, J. ; MORAES, M. . Formalismo Adaptativo Aplicado ao Reconhecimento de Linguagem Natural. In: Conferência Iberoamericana em Sistemas, Cibernética e Informática - CISCI

2002, Orlando - Flórida. Anais da Conferência Iberoamericana em Sistemas, Cibernética e Informática - CISCI, 2002.

JULIA, R. M. S. ; SEABRA, J. R. ; SEMEGHINI-SIQUEIRA, I. . An Intelligent Parser that Automatically Generates Semantic Rules during Syntatic and Semantic Analysis. In: IEEE International Conference on Systems, Man and Cybernetics, 1995, Vancouver. v. i. p. 806-811.

KAY, Martin; SPARCK JONES, Karen, Automated Language Processing, Annual Review of Information Science and Technology, Vol. 6, Pag. 141-166, 1971.

KINOSHITA, J. . An Example based Machine Translation System working on trigrams. Societas Linguistica Eropaea. 32nd Annual Meeting, ljubljana, 8-11 july, 1999. In: Societas Linguistica Eropaea. 32nd Annual Meeting, , 1999, ljubljana, 1999.

KINOSHITA, J. ; SALVADOR, L. N. ; MENEZES, C. E. D. ; SILVA, W. D. C. M. . COGROO - An Openoffice Grammar Checker. In: International Conference on Intelligent Systems Design and Applications, 2007, Rio de Janeiro - Brasil. ISDA'07 - Seventh Iternational Conference on Intelligent Systems Design and Applications, 2007. p. 525-530.

LANCASTER, Frederick Wilfrid. Indexação e resumos: teoria e prática. Tradução de Antonio Agenor Briquet de Lemos. Brasília: Briquet de Lemos/Livros, 2003.

LAROCCA NETO, J. ; SANTOS, A. D. ; KAESTNER, C. A. A. ; FREITAS, A. A. ; NIEVOLA, J. C. . A Trainable Algorithm for Summarizing News Stories. In: PKDD'2000 Workshop, 2000, Lyon, France. . Proc. PKDD'2000 Workshop on Machine Learning and Textual Information Access, 2000.

LE COADIC, Yves-François. A ciência da informação. Brasília: Briquet de Lemos/Livros, 1996.

LEFFA, V. J. . O uso do dicionário eletrônico na compreensão do texto em língua estrangeira. In: Anais do XI Congresso da Sociedade Brasileira de Computação, 1991. Santos, SP. p. 187-200.

MARTINS, Claudia A ; MONARD, M. C. ; MATSUBARA, E. T. . Reducing the Dimensionality of Bag-of-words Text Representation used by Learning Algorithms. In: Artificial Intelligence and Applications, 2003, Espanha. Proceedings AIA 2003. EEUU : Acta Press, 2003. v. 1. p. 49-58.

MASCARENHAS SILVA, F. Organização da informação em sistemas eletrônicos abertos de Informação Científica & Tecnológica: Análise da Plataforma Lattes., Tese de Doutorado, Escola de Comunicações e Artes da USP, 2008, Acessado em março de 2010, disponível em WWW.teses.usp.br/teses/disponiveis/27/27151/tde-17032008-095556/publico/lattes.pdf

MATSUBARA, E. T. ; MONARD, M. C. ; BATISTA, G. E. A. P. A. . Aprendizado Semi-Supervisionado Multi-Visão para a Classificação de Bases de Texto . IN: Workshop in Artificial Intelligence, 2004, Arica. Jornadas Chilenas de Computacion. Arica : Sociedad Chilena de Ciencias de la Computacion, 2004. v. 1. p. 1-9.

MIRANDA, Antonio; BARRETO, Aldo de Albuquerque. Pesquisa em ciência da informação no Brasil: síntese e perspectiva. Revista de Biblioteconomia de Brasília, Brasília, v.23/24, n.3, p. 277-292, 2000.

MONTGOMERY, Christine A., Automated Language Processing, Annual Review of Information Science and Technology, Vol. 4, Pag. 145-174, 1969.

