Universidade de Brasília

Faculdade de Ciência da Informação

Programa de Pós-Graduação em Ciência da Informação - PPGCI

ERNESTO CARLOS BODÊ

Memória, Mudança Linguística *versus* Recuperação em Documentos de Arquivo no longo prazo

Brasília

2015

ERNESTO CARLOS BODÊ

Tese apresentada ao Curso de Doutorado em Ciência da Informação da Faculdade de Ciência da Informação da Universidade de Brasília, como requisito parcial para a obtenção do título de Doutor em Ciência da Informação.

Orientador: Prof. Dr. Renato Tarciso Barbosa de Sousa

Memória, Mudança Linguística *versus* Recuperação em Documentos de Arquivo no longo prazo

Bodê, Ernesto Carlos
BB666m **Memória, Mudança Linguística versus Recuperação em Documentos de Arquivo no longo prazo / Ernesto Carlos Bodê; orientador Renato Tarciso Barbosa Sousa. -- Brasília, 2015.**
212 p.

Tese (Doutorado - Doutorado em Ciência da Informação) -- Universidade de Brasília, 2015.

1. Memória. 2. Linguística. 3. Preservação. 4. Representação da Informação. 5. Recuperação da Informação. I. Sousa, Renato Tarciso Barbosa, orient.
II. Título.

FOLHA DE APROVAÇÃO

Título: "Memória, Mudança Linguística *versus* Recuperação em Documentos de Arquivo no longo prazo".

Autor (a): Ernesto Carlos Bodê

Área de concentração: Gestão da Informação

Linha de pesquisa: Organização da Informação.

Tese submetida à Comissão Examinadora designada pelo Colegiado do Programa de Pós-graduação em Ciência da Informação da Faculdade em Ciência da Informação da Universidade de Brasília como requisito parcial para obtenção do título de **Doutor** em Ciência da Informação.

Tese aprovada em: 10 de março de 2016.

Prof. Dr. Renato Tarciso Barbosa de Sousa
Presidente (UnB/PPGCINF)

Profª Drª Dulce Maria Baptista
Membro Interno (UnB/PPGCINF)

Prof. Dr. Rogério Henrique de Araujo Junior
Membro Interno (UnB/PPGCINF)

Profª. Drª. Marisa Brascher Basilio Medeiros
Membro Externo (UFSC)

Prof. Dr. Miguel Ángel Márdero Arellano
Membro Externo (Inst. Brasileiro de Informação Ciênc. e Tecnologia)

Profª Drª Fernanda Passini Moreno
Suplente (UnB/PPGCINF)

DEDICATÓRIA

Minha mãe foi uma mulher simples. Não somente simples pelas roupas discretas e pela postura pessoal despretensiosa. Quero dizer que ela foi simples pelo meio social em que nasceu e pelos poucos anos em que pôde estudar numa escola rural.

Não foi uma dessas mães que te inspira sobre o saber, os princípios morais e éticos, o exemplo do estudo. Estudo ela não possuía e sua ética era a ética de tantos outros brasileiros, mais submissos à pobreza que cientes de Kant ou Nietsche. Mais preocupados em sobreviver que nas ideias de Marx.

Eu a amo por ser uma pessoa boa, cordial, carinhosa, caridosa. E admiro por ter sido capaz de sacrifícios por outros. Eu a amo e admiro por ter sido autêntica no que um ser humano melhor pode ser.

Agora, preiteando meu quarto diploma numa importante universidade brasileira, o de doutô, não posso deixar de lembrar que, em grande parte, todo o trabalho e sacrifícios até aqui foram em grande parte inspirados no desejo de poder (re)tribuir a ela o que me deu, devolvendo um pouco ao menos de todo carinho que me deu e dos sacrifícios que fez por mim.

Um golpe do destino a tirou deste mundo no último semestre desta pesquisa.

Não poderia dedicar esse trabalho a outra pessoa.

Palmira! Te amo!

AGRADECIMENTOS

Impossível agradecer a todos que nos ajudaram. Uma pergunta de um colega na sala de aula, a provocação de um professor, o incentivo de um colega ou amigo. Até alguns personagens de filmes e literatura, todos a seu modo ajudaram a formar o que somos, a estabelecer nosso entusiasmo e fortaleza diante de tantos desafios, inclusive no trabalho de elaboração desta pesquisa.

Quero agradecer à Professora Miriam Manini pelo incentivo inicial e apoio para ingressar no programa. Quero agradecer também a meu orientador, que tem sido bastante paciente com minhas idiossincrasias.

As primeiras leituras e críticas, além de meu orientador, foram feitas pela Professora Johanna Smit (PPGCI USP) e pelo Professor José Augusto Guimarães (PPGCI UNESP), suas críticas foram fundamentais para direcionar o trabalho inicial de pesquisa e focar em pontos cuja importância não tinha percebido.

As críticas dos membros das bancas, Professora Dulce Baptista, Marisa Bräscher, Rogério de Araújo, Miguel Arellano e Fernanda Passini também foram substanciais e permitiram grandes melhorias para esta pesquisa.

Agradeço também a minha revisora de texto Mari Lúcia del Fiaco e a minha revisora de normas para apresentação de trabalhos acadêmicos Karla Chaves Gentil. Ambas contribuíram muito com a qualidade final em relação à forma deste trabalho, todos os erros que porventura persistam são de minha responsabilidade.

Enfim, agradeço também à paciência de minha família, especialmente aos mais próximos de mim, minha mulher, Ivenice, e meu filho, Yves.

Obrigado!

RESUMO

O trabalho apresenta resultados que permitam a melhor compreensão dos efeitos do fenômeno da mudança linguística na recuperação da informação que ocorrerá no futuro de longo prazo por pessoas utilizando estados posteriores da língua portuguesa em relação ao estado de língua utilizado na criação dos documentos. O escopo definido compreende documentos de arquivo históricos produzidos contemporaneamente, os quais precisarão ser recuperados através de sistemas informatizados, ao longo do tempo em que serão utilizados os novos estados da língua portuguesa. A metodologia utilizada é caracterizada como uma pesquisa com métodos estatísticos básicos num nível exploratório. Os métodos específicos que foram adotados compreendem revisão bibliográfica para apresentação de um quadro sinóptico com os conceitos mais relevantes para a compreensão do problema, incluindo suas relações mútuas e seu contexto, tudo num nível teórico. Em seguida, é executada uma pesquisa documental num acervo composto de documentos de arquivo produzidos no período imperial brasileiro. Os resultados finais obtidos são o referido quadro sinóptico e dados que permitem a caracterização da mudança linguística, em vários aspectos da língua, num acervo considerado e que indicam problemas potenciais relativos à recuperação da informação futura. Com base nos dados obtidos e associados à revisão de literatura, as conclusões do trabalho permitem compreender, com mais exatidão, os efeitos da mudança linguística na recuperação da informação de longo prazo. Após as conclusões, são apresentados elementos para a continuidade desses estudos no futuro.

Palavras-chave: memória, documento de arquivo, preservação, representação da informação, recuperação da informação, mudança linguística.

ABSTRACT

The paper presents results that allow better understanding of the effects of the phenomenon of linguistic change in information retrieval that will occur in the long-term future by people using subsequent states of the Portuguese language in relation to the state of language used in the creation of documents. The defined scope comprises of historical records produced contemporaneously, which need to be recovered through computer systems, over time using new states of the Portuguese language as well. The methodology is characterized as a survey of basic statistical methods in an exploratory level. Specific methods that have been adopted include literature review to present a synoptic table with the most relevant concepts for the understanding of the problem, including their mutual relations and context, all on a theoretical level. Then a documentary research in a collection composed of records produced in the Brazilian imperial period is performed. The final results obtained are the above synoptic table and data to allow the characterization of linguistic change in a collection considered and possibly indicating problems of information retrieval. Based on the data obtained and associated with the literature review, the conclusions of the work allow us to understand, more accurately, the effects of language change in the recovery of long-term information. After the conclusions are presented elements for the ongoing of these studies.

Keywords: memory, archive document, preservation, information representation, information retrieval, linguistic change.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE SIGLAS

TICs	Tecnologias da Informação e Comunicação
SRI	Sistema de Recuperação da Informação
ML	Mudança Linguística
CI	Ciência da Informação
RI	Recuperação da Informação
IR	Information Retrieval
PLN	Processamento da Linguagem Natural
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
NILC	Núcleo Interinstitucional de Linguística Computacional
UNESCO	Organização das Nações Unidas para a Educação, Ciência e Cultura
TVM	Teoria da Variação e Mudança
SQ	Sociolinguística Quantitativa

SUMÁRIO

1 INTRODUÇÃO

O tema de pesquisa nesta tese, como será abordado na seção correspondente, é: a relação da recuperação da informação – utilizando sistemas de recuperação da informação – para a preservação de documentos de arquivo textuais de valor permanente. Portanto, relaciona-se à questão da memória e de sua preservação, mais especificamente em relação ao patrimônio documental.

O problema da pesquisa, como será abordado na seção correspondente, é definido desta forma: como os efeitos da mudança linguística podem afetar a recuperação de documentos de arquivo textuais de guarda permanente registrados no estado atual da língua portuguesa; recuperação que será feita por pessoas que utilizarão futuros estados dessa língua mediante sistemas de recuperação da informação.

Destaca-se que se trata de uma abordagem em relação aos documentos de arquivo de guarda permanente, portanto com valor histórico cultural, produzidos atualmente. O que se pretende é explorar e compreender como os efeitos da mudança linguística – que ocorre contínua e permanentemente, conforme se verá na revisão de literatura – afetarão a recuperação das informações nesses documentos através de sistemas de recuperação (informatizados) que serão utilizados no futuro.

Por se tratar de efeitos que ocorrerão no futuro em sistemas igualmente usados no futuro, com tecnologia de aplicação desconhecida do ponto de vista atual, a proposta de exploração e compreensão dos referidos efeitos firma-se em três partes, com métodos distintos para coleta de dados. Na primeira, com base na revisão de literatura, é apresentado um quadro sinóptico com os principais conceitos e relações entre eles que permitam a compreensão do problema de pesquisa. Na segunda, esse quadro sinóptico é confrontado com a produção de ciência da informação no Brasil, a fim de identificar como esses conceitos são tratados atualmente, pelo menos nos limites de uma exploração por ocorrência de termos. Na terceira, essa base de referência é então adicionada a uma pesquisa documental, que analisa um acervo documental da época do império brasileiro. A estratégia de exame de um acervo de documentos de arquivo do passado permite visualizar, por meio dos exemplos de análise da mudança nos elementos da língua ali registrados, como se dá a mudança linguística. Por exemplo, são investigados possíveis indicadores numéricos da mudança efetiva ocorrida em função do tempo transcorrido até hoje. Tais dados permitem extrapolar os efeitos da mudança linguística para acervos documentais atuais e possíveis efeitos no futuro.

Diante da complexidade da matéria, que envolve teorias da filosofia da linguagem, linguística, ciência da computação, além da própria ciência da informação, cujo ponto de vista é o principal, e também diante da complexidade de uma língua, falada ou escrita, optou-se por desenvolver uma pesquisa exploratória acerca do problema. Apresentam-se, nas conclusões, dados e observações que permitem melhor compreender esse problema. Ainda nas conclusões, há uma hipótese de pesquisa a ser desenvolvida em estudos futuros.

Conforme visto, esta pesquisa explora teorias de várias disciplinas e nelas se baseia. Aqui estão resumidas as principais delas. Em filosofia da linguagem, explora-se o conceito de linguagem propriamente. Em linguística, teorias sobre variação linguística (sociolinguística), mudança linguística e também linguística histórica e semântica. Em ciência da computação, são exploradas as teorias sobre recuperação da informação e modelos específicos para esta finalidade. Em ciência da informação, trata-se de teorias sobre representação da informação, incluindo instrumentos específicos para seu uso e teorias sobre o documento de arquivo e descrição de seu conteúdo.

O maior interesse desta pesquisa é colaborar para a diminuição de riscos em relação à preservação da memória em patrimônio documental do tipo documentos históricos de arquivo. Ao melhor explorar a problemática aqui proposta, abrem-se novos encaminhamentos de estudos com bases mais sólidas. Futuramente, pretende-se dar continuidade a este trabalho na forma de testes e avaliações práticas dos produtos apresentados nas conclusões.

Estruturalmente, o trabalho está dividido em quatro partes. A primeira (apresentação geral) aborda em detalhes o tema de pesquisa, problema e objetivos, além desta introdução. A segunda parte (revisão de literatura) apresenta a exploração nas disciplinas e respectivas teorias que sustentam a pesquisa. A terceira parte (metodologia & métodos) compreende o detalhamento metodológico em si e os métodos e procedimentos aplicados para obter dados. Aliados à revisão de literatura e seu aporte teórico, esses dados permitem a apresentação da quarta e última parte (conclusões gerais). Após as referências utilizadas no trabalho, há anexos e apêndices disponíveis.

Como os procedimentos utilizados na terceira parte desta tese implicaram o uso de dados linguísticos analisados através de vários aplicativos de *software* diferentes, optou-se por apresentar também como anexo um disco com a gravação dos arquivos que permitem sua verificação, testes e reproduções das análises.

2 TEMA DE PESQUISA

Umberto Eco, em seu livro que trata de produção científica (*Como se faz uma tese*), elenca algumas regras para quem pretende produzir e comunicar produtos do trabalho científico. A primeira delas é que o tema escolhido para pesquisa "responda aos interesses do candidato" (ECO, 2006). Isso significa que o tema deve estar ligado "ao tipo de exame quanto às suas leituras, sua atitude política, cultural ou religiosa" (ECO, 2006, p. 6). No caso específico deste trabalho, a escolha do tema foi o resultado da trajetória acadêmica e experiências pessoais, principalmente da vida profissional, pois foi ela que permitiu uma "consciência da problemática específica relacionada com o tema abordado de determinada perspectiva" (SEVERINO, 1992, p. 11).

O tema relaciona-se com o problema da memória, ou melhor, da preservação da memória contida ou registrada em documentos de arquivo considerados de valor histórico cultural que justifique serem mantidos por longos períodos.

Do ponto de vista de alguns historiadores, o cenário ideal seria aquele no qual todas as informações registradas (com seus respectivos suportes materiais) fossem mantidas, armazenadas e até tratadas (catalogadas, descritas e/ou indexadas) indefinidamente. Provavelmente, essa posição deriva do fato incontestável de que a história é contada com base nos documentos que registram informações ligadas a fatos, decisões, dados biográficos, contábeis etc. Infelizmente, nossa civilização nunca manteve todos os seus registros documentais e muito do que foi mantido não o foi sem adulterações – intencionais ou não.

Ao longo da história da civilização, além da falta de interesse sistemático pelo armazenamento e tratamento de todos os registros, pelo menos com a intenção de mantê-los indefinidamente, esses documentos estão sujeitos a várias ameaças e até desastres, naturais ou não. Para citar um exemplo atual, destaca-se a destruição de arte e de relíquias nas guerras no Oriente Médio.

Hoje, diante do problema da perda da memória registrada, a Unesco mantém várias iniciativas de apoio a projetos de preservação. Um deles, o *BlueShield* (escudo azul), através da associação com várias outras entidades, vem promovendo esforços para proteção de muitos sítios ameaçados[1] na atualidade.

No ambiente digital, a situação não é diferente e talvez seja até pior diante da grande quantidade de informações ali disponíveis. Uma prova disso, que pode ser retirada da rede internet, é a quantidade de endereços/*links* que não remetem a páginas com conteúdo.

[1]Ver <http://www.ancbs.org/cms/en/home>.

Eles somem sem deixar vestígios. Isso ocorre, no mais das vezes, em função dos diferentes custos para mantê-los na rede. Além do custo financeiro (servidores, energia elétrica, estrutura de rede e outros recursos), há os custos com recursos humanos, por exemplo, alguém que mantenha atualizado um blog de notícias, que mantenha a administração dos servidores na rede etc.

Para o universo dos documentos de arquivo, os princípios atuais[2] de gestão de documentos[3] especificam procedimentos para verificar o que e por quanto tempo deve ser mantido por razões legais, administrativas ou históricas.

O ponto de partida é a etapa em que se decidiu manter, armazenados e tratados, documentos digitais com valor histórico por longo período ou por tempo indefinido, sejam quais tenham sido os critérios adotados para essa decisão.

Há muitos aspectos que podem ser examinados em relação à preservação da memória por longo período contida na informação registrada em documentos com valor histórico. Desde a gestão documental (no sentido dos procedimentos que afetam aqueles que podem vir a receber o *status* de documento com valor histórico e de guarda por longo período) até os procedimentos de organização de repositórios digitais para preservação (quer se trate de arquivo histórico como instituição ou no sentido de acervos documentais).

Há também uma relação direta com a área de recuperação da informação (RI). Documentos históricos são úteis à medida que suas informações possam ser acessadas por pessoas interessadas nos aspectos culturais ali registrados. As modernas tecnologias permitem acesso a mais pessoas e de maneira mais eficiente, tanto através de redes como a internet como em sistemas informatizados disponíveis em instituições de arquivo, museus ou outros centros de memória. Nesse contexto, a representação dos documentos históricos, sua busca e recuperação são elementos importantes. Sobre a relação entre documentos históricos – e de resto qualquer outro artefato que contenha herança cultural – e a recuperação da informação, a citação desses pesquisadores é esclarecedora:

> Preservar e dar acesso à herança cultural se faz através da coleta de informação sobre objetos com herança cultural (ou objetos arqueológicos ou vidas de pessoas), armazenada e organizada em sistemas de informação como catálogos de bibliotecas, instrumentos de descrição de arquivos e registros de museus. Esses sistemas não dão acesso direto aos objetos físicos por si sós. Enquanto mecanismos de busca na Internet podem dar acesso a páginas web diretamente através de hiperlinks que conduzem a páginas que contém palavras na query, os objetos em coleções com herança cultural não podem ser diretamente acessados, de maneira que sistemas de

[2]Trata-se, aqui, principalmente da avaliação e seleção documental.
[3]No Brasil e em vários outros países, há legislação específica que trata da "avaliação documental" e protege a informação e seus respectivos documentos históricos.

> informação tem que lidar com representações (textuais), frequentemente na forma de objetos de registro. Uma das atividades chave nas instituições com herança cultural é, portanto, fazer descrições detalhadas de seus objetos em catálogos de bibliotecas, instrumentos de pesquisa de arquivo e registros museológicos. Sem essas descrições, organizadas em alguma forma sistemática, um objeto é quase completamente inacessível. (KOOLEN; KAMPS; KEIJZER, 2009, p. 274).

Essa importante relação de documentos históricos com a recuperação da informação não se aplica somente aos documentos históricos do passado, também chamados de legado, mas também àqueles que são produzidos hoje e que precisarão ser recuperados pelas próximas gerações, inclusive considerando o fato de que a maior eficiência e facilidade de uso desses sistemas, que se observa atualmente, tendem a aumentar seu uso ao longo do tempo.

O tema desta pesquisa engloba essa relação, formalmente:

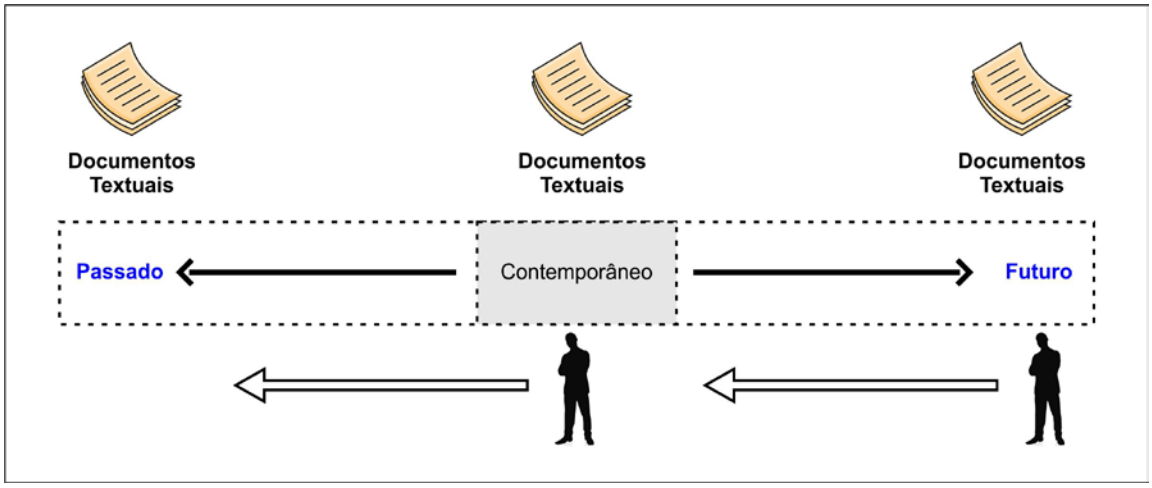
Tema de pesquisa = *A relação da recuperação da informação – utilizando sistemas de recuperação da informação – para a preservação de documentos de arquivo textuais de valor permanente*.