MORAES, S. M. W. ; STRUBE DE LIMA, V. L. . Um estudo sobre Categorização Hierárquica de uma grande coleção de textos em Língua Portuguesa. In: V Workshop em Tecnologia da

Informação e da Linguagem Humana, 2007, Rio de Janeiro. Anais do Congresso da Sociedade Brasileira de Computação. Rio de Janeiro, 2007. v. 1. p. 1659-1668.

MORAIS, E. A. M. ; AMBROSIO, A. P. . Automatic Domain Classification of Jurisprudence Documents. In: EATIS 2008 - Euroamerican Conference on Telematics and Information Systems, 2008, Aracaju - SE. Anais do EATIS 2008, 2008.

MUELLER, S. P. M. e PERCEGUEIRO, C. M. P. A, O periódico Ciência da Informação na década de 90: um retrato da área refletido em seus artigos, Ciência da Informação, vol.30 no.2 Brasília May/Aug. 2001..

MUELLER, S. P. M.; CAMPELO, B. S.; DIAS, E. J. W. Disseminação da pesquisa em ciência da informação e biblioteconomia no Brasil. Ciência da Informação, Brasília, v. 25, n. 3, p. 337-352, set./dez. 1996.

MUELLER, Suzana P.M.; MIRANDA, Antonio; SUAIDEN, Emir. A pesquisa em Ciência da Informação no Brasil: análise dos trabalhos apresentados no IV ENANCIB. Revista de Biblioteconomia de Brasília, v. 23/24, p. 293-308, 2000..

OLIVEIRA, C. M. G. M. ; GARRAO, M. U. ; AMARAL, L. A. M. . Complex Prepositions Prep+N+Prep as Negative Patterns in Automatic Term Extraction from Texts. In: 7th Conference on Computational Lexicography and Text Research, 2003, Budapest. Proceedings of the 7th Conference on Computational Lexicography and Text Research, 2003.

OLIVEIRA, Itamar Leite de ; WAZLAWICK, R. S. . A Modular Connectionist Parser for Resolution of Pronominal Anaphoric References in Multiple Sentences. In: International Joint Conference on Neural Network, 1998, anchorage, alaska. IEEE World Conference on Computational Intelligence - IEEE/WCCI-98, 1998. v. 2. p. 1194-1199.

PADILHA, E. G. ; VICCARI, R. M. . Morfologia da Língua Portuguesa com Máquinas de Estados Finitos. In: 5o. Workshop de Processamento da Língua Portuguesa Falada e Escrita (PROPOR-2000), 2000, Atibaia. Anais do 5o. PROPOR - Workshop de Processamento da Língua Portuguesa Falada e Escrita, 2000.

PARDO, Thiago A S ; MARCU, Daniel ; NUNES, M. G. V. . Um Modelo Estatístico Gerativo para o Aprendizado Não Supervisionado da Estrutura Argumental dos Verbos. IN: III Workshop em Tecnologia da Informação e da Linguagem Humana, TIL 2005. , 2005, São Leopoldo. Anais do XXV Congresso Brasileiro da Sociedade Brasileira de Computação (CD-ROM). São Leopoldo : SBC, 2005. v. 1. p. 1-10.

PARDO, THIAGO A. S. ; RINO, LUCIA H MACHADO ; NUNES, M. G. V. . NeuralSumm: Uma Abordagem Conexionista para a Sumarização Automática de Textos. In: Encontro Nacional de Inteligência Artificial, 2003, Campinas. Anais do ENIA'2003, 2003. v. 1.

PARDO, Thiago Alexandre Salgueiro ; RINO, L. H. M. . DMSumm: Um Gerador Automático de sumários. In: I Workshop de Teses e Dissertações em Inteligência Artificial (WTDIA'2002), 2002, Porto de Galinhas - PE. anais do i WTDIA'2002. porto de galinhas - PE : UFPE e Sociedade Brasileira de Computação, 2002. v. 1. p. 1-10.