3 PROBLEMA DE PESQUISA

O ponto de partida para o desenvolvimento e formulação do problema de pesquisa são os processos de representação e de recuperação da informação em documentos com valor histórico-cultural os quais, em função desse valor, serão potencialmente alvo de recuperação futura para muito além do momento de sua criação, inserção ou arquivamento[1]. É importante frisar que “muito além” é mensurado em várias décadas ou mesmo séculos.

O lapso temporal entre a data de criação de um documento de arquivo textual com valor permanente e sua recuperação pode ser medido a partir de três perspectivas. Na primeira, documentos são criados no passado e recuperados contemporaneamente. Na segunda, documentos são criados e recuperados contemporaneamente. Na terceira e última, documentos são criados contemporaneamente e recuperados no futuro, pelas próximas gerações. A *figura 1* ilustra estas perspectivas.

Figura 1 – Perspectivas de produção e recuperação documentos



Fonte: Elaboração própria.

Do ponto de vista da(s) pessoa(s) que busca(m) e recupera(m) informações registradas naqueles documentos, a ação de buscar e recuperar sempre ocorre em relação às informações que foram registradas no passado, ou seja, que tenham sido previamente criadas por outras pessoas. Se o tempo transcorrido entre o momento da criação do documento e o ato de busca e recuperação for de poucos anos ou décadas, considera-se haver contemporaneidade

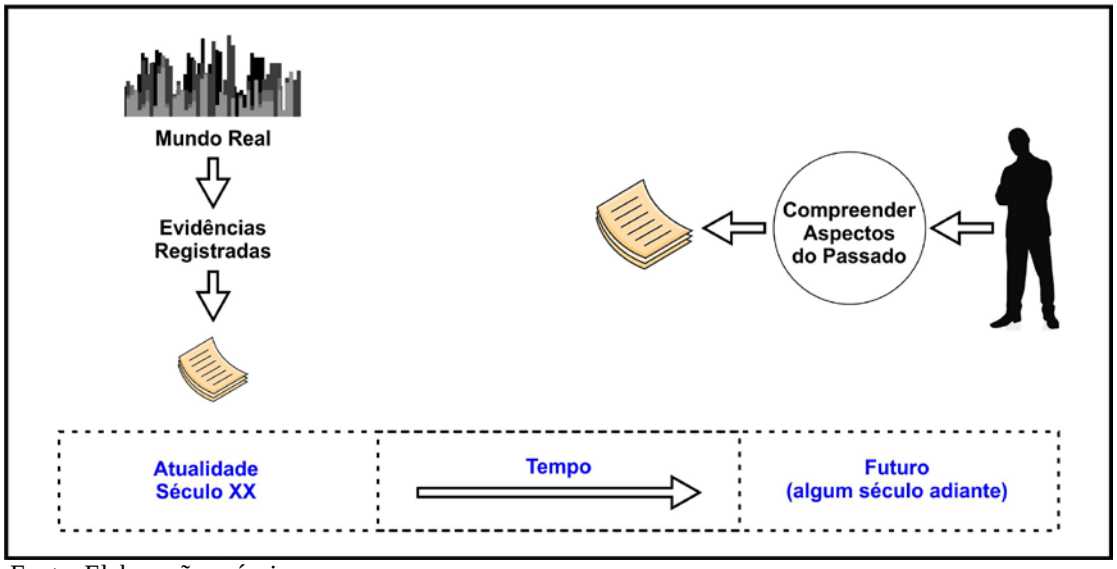
[1]A teoria arquivística prevê um ciclo de vida para os documentos, desde sua criação e associa "valores" às fases em que um documento arquivístico está sendo utilizado. Assim, o "valor primário é a qualidade inerente às razões de criação de todo documento" (CAMARGO; BELLOTTO, 1996, p. 78). Um documento com valor *histórico-cultural* é equivalente ao valor secundário, qualidade informativa que um documento pode possuir para além de seu valor primário" (CAMARGO; BELLOTTO, 1996, p. 78).

com aqueles documentos. Do contrário, fala–se de recuperação de documentos antigos. Essa antiguidade será mais bem caracterizada se o tempo transcorrido for de vários séculos. Da mesma forma, isso tudo implica aceitar que os documentos textuais históricos que são produzidos atualmente (início do século XXI) serão considerados antigos pelas pessoas que tentarem buscá-los e recuperá-los no futuro.

O valor histórico-cultural de documentos criados no passado tende a ser mais relevante proporcionalmente ao quanto esse passado está mais distante, pois esses documentos podem ser os melhores e até os únicos meios de acesso àquele legado cultural (evidências registradas no mundo real em uma determinada época), já que há uma forte tendência de perda de documentos e suas informações por vários motivos, desde o simples desgaste pelo uso até desastres naturais ou não. O mesmo raciocínio pode ser aplicado aos documentos que registram a cultura hoje e que terão seu valor aumentado proporcionalmente em relação a hoje e ao lapso temporal entre as próximas gerações, pois, da mesma forma, eles tendem a ser o melhor meio de compreensão da cultura atual pelas pessoas no futuro.

Em relação ao escopo deste problema de pesquisa, a perspectiva escolhida é a terceira, ou seja, a preocupação com o legado cultural – na forma específica de documentos de arquivo com valor histórico – e sua futura busca e recuperação pelas próximas gerações. A *figura 2* procura ilustrar especificamente esta perspectiva.

Figura 2 – Perspectiva do problema de pesquisa



Fonte: Elaboração própria.

As tecnologias da informação e comunicação (TICs) e processos técnicos utilizados hoje tanto para a representação (em biblioteconomia e arquivologia) de documentos

históricos quanto para sua recuperação através de sistemas de recuperação da informação são largamente dependentes de aspectos e componentes da língua(gem)[2]. As relações com a língua(gem) são essenciais em qualquer sistema de recuperação da informação (SRI) e até mesmo para os sistemas manuais[3], embora os últimos estejam fora do escopo desta pesquisa. A tecnologia atual para a interação pessoa-máquina (sistema) para os processos de recuperação não prescinde de comandos para a busca e recuperação (*queries*). E mesmo em casos em que sistemas permitem o reconhecimento de voz humana, isso ocorre para transformar esse som em comandos de texto a serem interpretados por um sistema informático.

Até mesmo a recuperação de documentos não textuais é também dependente da língua(gem), pois processos técnicos como descrição de documentos de arquivo, catalogação, indexação ou preparação de resumos precisam ocorrer (registrando a língua de alguma forma) antes da etapa de recuperação de documentos com imagens fixas ou em movimento e até no caso dos documentos sonoros. Os representantes (na forma da língua escrita) daqueles documentos não textuais, por exemplo, em algum padrão de *metadados*, serão utilizados tipicamente em bancos de dados a serem acessados pelos sistemas de recuperação.

A escrita possui, obviamente, uma forte relação com a língua falada. Ocorre, porém, que as línguas sofrem mudanças ao longo do tempo, mais drásticas em função do maior decorrer de tempo. A mudança linguística (ML) é um fenômeno que vem sendo estudado e discutido com base na análise de documentos escritos, pois a fixação de determinada língua numa forma específica tem permitido a comparação dessas formas ao longo dos séculos. Na verdade, desde a invenção da escrita há registros que podem ser comparados.

Um conceito linguístico importante relacionado à ML é o de estado de uma língua, que por ora pode ser simplificado como as características gerais de uma língua quando comparada com ela mesma num determinado momento do passado, de maneira que houve vários estados da língua portuguesa, por exemplo. Da mesma forma, é possível deduzir que existirão vários outros estados no futuro.

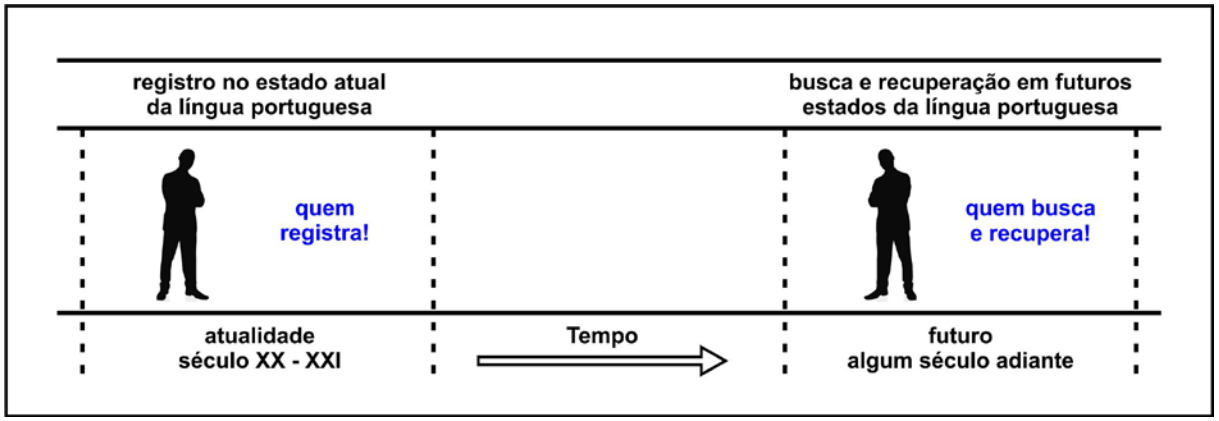
---

[2]Marcos Bagno, na tradução do livro "História concisa da Linguística" (WEEDWOOD, 2002), na Introdução, opta por grafar em português o original inglês Language na forma Língua(gem). Considero uma boa alternativa, pois o termo linguagem é utilizado também significando língua falada ou a capacidade humana de linguagem em suas variadas formas (que inclui a língua falada), além de outras acepções semióticas. Assim, adotou-se essa forma: língua(gem) e, como aquele autor, quando é utilizada, significa "capacidade humana de se comunicar por meio da fala e da escrita" (WEEDWOOD, 2002, p. 9).

[3]Cita-se como exemplo o velho sistema manual de fichas para busca de livros em bibliotecas. O texto estava lá indicando título, autor, às vezes algum tipo de resumo e a localização daquele livro.

Neste cenário, no futuro haverá pessoas buscando recuperar informação por meio do uso de um estado da língua posterior ao atual. De fato, ao longo do tempo de guarda permanente de documentos históricos, serão pessoas em vários estados futuros da língua considerada. E esses estados da língua serão diferentes do atual no qual os documentos históricos são registrados. A *figura 3* procura ilustrar este cenário.

Figura 3 – Busca e Recuperação entre diferentes estados de língua



Fonte: Elaboração própria.

Considerando que essas mudanças (entre os estados de uma língua) se refletem na respectiva escrita associada e que existe a importante presença de ambas no processo de representação da informações no processo de recuperação, é pertinente perguntar para o caso específico da língua portuguesa à qual se limita a análise:

Problema de pesquisa = Como os efeitos da mudança linguística podem afetar a recuperação de documentos de arquivo textuais de guarda permanente registrados no estado atual da língua portuguesa.

4 OBJETIVOS

Nesta seção, são apresentados os objetivos (geral e específicos) que conduzirão as atividades concretas de trabalho em relação ao problema de pesquisa apresentado.

4.1 Objetivo Geral

Derivado do problema de pesquisa e coerente com ele, o objetivo geral pode ser assim apresentado:

Descrever como os efeitos da mudança linguística podem afetar a recuperação de documentos de arquivo textuais de guarda permanente registrados no estado atual da língua portuguesa.

A formulação do objetivo geral procura enfatizar o caráter descritivo desta pesquisa. Esse objetivo relaciona-se com um problema complexo, pois conecta teorias da ciência da informação (representação, documento histórico, herança cultural), linguística (língua, escrita, mudança linguística) e ciência da computação (linguística computacional, busca e recuperação da informação, sistemas de recuperação da informação), para citar as teorias mais importantes aqui tratadas. Tal complexidade precisa ser explorada, descrita e analisada para que possa ser tratada com maior clareza e eficiência em futuros estágios de pesquisa, isso será feito inicialmente por meio da respectiva revisão de literatura no início da tese e depois através dos procedimentos práticos com documentos.

A pesquisa exploratória aqui desenvolvida limita-se a explorar a teoria linguística sobre mudança linguística, em relação a documentos históricos de guarda permanente, e os efeitos de tais mudanças em sistemas de recuperação da informação (informatizados) no futuro. Conforme os objetivos específicos a seguir, a exploração limita-se a descrever os conceitos mais relevantes ao problema e caracterizar como ocorre a mudança linguística num acervo de documentos (nos diferentes aspectos da língua).

4.2 Objetivos Específicos

Por intermédio dos objetivos específicos abaixo, é possível compreender os desdobramentos da pesquisa. Eles são apresentados em ordem lógica de complexidade de maneira a, cumulativamente, desenvolver o objetivo geral:

1. Apresentar os conceitos teóricos essenciais à compreensão das relações entre mudança linguística e recuperação da informação;
2. Caracterizar a mudança linguística em relação à atualidade num acervo de documentos produzidos no período imperial brasileiro;

O primeiro objetivo será desenvolvido a fim de apresentar um quadro sinóptico dos conceitos fundamentais para a compreensão das relações entre mudança linguística e recuperação da informação. Os procedimentos metodológicos objetivam identificar e eleger os conceitos mais importantes, apresentar sua conceituação baseada na literatura científica e estabelecer as relações entre esses conceitos. O quadro apresentado então permitirá a compreensão com base em conceitos teóricos.

Este primeiro objetivo estabelece as teorias fundamentais para as demais etapas de pesquisa e o método de trabalho será a pesquisa bibliográfica na própria revisão de literatura da tese, conforme será detalhado na seção que trata de metodologia e métodos utilizados.

O segundo objetivo específico consiste na verificação empírica dos efeitos da mudança linguística num acervo de documentos históricos de arquivo. Pretende-se melhor caracterizar o que significam efeitos da mudança linguística e assim visualizar melhor as possíveis consequências na recuperação futura nos documentos. O escopo desta pesquisa relaciona-se à perspectiva de documentos históricos produzidos na contemporaneidade e sua recuperação futura. No entanto, não é possível, por motivos práticos, verificar os efeitos da mudança linguística em documentos produzidos atualmente, pois tais efeitos só se mostrarão de maneira significativa durante os próximos séculos. Assim, torna-se necessário utilizar documentos produzidos no passado e considerar os limites desse levantamento. Em outras palavras, os dados obtidos permitem uma caracterização relativa.

O *corpus* documental utilizado neste objetivo consiste em decretos legislativos produzidos no período imperial brasileiro, a cerca de cento e oitenta e dois anos atrás. Considerando então o tempo transcorrido até hoje, será possível observar os efeitos da mudança linguística em vários aspectos da língua portuguesa.

5 JUSTIFICATIVA

Os motivos que levam à justificação de nossa pesquisa podem ser apresentados a partir da resposta a duas perguntas. A primeira pergunta é: qual a importância do problema de pesquisa para a sociedade e para a academia? A segunda é: quais as contribuições da pesquisa para a ciência da informação? Ou seja, como os resultados esperados se alinham a outras abordagens correlatas em ciência da informação.

A resposta à primeira pergunta corresponde à importância do conceito patrimônio cultural para nossa sociedade e consequentemente também para a academia como tema a ser explorado e compreendido. Pelo menos no caso brasileiro, o conceito de patrimônio cultural é abordado diretamente em nossa constituição federal, na qual o caput do Art. 216 explica que constituem o patrimônio cultural brasileiro "os bens de natureza material e imaterial, tomados individualmente ou em conjunto, portadores de referência à identidade, à ação, à memória dos diferentes grupos formadores da sociedade brasileira" (BRASIL, 1988). Cita como exemplos: "as formas de expressão", "os modos de criar, fazer e viver", "as criações científicas, artísticas e tecnológicas", "as obras, objetos, documentos, edificações e demais espaços destinados às manifestações artístico-culturais" e "os conjuntos urbanos e sítios de valor histórico, paisagístico, artístico, arqueológico, ecológico e científico" (BRASIL, 1988, art. 216). O mesmo artigo referido também estabelece a proteção desse patrimônio pelo poder público que "com a colaboração da comunidade, promoverá e protegerá o patrimônio cultural brasileiro, por meio de inventários, registros, vigilância, tombamento e desapropriação, e de outras formas de acautelamento e preservação" (BRASIL, 1988, art. 216, §1º).

Esta pesquisa trata dos efeitos da mudança linguística na recuperação futura. Tais efeitos só se tornam pertinentes o suficiente para evidenciarem possíveis problemas entre intervalos de várias décadas ou séculos, daí ser nosso objetivo de pesquisa um acervo de documentos históricos. Este tipo de acervo documental faz parte do que é chamado patrimônio documental, uma categoria de patrimônio documental. Com relação à sua importância social, a própria constituição federal procura protegê-lo "cabem à administração pública, na forma da lei, a gestão da documentação governamental e as providências para franquear sua consulta a quantos dela necessitem" (BRASIL, 1988, art. 216, §2º).

De acordo então com nossa última carta constituinte, é inegável a importância atribuída ao nosso patrimônio documental. Essa importância é também reflexo de uma visão contemporânea acerca do patrimônio cultural de uma nação e sua importância. Note-se que nossa carta tem menos de três décadas de promulgação. A partir do século XVIII com a noção

de Estado e Nação inicia-se a ideia de bens dessa nação, mas apenas no século XX esse "patrimônio começa a se constituir como fruto da memória da sociedade" (RIBEIRO; PIRES, 2015, p. 2).

A importância contemporânea que o patrimônio cultural de uma nação recebe está diretamente relacionada, entre outros fatores, à noção de identidade de pessoas e grupos sociais que pode ser reforçada e preservada através dos diferentes tipos de patrimônio cultural, inclusive os imateriais como uma língua, "a noção de identidade, tal qual o estatuto ontológico da memória, está centrada em discursos, objetos, práticas simbólicas que nos posicionam no mundo, que dizem nosso lugar em relação ao outro" (SILVEIRA, 2012, p. 3).

A categoria de patrimônio cultural que tratamos nessa pesquisa é especificamente o patrimônio cultural documental de uma nação, ou simplesmente patrimônio documental. Estamos, portanto, tratando de documentos com informações registradas que os tornam representantes de aspectos de nossa cultura. A UNESCO reconhece como patrimônio documental, entre outros, "itens textuais tais como manuscritos, livros, jornais, cartazes etc.” O conteúdo textual pode ter sido inscrito a tinta, lápis, pintura ou outro meio. O suporte pode ser de papel, plástico, papiro, pergaminho, folhas de palmeira, cortiça, pano, pedra, etc. (EDMONDSON, 2012), mas também "documentos virtuais, tais como os sítios de Internet, armazenados em servidores: o suporte pode ser um disco rígido ou uma fita e os dados eletrônicos são o conteúdo" (EDMONDSON, 2012, p. 11).

O conceito de herança cultural está muito relacionado ao de patrimônio documental e remete ao problema da memória, no sentido de que o patrimônio deve ser herdado pelas gerações futuras, preservando assim os aspectos culturais de uma nação. O ponto de vista da UNESCO é emblemático sobre isso, em um de seus programas voltados à proteção do patrimônio documental, ali preservação "é a soma das medidas necessárias para garantir a acessibilidade permanente - para sempre - do patrimônio documental." (EDMONDSON, 2012, p. 15).