PILTCHER, Gustavo ; BORGES, Thyago ; LOH, S. ; LITCHNOW, Daniel ; SIMOES, Gabriel . Correção de Palavras em Chats: Avaliação de bases para Dicionários de Referência. IN: Workshop de Tecnologia da Informação e Linguagem, 2005, São Leopoldo. Anais Congresso SBC 2005, 2005. p. 2228-2237.

PINHEIRO, L. V. R. e LOUREIRO, J. M. M., Traçados e limites da Ciência da Informação Ciência da Informação, Brasília, v.24, n.1, p. 42-53, 1995.

PINHEIRO, Lena Vânia Ribeiro. Infra-estrutura da pesquisa em Ciência da Informação. DataGramaZero, Rio de Janeiro, v. 1, n. 6, dez. 2000.

RINO, L. H. M. ; PARDO, Thiago Alexandre Salgueiro ; SILLA JR, Carlos Nascimento ; KAESTNER, Celso Antonio Alves ; POMBO, M. . A Comparison of Automatic Summarization Systems for Brazilian Portuguese Texts. IN: A. L. C. BAZZAN, S. LABIDI (eds.), Advances in Artificial Inteligence. XVII Brazilian Symposium on Artificial Intelligence - SBIA'04, 2004, São Luis, Maranhão. Lecture Notes in Computer Science. Germany : Springer-Verlag, 2004. v. 3171. p. 235-244.

RINO, L. H. M. ; SENO, Eloize Rossi Marques . A importância do tratamento co-referencial para a sumarização automática de textos. Estudos Linguísticos (São Paulo), v. XXXV, p. 1179-1188, 2006.

RIPOLL, L. M. B. ; MENDES, S. B. T. . Um modelo conexionista para tratamento da ambiguidade verbal de um sub-conjunto do português. In: XV SEMISH - Seminário Integrado de Software e Hardware, 1988, Rio de Janeiro. Anais do XV SEMISH, 1988.

ROBIN, J. P. L. . Automatic Generation and Revision of Natural Language Report Summaries Providing Historical Background. In: XI Simpósio Brasileiro de Inteligência Artificial, 1994, Fortaleza, CE, Brasil. p. 0-0.

ROCHA, A. F. ; GUILHERME, I. R. ; THEOTO, M. ; MIYADAHIRA, A. M. K. ; KOIZUMI, M. S. . A neural net for extracting knowledge from natural language data bases. IEEE Transactions on Neural Networks, v. 3, n. 5, p. 819-828, 1992.

ROCHA, R. A. ; ROCHA, B. H. S. C. ; HUFF, S. M. . Automated Translation between Medical Vocabularies using a Frame-based Interlingua. In: Seventeenth Symposium on Computer Applications in Medical Care, 1993, 1993. p. 690-694.

ROSA, J. L. G. . A Thematic Connectionist Approach to Portuguese Language Processing. In: Iasted International Conference on Artificial Intelligence and Soft Computing, 1997, banff. Proceedings of the Iasted International Conference on Artificial Intelligence and Soft Computing, 1997.

ROSSI, D. ; PINHEIRO, Clarissa ; FEIER, Nara Bressane ; VIEIRA, Renata . Resolução Automática de Correferência em textos da Língua Portuguesa. REIC. Revista Eletrônica de Iniciação Científica, http: //www. sbc. org. br/reic/, v. 1, n. 2, p. 1-9, 2001.

SALLES, T. ; ROCHA, L. C. ; MOURAO, F. H. J. ; CUNHA, L. ; PAPPA, G. L. ; GONCALVES, Marcos Andre ; MEIRA JUNIOR, Wagner . Classificação Automática de Documentos Robusta Temporalmente. In: Simpósio Brasileiro de Banco de Dados, 2009, Fortaleza. Anais do XXIV Simpósio Brasileiro de Banco de Dados, 2009.

SALTON, Gerard, Automated Language Processing, Annual Review of Information Science and Technology, Vol. 3, Pag. 169-199, 1968.

SARACEVIC, T.. Information science. Journal of the American Society for Information Science, 50 (12), 1051-1063.

SARACEVIC, T. Ciência da Informação: origem, evolução, relações. Perspectivas em Ciência da Informação, Belo Horizonte, v. 1, n. 1, p. 41-62, jan/jun 1996.

SCHULZ, S. ; NOHAMA, P. ; BORSATO, E. P. ; MATIAS, L. J. D. . Indexação e Recuperação Automática de textos médicos. In: CBIS'2002 - VIII Congresso Brasileiro de Informática em Saúde, 2002, Natal. Anais do CBIS'2002 - VIII Congresso Brasileiro de Informática em Saúde, 2002. v. 1. p. 1-4.

SEMEGHINI-SIQUEIRA, Idmea. ; COSTA, A. ; COHN, P. G. . Uma Gramática Conexionista: Propriedades e Aplicações. In: III Simpósio Brasileiro de Inteligência Artificial, 1986, Rio de Janeiro. III Simpósio Brasileiro de Inteligência Artificial. Rio de Janeiro : IME-RJ, 1986. p. 113-125.

SENO, Eloize R M ; NUNES, M. G. V. . Fusão Automática de Sentenças Similares em Português. In: VII Simpósio Brasileiro em Tecnologia da Informação e da Linguagem Humana, 2009, São Carlos. STIL 2009 - Anais, 2009. p. 1-10.

SENO, Eloize R M ; NUNES, M. G. V. . Some Experiments on Clustering Similar Sentences of Texts in Portuguese. In: International Conference on Computational Processing of the Portuguese Language - PROPOR, 2008, Aveiro. Lecture Notes in Computer Science - Computational Processing of the Portuguese Language. Berlin / Heidelberg : Springer, 2008. v. 5190. p. 133-142.

SILVA, Cassiana Fagundes da ; VIEIRA, Renata ; OSORIO, Fernando Santos . Evaluating the Use of Linguistic Information in the Preprocessing Phase of Text Mining - Iberoamerican Journal of Artificial Intelligence. Inteligência Artificial, Espanha, v. 9, n. 26, p. 59-66, 2005.

SILVA, Nilza Nunes, Amostragem Probabilística, Ed. USP, 120 p., 1998.

SIMMONS, Robert F., Automated Language Processing, Annual Review of Information Science and Technology, Vol. 1, Pág. 137-169, 1966.

SOUZA, R. R., Uma proposta para metodologia para escolha automática de descritores utilizando sintagmas nominais, tese de Doutorado, Escola de Ciência da Informação, UFMG, 2005.

SPECIA, L. ; NUNES, Maria das Graças Volpe . Um modelo para a Desambiguação lexical de sentido na Tradução Automática. IN: WTDIA - Workshop de teses e dissertações em Inteligência Artificial, 2004, São Luis. Anais do II Workshop de teses e dissertações em Inteligência Artificial, 2004. p. 81-90.

SPECIA, L. ; NUNES, Maria das Gracas Volpe ; STEVENSON, Mark . Mining rules for Word Sense Disambiguation. IN: III TIL - Workshop em Tecnologia da Informação e da Linguagem Humana, 2005, São Leopoldo. Anais do III TIL - Workshop em Tecnologia da Informação e da Linguagem Humana, 2005.

SPECIA, L. ; STEVENSON, Mark ; NUNES, Maria das Gracas Volpe . Learning Expressive Models for Word Sense Disambiguation. IN: ACL-2007 - 45th Annual Meeting of the Association for Computational Linguistics, 2007, Prague. Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, 2007.

STRUBE DE LIMA, V. L. . Tratamento automatizado da língua natural: rumo a correção automática?. Letras de hoje, Porto Alegre, v. 25, n. 4, p. 41-56, 1990.

VILLAVICENCIO, ALINE . Representing a System of Lexical Types using Default Unification. In: Meeting of the European Chapter of the Association for Computational Linguistics (EACL-99), 1999, Bergen. Proceedings of European Chapter of the Association for Computational Linguistics (EACL-99). bergen, 1999.

WALKER, Donald E., Automated Language Processing, Annual Review of Information Science and Technology, Vol. 8, Pag. 69-119, 1973.