Para responder à segunda pergunta que inicialmente fizemos, ou seja, "quais as contribuições da pesquisa para a ciência da informação?" é preciso identificar as relações entre patrimônio documental e herança cultural com pesquisas em ciência da informação que se alinham com aqueles conceitos. A ciência da informação é inclinada à interdisciplinaridade. Assim, estabelecer relações com áreas que tratam naturalmente da memória como a história, não é algo novo. Podem-se estabelecer relações também com a preservação de documentos e informações. O patrimônio documental é assim um elemento comum à história, área de preservação e ciência da informação.

Sobre a relação memória e informação, as mesmas "não são palavras de cunho ingênuo, elas se instituem, tem função, desdobram-se, põem-se em rede, são mutáveis, acompanham e produzem efeito na sociedade." (VERRI, 2012, online). E da mesma autora, de um ponto de vista da ciência da informação, "as informações registradas em diferentes suportes, selecionadas, agrupadas e organizadas em bibliotecas, arquivos e museus, forma os lastros do conhecimento, dos saberes estruturadores de indivíduos e de coletividades" (VERRI, 2012, online).

Com relação à preservação do patrimônio documental, a ciência da informação tem dispensado esforços tanto com relação à preservação de suportes tradicionais como também para os documentos digitais no bojo das novas tecnologias, sobre isso:

> Na área da ciência da informação, o uso da tecnologia digital que toma o lugar dos tradicionais meios de preservação, como a microfilmagem, trouxe consigo a preocupação com as normas para o uso das técnicas digitais e sua prontidão na tarefa da preservação a longo prazo. (CHEPESUIK, 1997 apud ARELLANO, 2004, p. 16).

Finalmente, podemos concluir sobre as relações entre a ciência da informação e o patrimônio documental com base nas conclusões de uma pesquisa que investigou a "inserção de pesquisas sobre patrimônio cultural no universo da Ciência da Informação" (SOUZA; CRIPPA, 2010, p. 2):

> Mesmo não se apresentando como um dos temas “dominantes” na Ciência da Informação, o que influencia na quantidade de trabalhos publicados sobre o assunto nas revistas do campo, o patrimônio cultural tem aumentando sua representatividade, principalmente, nas pesquisas de pós-graduação. A presença da temática no ENANCIB, assim como a formação de grupos de trabalho e debates nos últimos anos, apontam para a continuidade desta discussão na CI. A presença de pesquisadores formados em CI também demonstra que o campo já incorporou discussões sobre o patrimônio que geralmente eram apresentadas por profissionais de outras áreas, como Arquitetura, História, Antropologia, com uma maior tradição de pesquisa em patrimônio cultural. (SOUZA; CRIPPA, 2010, p. 17).

Garantir que documentos históricos ou em nosso caso específico o patrimônio documental constituído por documentos de arquivo possa ser recuperado de forma satisfatória pelas próximas gerações é uma das possíveis ações de preservação dos acervos. Resta verificar se a relação entre mudança linguística e a recuperação da informação tem sido objeto de pesquisa em ciência da informação e se há lacunas de pesquisa a serem preenchidas neste ramo de pesquisas.

É fato verificável nos sítios de várias instituições na rede internet[1] que as facilidades técnicas e a diminuição de custos permitiram a digitalização maciça de documentos produzidos há séculos e sua disponibilização nas respectivas versões digitais para busca e recuperação. O conteúdo textual desses documentos foi, necessariamente, registrado utilizando estados anteriores das respectivas línguas em que foram produzidos em relação ao estado atual das mesmas línguas. E logo ficou claro que os usuários atuais, utilizando o estado atual de suas próprias línguas demonstram enfrentar dificuldades para a busca e recuperação plena desses documentos.

Do ponto de vista da Ciência da Informação e da recuperação de documentos antigos, é possível identificar críticas acerca da qualidade dos resultados da correspondente recuperação baseada em algoritmos de texto integral (*full-text-indexing*) quando aborda-se o universo de digitalização de livros antigos e o problema da mudança linguística no vernáculo inglês: "em algumas instâncias, palavras e frases no texto digitalizado têm *significados diferentes* do uso de hoje" (SOBEL; BEALL, 2011, p. 4, grifo nosso). Ainda quanto ao vernáculo inglês, documentos de arquivo da guerra civil estadunidense (em grande parte composta por manuscritos) foram digitalizados para pesquisa e recuperação do público. Nesses documentos percebem-se, “idiossincrasias das fontes primárias, incluindo grafias alternativas, abreviações, uso regional ou *obsoleto de palavras*, expressões idiomáticas e omissões fazem a pesquisa por texto integral difícil, no mínimo” (BAIR; CARLSON, 2008, p. 2, grifo nosso).

Os dois exemplos no parágrafo anterior são emblemáticos em relação aos problemas como ortografia antiga ou léxico ultrapassado, alguns dos problemas identificados em relação a mudanças numa língua. Vários problemas relacionados ao registro antigo de uma língua para uso atual podem ser contornados ou pelo menos remediados por meio de estudos linguísticos ou com o uso de linguagens documentárias. Mas isso só pode ser feito a um alto custo, principalmente considerando a enorme quantidade de documentos digitalizados e disponibilizados na atualidade. A busca e a recuperação por meio de sistemas informatizados baseados em algoritmos automáticos para texto integral têm sido apontadas como uma solução ainda não adequada neste cenário (GARRETT, 2006).

Os sistemas de recuperação da informação atuais já lidam há muito tempo com os efeitos do que é chamado problema do vocabulário:

[1]Como exemplo, são utilizados documentos disponíveis no Arquivo e Biblioteca da Câmara dos Deputados em Brasília.

> Muitas funções da maioria de sistemas de grande porte dependem de usuários digitando as palavras corretas. Usuários novos ou intermitentes frequentemente usam as palavras incorretas e falham em conseguir as ações ou informações que precisam. (FURNAS et al., 1987, p. 964).

Em sua quase totalidade, as tecnologias utilizadas nos sistemas de RI consideram a necessidade de informação de um usuário como estática e, por meio de *feedback*, esse usuário deve corrigir sua formulação (GRETE, 2014). No entanto, não apenas as necessidades dos usuários não são estáticas, há um processo contínuo de busca e aclaramento de dúvidas, pois a língua – necessariamente utilizada para formular as necessidades de informação – também não é estática ou homogênea ao longo do tempo.

É possível identificar nos últimos anos novas abordagens tecnológicas que tentam identificar e contornar essas características tão próprias das pessoas e de suas línguas: "Nós então propomos um framework para explorar de várias perspectivas a mudança lexical, isto é, alterações no significado de palavras no tempo" (JATOWT; DUH, 2014, p. 2). Um relatório recente procura explorar o problema da mudança da língua frente aos sistemas de busca (MORSY; KARYPIS, 2015).

Em função da idade dessas pesquisas (dois últimos anos) e outras que têm surgido, ainda não está claro se essas propostas recentes para automatizar a identificação e o tratamento dos efeitos da mudança linguística em sistemas de recuperação da informação terão sucesso. Cabe destacar que o sucesso da indexação automática até hoje não é uma realidade plena, pois depende de textos com características bem definidas e padronizadas (LIMA; BOCCATO, 2009).

Diante das evidências de problemas neste cenário, procuramos obter, preliminarmente, dados que indiquem como o problema dos efeitos da mudança linguística tem sido abordado do ponto de vista da ciência da informação e assim obter dados que justifiquem sua exploração mais aprofundada.

Para este fim, definimos uma amostra com teses de doutorado em departamentos de ciência da informação, no Brasil. Para maior homogeneidade, consideramos apenas cursos autorizados pela CAPES com nota mínima quatro. O recorte temporal foi entre 2000 e 2015, na prática foram localizadas teses entre 2002 e 2015, oitenta e duas no total. Foram selecionadas teses que tivessem em seus títulos ou resumos pelo menos uma das seguintes palavras chave: representação da informação, recuperação da informação, memória ou documento arquivístico. A tabela 1 resume estes dados. O anexo A contém os títulos e anos das teses selecionadas, ordenadas pelo nome do arquivo e departamentos de pós-graduação em ciência da informação.

Tabela 1 – Teses em departamentos PPGCI's com nota 4

Universidade	Nota	#Teses
UFMG	4	23
UFRJ-IBICT[2]	4	02
Unb	4	13
UNESP	4	26
USP	4	18

Fonte: Capes (20/3/2015).

Efetivamos então um levantamento da **ocorrência de termos** e a correspondente análise gráfica dos resultados. Para o levantamento de ocorrência de termos, foram eleitos quatro grupos de conceitos: História, Linguística, Computação e Ciência da Informação. Com base nos objetivos desta pesquisa e sondagens iniciais, foram especificados os conceitos mais prováveis que deveriam ser tratados e inter-relacionados em pesquisas. Os grupos e conceitos escolhidos estão no *quadro1*.

Quadro 1 – Termos pesquisados por ocorrências, agrupados

Grupos	Termos correspondentes
História	memória, preservação, documento de arquivo
Linguística	linguagem, linguística, linguística histórica, mudança linguística, variação linguística, sociolinguística, diacronia, língua escrita
Computação	Sistema de recuperação da informação, processamento de linguagem natural
Ciência da informação	Arquivologia, representação da informação, recuperação da informação, tesauro, ontologia, vocabulário controlado

Fonte: Elaboração própria.

O termo "documento de arquivo" foi inserido no grupo História, pois havia interesse em documentos de arquivos históricos. Em ciência da informação e arquivologia, esse termo tanto pode se referir a "documento histórico" como também a documentos de arquivo sem valor histórico. Poder-se-ia empregar somente o termo "documento histórico", mas o mais usual é "documento de arquivo". O termo "ontologia" poderia estar no grupo Computação, mas havia interesse em pesquisas do ponto de vista da ciência da informação que tratassem desse assunto, daí porque ele foi inserido no grupo Ciência da Informação.

A partir desse agrupamento de termos, efetuou-se uma busca por ocorrência dos termos em cada uma das teses (texto completo, não apenas resumos). Foi utilizado para esta

[2]No caso da UFRJ-IBICT a maior parte das teses não está disponível em meio digital, daí o número reduzido de itens na amostra.

finalidade um *software* para busca de termos em documentos[3]. Foram consideradas variações como plural, termos compostos e sinônimos: documento de arquivo OR documento histórico; linguística histórica OR linguística diacrônica; diacronia OR diacrônico; arquivologia OR arquivística. Os arquivos correspondentes a estes dados e as planilhas utilizadas estão disponíveis no apêndice B em disco.

O resultado da pesquisa por ocorrência de termos foi uma tabela na qual os números de ocorrências dos termos foram agrupados em cores. Em função da quantidade de informações na tabela, optou-se por colocá-la no *apêndice A*. Os menores valores estão em um tom de vermelho, passando por amarelo e verde, conforme a escala aumenta. À direita na tabela do apêndice A foram identificadas teses nas quais ocorrem simultaneamente termos em quatro grupos, (1) Memória + Linguística, (2) conceitos do grupo 1 e Sistemas de recuperação da informação + Recuperação da informação, (3) todos os conceitos anteriores + Mudança linguística e (4) conceitos nos grupos (1) e (2) mais variação linguística. A legenda com o número 1 e fundo verde indica as teses nas quais ocorrem simultaneamente estes termos. Como não ocorreu o termo mudança linguística, nenhuma tese está no grupo (3) e apenas uma no grupo (4).

Nessa apresentação gráfica, é possível visualizar algumas informações preliminares importantes sobre o universo pesquisado.

Primeiro, é possível identificar uma baixa ocorrência dos termos do grupo Linguística nessa amostra de teses[4]. Destaca-se também a não ocorrência da expressão "mudança linguística" e a quase não ocorrência de "variação linguística". Mesmo considerando que eles podem estar presentes nas teses com sinônimos que não tenham sido considerados pelo *software* de busca, fica evidente a pouca atenção que eles têm nesse universo considerado. Estes termos são utilizados atualmente nessa forma, pelo menos na área de Linguística, conforme será mais bem tratado na revisão de literatura. Os termos “linguística histórica”, “sociolinguística”, “diacronia” e “língua escrita” também ocorrem de maneira significativamente pequena. Por outro lado, os termos linguagem e linguística ocorrem de maneira significativa, conforme pode ser observado na tabela do apêndice A.

Esse cenário sugere que linguística e linguagem são conceitos relevantes no conjunto de teses considerado, mas os dados sugerem que não se estabelece relação com outros conceitos linguísticos que aparecem com baixa ocorrência.

---

[3]Utilizou-se o aplicativo Agent Ransack (versão 2014), disponível gratuitamente para uso não comercial em: <http://mythicsoft.com/agentransack>.

[4]Acompanhar as afirmações seguintes conjuntamente com o apêndice A.

Segundo, observou-se que o grupo História é relativamente bem representado em relação às ocorrências de termos. Isso sugere que o assunto memória, preservação e documento de arquivo (esse bem menos representado) despertam interesse no conjunto de teses considerado.

Terceiro, as maiores ocorrências de termos estão no grupo Ciência da Informação, o que é perfeitamente coerente com o fato de que todas as teses consideradas nessa sondagem são de departamentos de Ciência da Informação.

A *tabela 2* elenca o total de ocorrências de termos em ordem relativa, bem como a média, valor mínimo e máximo de ocorrências, se comparadas entre as teses pesquisadas. O termo "linguagem" é aquele que mais aparece no cômputo geral de termos, ele foi computado 8.280 vezes. Os que menos aparecem são "Linguística Histórica" e "mudança linguística".

Tabela 2 – Termos em ordem numérica de ocorrência

Termo	Total	Média	Mín.	Máx.
Linguagem	8280	101	0	788
Ontologia	5650	69	0	1546
Tesauro	2373	29	0	805
Arquivologia	2011	25	0	696
Memória	1648	20	0	311
Recuperação da informação	1057	13	0	204
Linguística	988	12	0	138
Preservação	845	10	0	502
Proc. linguagem natural	524	6	0	176
Representação da informação	261	3	0	24
Vocabulário controlado	254	3	0	58
Sistema de rec. da informação	168	2	0	38
Documento de arquivo	53	1	0	15
Diacronia	30	0	0	15
Sociolinguística	11	0	0	3
Variação linguística	4	0	0	2
Língua escrita	4	0	0	1
Linguística Histórica	0	0	0	0
Mudança linguística	0	0	0	0

Fonte: Elaboração própria.

Os dados obtidos fornecem subsídios para acreditar que há relativa pouca atenção para o conceito mudança linguística. Outra conclusão é o pouco tratamento concomitante encontrado, ou seja, relações simultâneas estabelecidas entre os conceitos de memória, linguística (em algum aspecto), mudança linguística, representação da informação,

recuperação da informação e sistemas de recuperação da informação. A abordagem concomitante para todos esses aspectos é necessária para indicar uma postura adequada em relação ao problema dos efeitos da mudança linguística na recuperação de informações no longo prazo.

Com base então na importância do patrimônio documental para a sociedade e da relativa pouca atenção que tem sido dada, ainda que apenas com base em dados preliminares, às relações entre mudança linguística e recuperação da informação no quadro de pesquisa em ciência da informação no Brasil (no recorte considerado), o que também indica a necessidade de maior aprofundamento nesta direção; acreditamos que apresentamos dados suficientes para a justificação desta pesquisa.

.

6 REVISÃO DE LITERATURA

As seções seguintes apresentam os resultados das investigações sobre os conceitos fundamentais tratados nesta pesquisa, bem como suas relações, que permitiram a melhor compreensão do problema. Tais resultados também estabeleceram os conceitos utilizados na pesquisa documental efetuada.

A revisão de literatura compreendeu a pesquisa e análise de:

- Linguagem, de um ponto de vista da filosofia da linguagem, semiótica e linguística;
- Mudança linguística;
- A língua escrita;
- Representação e Recuperação da informação;

Ao final, apresentamos um quadro sinóptico com todos os principais conceitos analisados.

7 LÍNGUA(GEM)

Todo processo de pesquisa, raciocínio ou comunicação não pode prescindir do uso de algum tipo de língua(gem). De fato, é possível construir uma classificação hierárquica do conceito de língua(gem), considerando campos de estudo e outros (sub)conceitos (inter)relacionados. Nesta seção, o problema consiste em explorar o conceito de linguagem à luz da Filosofia da Linguagem, Semiótica e Linguística.

Um caminho possível para iniciar a discussão científica acerca de um conceito é apresentar definições prévias de outros autores. No entanto, nos limites deste trabalho, isso traria mais problemas do que ajuda. A língua(gem) é um conceito que instiga e instigou uma parcela importante de filósofos e cientistas e serviu para originar visões díspares quando não conflitantes. É por essa razão que foi empregado outro método para dar cabo desse problema. Um método que não passe somente (ou essencialmente) pelo (re)exame de conceitos anteriores. O caminho alternativo foi apresentar uma análise do conceito de língua(gem) a partir de áreas, de maneira ampla, e não das ideias de autores específicos nessas áreas.

Escolheram-se as áreas que mais se destacam em relação ao estudo do problema da língua(gem). Trata-se da *Filosofia da Linguagem*, *Semiótica* e *Linguística*. Há vantagem na escolha dessas áreas, pois as três tiveram grande evolução durante o século XX, ou seja, há pouco tempo, o que permite considerar os últimos avanços. Essas áreas estão muito interligadas, inclusive alguns expoentes atuam em mais de uma área ou nos limites das áreas. Finalmente, a problemática da língua(gem) é o objeto principal nos três campos de conhecimento.

A seguir, apresenta-se um resumo histórico de cada um, destacando os elementos importantes.

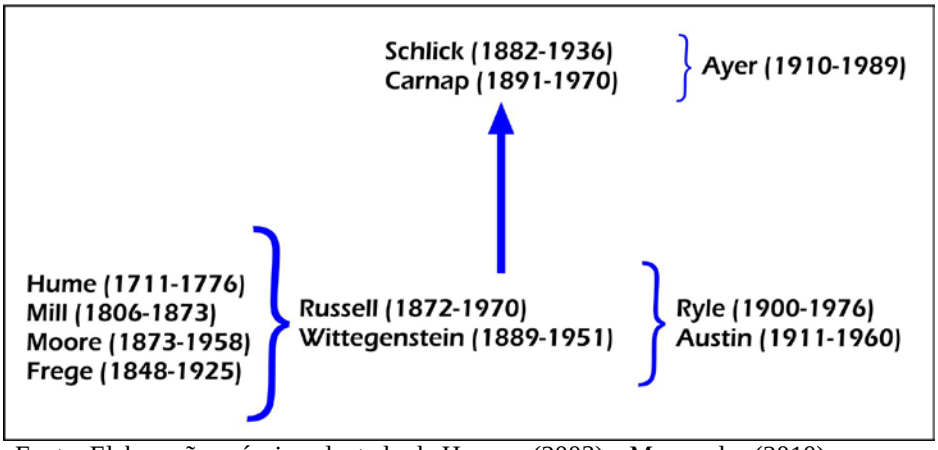
7.1 Filosofia da Linguagem

Se se considerar que a linguagem humana, em suas manifestações específicas (línguas), é imprescindível para o desenvolvimento do conhecimento humano, então uma filosofia da linguagem pode, pelo menos em alguns aspectos, ser importante para qualquer outro ramo da Filosofia geral, pois, segundo Miller, “os filósofos têm sido motivados por um desejo de dizer algo sistemático sobre essas noções de compreensão linguística, significado e conhecimento” (MILLER, 2010, p. 5). Isso vale, potencialmente para qualquer outro ramo da ciência.

Não é trivial definir o que é *filosofia da linguagem*, pois pelo menos algum aspecto da língua(gem) (humana ou não) é objeto de vários ramos da Filosofia ou está de algum modo relacionado a algum deles. Além disso, podem-se identificar problemas filosóficos que envolvem a língua(gem) desde os gregos, pelo menos. De modo que é preciso delimitar o que se irá tratar nesta seção sobre *Filosofia da Linguagem*.