WARNER, A.J., Natural Language Processing, Annual Review of Information Science and Technology, Vol. 22, Pág. 79-108, 1987.

WITTEN; I. *et al*.. Managing Gigabytes. Morgan Kaufmann Publishers, Inc. Second Edition. 1999.

WIVES, L. K. Utilizando conceitos como descritores de textos para o processo de identificação de conglomerados (clustering) de documentos, tese de Doutorado, Universidade Federal do Rio Grande do Sul, 2004.

ZAVAGLIA, C. . Homonímia no Português: tratamento semântico segundo a estrutura Qualia de Pustejovsky com vistas a implementações computacionais. Revista Alfa, v. 42, n. 47, p. 77-99, 2003.

ZIVIANI, N. ; ALBUQUERQUE, L. C. A. . Um novo método eficiente para recuperação em textos. In: VII Congresso da Sociedade Brasileira de Computação, 1987, Salvador. Anais do VII Congresso da Sociedade Brasileira de Computação, 1987. p. 175-187.

APÊNDICE A – ARTIGOS DE REVISÃO DO ARIST ANALISADOS

Para a construção do critério de seleção automática das publicações atinentes à área de PLN, os onze artigos de revisão listados a seguir foram analisados.

Chowdhury, Gobinda C, Natural Language Processing, Annual Review of Information Science and Technology, Vol. 37, Pág. 51-89, 2003.

Haas, Stephanie W, Natural Language Processing: Toward large-scale, robust systems, Annual Review of Information Science and Technology, Vol. 31, Pág. 83-119, 1996.

Warner, A.J., Natural Language Processing, Annual Review of Information Science and Technology, Vol. 22, Pág. 79-108, 1987.

Becker, David, Automated Language Processing, Annual Review of Information Science and Technology, Vol. 16, Pág. 113-138, 1981.

Damerau, Fred J., Automated Language Processing, Annual Review of Information Science and Technology, Vol. 11, Pág. 107-161, 1976.

Walker, Donald E., Automated Language Processing, Annual Review of Information Science and Technology, Vol. 8, Pag. 69-119, 1973.

Kay, Martin; Jones, Karen Sparck,Automated Language Processing, Annual Review of Information Science and Technology, Vol. 6, Pag. 141-166, 1971.

Montgomery, Christine A., Automated Language Processing, Annual Review of Information Science and Technology, Vol. 4, Pag. 145-174, 1969.

Salton, Gerard, Automated Language Processing, Annual Review of Information Science and Technology, Vol. 3, Pag. 169-199, 1968.

Bobrow, D.G.; Fraser, J.B.; Quillian, M.R., Automated Language Processing, Annual Review of Information Science and Technology, Vol. 2, Pag. 161-186, 1967.

Simmons, Robert F., Automated Language Processing, Annual Review of Information Science and Technology, Vol. 1, Pág. 137-169, 1966.

APÊNDICE B - LISTA DOS TERMOS INDEXADORES

Os termos indexadores usados no critério de seleção automática estão listados a seguir, conforme codificados no programa desenvolvido pela doutoranda no escopo desta tese. Para cada termo, utilizou-se também as variações de número e idioma (para o inglês).

Conceitos Computacionais	Conceitos Linguísticos	Aplicações	Técnicas/Métodos
	portugues		
	linguagem natural		
	palavra		
automatico	linguistica		
computacional	lexical	traducao	
sistema	verbo	sumarizacao	
inteligencia artificial	ingles	indexacao	lexico
processamento	adjetivo	classificacao	gramatica
tecnologia	pronome	recuperacao	parser
parser	adverbio	stemming	corpus
algoritmo	ambiguidades	etiquetagem	dicionario
redes neurais	sufixo	respondedor	thesauro
implementacao	homografo	analise de conteudo	analise de radicais
arvore	sintagmas	spelling	palavra-chave
grafo	morfema	analise de estilo	
inferencia	sinonimo	analise de discurso	
	antonimo		
	hiponimia		
	meronimia		