A *figura 4* apresenta os principais atores sobre Filosofia da Linguagem no séc. XIX e XX. Essa figura sugere o papel importante e precursor de Russell no desenvolvimento da Filosofia da Linguagem. Entre seus trabalhos mais relevantes, estão as obras *Princípios da Matemática (1903)*, A *filosofia do atomismo lógico (1918)* e *Investigação de significado e verdade (1940)*. Russell sofreu várias influências – na figura, relacionamos Moore e Frege, mas antes desses, Stuart Mill e Hume.

Figura 4 – Correntes filosóficas



Fonte: Elaboração própria, adaptado de Hunnex (2003) e Marcondes (2010).

A *figura 4* também procura evidenciar o importantíssimo papel de Ludwig Wittgenstein no cenário do desenvolvimento da Filosofia da Linguagem. Seu trabalho fundamental conhecido como o primeiro Wittgenstein foi o *Tractatus Logico-Philosophicus* (1921). O que ficou conhecido como Círculo de Viena, por exemplo, foi explicitamente influenciado pelos seus trabalhos. Os trabalhos tardios de Russell e os de Wittgenstein dão grande importância à linguagem ordinária e sua possível redução a formas precisas para comunicação em filosofia e ciência.

A corrente dos *positivistas lógicos* ou *empirismo lógico* levou adiante essas ideias ao distinguir os critérios de verificabilidade e falsidade para as proposições apresentadas na linguagem. Schlick e Carnap são dois importantes representantes nessa linha de pensamento. Do último autor, são obras importantes A *sintaxe lógica da linguagem* (1937), *Introdução à Semântica* (1942) e *Significado e necessidade* (1947).

Mais tarde, no segundo Wittgenstein (*Investigações Filosóficas*, 1953), algumas ideias relevantes que serviram de base tanto para o atomismo lógico quanto para o positivismo lógico são rejeitadas. Wittgenstein propõe a análise da linguagem como apresentada em sua forma e contexto originais:

> Assim, ao aceitar a linguagem essencialmente como ela é e ao buscar somente aclarar e esclarecer o que estava diante dele, ele procurou – como Moore – levar a linguagem do seu uso filosófico e frequentemente problemático de volta ao seu uso natural e ordinário. (HUNNEX, 2003, p. 130).

O Positivismo Lógico, ou Empirismo Lógico, foi uma corrente filosófica[1] que teve uma preocupação especial com enunciados ou proposições, parte essencial da linguagem utilizada em ciência e/ou filosofia. Um dos pontos mais característicos daquele grupo de filósofos é sua postura em relação aos enunciados metafísicos. Eles sustentam "a tese de que todos os enunciados metafísicos são desprovidos de sentido, porque não verificáveis empiricamente" (ABBAGNANO, 2007, p. 381).

O mesmo autor também sustenta que há, entre os representantes do Empirismo Lógico, duas teses em comum com as ideias do primeiro Wittgenstein. Primeiro, "os enunciados factuais, isto é, que se referem a coisas existentes, só têm significado se forem empiricamente verificáveis" (ABBAGNANO, 2007, p. 382). Segundo, "existem enunciados não verificáveis, mas verdadeiros com base nos próprios termos que os compõem (matemática e lógica)" (ABBAGNANO, 2007, p. 382). Esses elementos tiveram influência direta na forma como se faz ciência hoje, a possibilidade de verificar e testar.

O segundo[2] Wittgenstein teve influência decisiva na chamada Filosofia da Linguagem Comum[3], bem antes de sua segunda obra-prima, "Investigações Filosóficas", obra publicada em 1953. Em 1929, ele argumenta:

> Eu costumava acreditar que havia a linguagem ordinária que todos normalmente falamos e uma linguagem primária que expressava o que realmente conhecíamos, a saber, os fenômenos. Eu também falava de um primeiro e um segundo sistema. Agora, desejo explicar por que eu não mais me ligo a essa concepção. (WCV:45 apud CHILD, 2013, p. 92).

[1]Essa corrente teve início no que é conhecido historicamente como Círculo de Viena, "um grupo de filósofos analíticos com inclinação científica e matemática que se encontraram nos seminários privados de Moritz Schlick, nas manhãs de sábado, em Viena a partir de 1923" (MAUTNER, 2011, p. 147).

[2]Apesar da predominância da visão de dois Wittgenstein "Em pesquisas acadêmicas mais recentes, essa divisão tem sido questionada: alguns analistas têm visto unidade de pensamento, enquanto outros veem nuances mais discretas, adicionando estágios como o médio Wittgenstein e o terceiro Wittgenstein" (BILETZKI; MATAR, 2014, p. 1).

[3]A Filosofia da Linguagem Comum é uma linha filosófica que "coloca a linguagem comum como objeto" e "julga que sua análise é o instrumento mais apropriado para resolver importantes problemas filosóficos" (ABBAGNANO, 2007, p. 719).

Do que se viu sobre Filosofia da Linguagem nos parágrafos anteriores, alguns traços definidores podem ser derivados.

Primeiro, nota-se uma evolução na importância da linguagem comum em contrapartida com as linguagens artificiais. Explicando melhor, ainda em Russell[4] e alguns antecessores imediatos, a linguagem da lógica era a solução para transformar os enunciados importantes em uma linguagem artificial (proposições) que se pretendia perfeita para expressar ciência e filosofia por um lado, mas que, por outro lado, apresentou seus limites.

Tomando como referência o Círculo de Viena, houve uma evolução nas teses do empirismo no sentido de avaliar enunciados ainda na linguagem comum. Finalmente, nos meados do século XX, a linguagem comum ou ordinária voltou a ocupar o foco das atenções.

No entanto, como se verá nas abordagens de Semiótica, mas principalmente na Linguística, "linguagem comum" não significa uma coisa simples ou menos complexa do que a própria linguagem lógica.

7.2 Semiótica

Várias áreas e linhas de pensamento na ciência tentam explicar o mundo ou pelo menos partes dele. Assim é que a química procura explicar certos efeitos nos materiais, na água e nos seres vivos. A física procura dar conta do porquê de haver movimento, como a luz se propaga ou como a energia elétrica pode ser produzida. Na filosofia, detectam-se as mesmas metas explanatórias gerais, por exemplo, na *cosmologia (como o universo foi criado),* na *filosofia da linguagem (como funciona a linguagem)* ou na *teoria do conhecimento (o que é conhecimento).*

A Semiótica também tem a pretensão de explicar aspectos particulares do mundo e, para alguns, até o mundo como um todo[5]. É difícil situar a Semiótica de maneira definitiva como parte das ciências ou da filosofia, pois não é – esta conclusão é quase unânime entre os autores – totalmente homogênea em suas linhas de estudos[6]. Isso talvez como reflexo de sua curta existência[7].

---

[4]É famosa na biografia de Russell uma passagem na qual ele fica encantado com a proposta de Giuseppe Peano num congresso de Filosofia em 1900 sobre os primórdios do que é hoje a linguagem da lógica.

[5]Para Charles Sanders Peirce, a Semiótica possui um potencial imenso. Para ele, "trata-se (...) de um quadro de referência que engloba qualquer outro estudo" (DUCROT; TODOROV, 2010, p. 89).

[6]De acordo com Tzvetan Todorov, "apesar da existência desses trabalhos (após citar vários autores da América e da Europa) e de quase um século de história (e vinte séculos de pré-história), a Semiótica continua sendo mais um projeto do que uma ciência constituída" (DUCROT; TODOROV, 2010, p. 93).

[7]A Semiótica atual tem dois pontos distintos, mas muito bem definidos entre o final do século XIX e início do século XX com as ideias de Peirce e Saussure (cf. mais adiante). Portanto, como área de estudo, existe há pouco

Certamente, o objetivo da *Semiótica Geral* é algo que se estrutura num campo de estudos que envolve (ou está nos limites de) conceitos como *informação*, *comunicação*, *sinal*, *código*, *ícone*, *símbolo*, *significação* e *representação do conhecimento*. Esses são os termos em evidência em trabalhos publicados na área[8].

Os primeiros autores, hoje clássicos, que trataram das bases da Semiótica e contribuíram para o que ela é essencialmente hoje foram Saussure (1857-1913) e Peirce (1839-1914). Ambos propõem teorias sobre o signo, apesar das diferenças de pontos de vista de cada um. O primeiro tinha como área de trabalho a *Linguística* e o último estudava a *Lógica*[9].

Assim, um caminho para tratar da Semiótica é discutir o que é considerado um signo nessa área e saber sua relação com outros conceitos próximos, no mesmo campo de estudos.

Partindo do princípio de que as propostas e definições para signo estão ligados a vários fenômenos e situações do homem em sua relação com o mundo nos processos de troca e obtenção de conhecimento, toma-se esse cenário como ponto de partida para compreender melhor o que é um signo para a semiótica.

Um esquema muito simplificado do mundo – e útil aos nossos propósitos – exige a presença de pelo menos dois indivíduos para concretizar a comunicação das duas mentes por meio de algum meio e tipo de língua(gem). O que pode ser comunicado pode ser classificado de infinitas formas.

Imagine-se, para efeito do esquema aqui proposto, uma separação entre coisas naturais e artificiais (produzidas pelo homem). A partir das coisas naturais, há as orgânicas e inorgânicas. Para os orgânicos, existem os racionais (homem) e irracionais (demais criaturas vivas). A partir do ponto das coisas artificiais, pode-se subdividir em coisas para registrar o conhecimento e demais artefatos com outras funções primárias (ferramentas de pesca, marcenaria, abrigos, casas, veículos e assim por diante).

Nesse esquema simplificado do mundo, o homem, ao mesmo tempo em que está inserido numa posição taxonômica bem definida, é também o criador desse esquema (no caso do ponto de vista aqui esposado) e também um ser que se correlaciona com as demais coisas.

---

mais de um século.

[8]Umberto Eco, em seu Tratado Geral de Semiótica (ECO, 2012), na introdução da obra, procura definir o que acredita serem os limites e fins de uma teoria semiótica. Sua visão é interessante, pois, tendo publicado a obra na década de 1970, permite uma visão histórica importante em relação às primeiras ideias de Peirce e Saussure.

[9]De fato, Peirce às vezes sugere que sua Lógica e o que ele chama de Semiótica, para ele, são a mesma coisa: "Em seu sentido geral, a lógica é, como acredito ter mostrado, apenas um outro nome para semiótica" (PEIRCE, 2012, p. 45).

Primeiro com outros seres semelhantes, outros indivíduos, mas também com todas as demais coisas e processos diretamente relacionados a essas coisas: criação de artefatos artificiais, identificação e classificação de coisas naturais, ritos religiosos, processos de produção do conhecimento e tantos outros processos provavelmente impossíveis de ser descritos em sua totalidade.

Esse esquema representativo do mundo, apesar de simplificar informações e categorias, abrange boa parte do mundo e condiz com as primeiras ideias sobre Semiótica, na visão peirciana, ou seja, a Semiótica é algo que poderia abarcar a totalidade do que o homem sabe ou pode saber.

> Segundo Peirce, não só as palavras da linguagem e os signos não verbais produzidos intencionalmente pelos seres humanos para comunicar-se, mas qualquer evento, estado, objeto do mundo externo e qualquer evento ou estado mental (representação, emoção, sensação etc.) podem entrar numa relação semiósica desde que interpretados por algum interpretante como signo de qualquer outra coisa. (ABBAGNANO, 2007, p. 1034).

Em seu Tratado Geral de Semiótica, Umberto Eco declara que o objetivo de seu livro "é explorar as possibilidades e as funções sociais de um estudo unificado de todo e qualquer fenômeno de significação e/ou comunicação" (ECO, 2012, p. 1). Donde se conclui que esse esquema de mundo interessa enquanto existam "fenômenos de significação e/ou comunicação" (ECO, 2012, p. 1). Segundo esse autor, o que permite analisar esses fenômenos é a ideia de signo.

Uma das definições mais estudadas de signo é a proposta por Peirce: "é tudo aquilo que está relacionado com uma segunda coisa" (PEIRCE, 2012, p. 28). É conhecida como triádica, pois envolve necessariamente três elementos em sua composição: (1) um objeto (que, no esquema aqui proposto, pode ser qualquer coisa, inclusive os próprios indivíduos), (2) um interpretante (que não é necessariamente uma pessoa[10]) e (3) signo (algo que só existe na relação (1) com (2)).

Peirce desenvolve, então, em torno dessa proposta de signo toda uma complexa teoria; deriva daí uma classificação de signos entre ícones, índices e símbolos.

Se, para Peirce, a Semiótica é uma disciplina através da qual o mundo poderia ser tornado inteligível, ele não explica como exatamente seria esse entendimento; para muitos dos

[10]Muitas das ideias de Peirce em relação ao signo são polêmicas e às vezes obscuras. Isso talvez seja o resultado da falta de uma obra única e definitiva desse autor sobre Semiótica. Apesar da tradução brasileira de sua obra (PEIRCE, 2012) ter o título Semiótica em português, o título original é "The collected papers of Charles Sanders Peirce", um apanhado de vários de seus escritos, incluindo correspondência e resenhas.

demais teóricos da Semiótica, ela é algo aplicável somente nos processos de comunicação e língua(gem). De fato, há pesquisas até mesmo sobre a comunicação e língua(gem) de seres não racionais (abelhas, baleias e golfinhos são os mais citados).

É através dos processos de comunicação e uso da língua(gem) diretamente entre indivíduos racionais, ou até mesmo num mesmo indivíduo consigo mesmo, utilizando ou não alguma forma de registro do conhecimento nesses processos, que o signo semiótico aparece como um conceito útil para apoiar as explicações de fenômenos observados.

Saussure, há quase um século, entendia a Semiótica como uma disciplina de suporte à Linguística:

> Pode-se, então, conceber uma ciência que estude a vida dos signos no seio da vida social; ela constituiria uma parte da Psicologia Social e, por conseguinte, da Psicologia Geral; chamá-la-emos de Semiologia (do grego semeion, "signo"). Ela nos ensinará em que consistem os signos, que leis os regem. Como tal ciência não existe ainda [*escrito começo séc. XX*], não se pode dizer o que será. (SAUSSURE, 2012, p. 46, nossa nota).

Saussure não tentou, no entanto, desenvolver estudos semióticos, apenas procurou posicionar a Linguística num contexto maior em relação ao que entendia por semiótica. O conceito de Saussure, que é praticamente contemporâneo a Peirce (como confirmam as datas de seus trabalhos publicados), apesar de independente dele, aparece aplicado à língua humana concreta e fonética. É relevante também ressaltar que o signo de Saussure não é triádico, mas diádico.

Para ele, o signo linguístico:

> não une uma coisa e uma palavra, mas um conceito e uma imagem acústica. Esta não é o som material, coisa puramente física, mas a impressão (empreinte) psíquica desse som, a representação que dele nos dá o testemunho de nossos sentidos; tal imagem é sensorial e, se chegamos a chamá-la "material", é somente nesse sentido, e por oposição ao outro termo da associação, o conceito, geralmente mais abstrato. (SAUSSURE, 2012, p. 106).

Se o esquema apresentado tem o objetivo de associar a ideia de signo no mundo como um todo, outro esquema representaria a ideia de signo aplicada ao universo da língua e nos estudos de linguística. Isso na esteira das ideias de Saussure com relação a signos, ou seja, sua aplicação em relação à língua.

Desenvolvendo a ideia de signo a partir de Saussure, para Hjemslev, a ideia de signo está associada a planos que estruturam e permitem compreender fenômenos linguísticos. Em sua teoria glossemática, ele propõe o plano da expressão e o plano do

contexto. Para cada plano, o signo é composto por sentido (pensamento), forma (limites em cada língua) e a substância que associa sentido e forma (DUCROT; TODOROV, 2010).

Por último, mas não menos importante, cita-se outro conceito no universo da Semiótica: a Semiose.

Para Abbagnano, a

> Semiótica contemporânea, das origens, até hoje, caminhou progressivamente da ideia de que seu conceito fundamental é o signo, para a ideia de que ele é a Semiose. Isso fez com que a atenção se deslocasse das estruturas do signo e dos sistemas de significação para os processos de produção do sentido, as atividades da comunicação, os mecanismos de geração dos textos e de sua interpretação. (ABBAGNANO, 2007, p. 1032)[11].

Assim como sobre Filosofia da Linguagem, há muito mais que poderia ser dito, incluindo sobre o conceito de signo, contudo, nos limites dos interesses imediatos, este breve relato atende a necessidade.

7.3 Linguística

Assim como a Filosofia da Linguagem e a Semiótica, a Linguística também tem como objeto importante de estudos a língua(gem). Em relação à Filosofia da Linguagem, por exemplo, “os interesses que justificam a atenção da Filosofia da Linguagem são no mais das vezes diferentes e distintos dos interesses que estimulam e orientam a Linguística (mas que entre eles não há conflito nem antítese)” (ABBAGNANO, 2007, p. 721).

A Linguística, como é conhecida hoje, pode ser compreendida em seu processo de desenvolvimento histórico em pelo menos oito marcos temporais, apresentados no *quadro 2*.

E, apesar da sequência cronológica, não se trata de uma simples “evolução” entre fases. De fato, alguns marcos na tabela ocorrem simultaneamente e de maneira bastante independente, como o saussurianismo na Europa e o distribucionismo na América do Norte. Trata-se muito mais de ideias e teorias que sobreviveram e, em alguns casos, ainda sobrevivem paralelamente.

[11]E, nessa linha, Umberto Eco é um dos principais representantes.

Quadro 2 – Marcos da história da Linguística

Marco	Ano	Expoentes
Port-Royal	1660	Claude Lancelot
Linguística histórica	1816	Bopp
Neogramáticos	1850-1900	G. Curtius / H. Paul
Distribucionismo	1914	L. Bloomfield
Saussurianismo	1916	F. de Saussure
Escola de Praga	Década 30	Trubetzkói, Martinet, Jakobson
Glossemática	1943	L. Hjelmslev
Gerativismo	1950	N. Chomsky

Fonte: Ducrot e Todorov (2010).
Nota: A coluna "ano" indica o ano de publicação das principais obras em cada fase.

É importante notar que as influências das obras em cada marco, em diferentes graus, até hoje são percebidas. Notadamente o saussurianismo: a obra "Curso de Linguística Geral" (SAUSSURE, 2012) ainda hoje é impressa e estudada nas universidades brasileiras, inclusive. Além disso, o gerativismo também é uma importante fase ainda presente nas universidades, notadamente na América do Norte através de Chomsky, que ainda atua na área.

Apesar de possuírem a língua(gem) como objeto maior comum, qual é o interesse específico da Linguística que a distingue de Filosofia da Linguagem e da Semiótica?

A resposta para essa pergunta é a *língua humana* produzida naturalmente no seio social[12]. Utilizou-se, até agora, intencionalmente, somente o termo língua(gem) ou "tipos de linguagem". Na verdade, parece ser o mais adequado quando se trata de Filosofia da Linguagem e Semiótica, visto que essas áreas não se ocupam somente da língua humana natural, mas também de linguagens ou línguas artificiais como as proposições lógicas, a linguagem matemática ou as linguagens de sinais. A Linguística focaliza sua atenção na língua humana especificamente.

É possível elencar algumas características das línguas humanas que oferecem uma imagem da complexidade das línguas:

> (...) são duplamente articuladas, suas unidades definem-se umas por oposição às outras, seus signos são arbitrários; a redundância está sempre presente (...); apresentam ambiguidades, dissimetrias e irregularidades; permitem a recursividade; estão em perpétua mudança; permitem a inventividade, a criatividade, o deslocamento de sentido (como nas metáforas, por exemplo), os jogos com os sons e os sentidos; são estruturadas em três níveis (o dos sons, o da gramática e o dos sentidos); seus significantes são lineares nas manifestações orais das línguas (não podemos esquecer-nos de que temos também as línguas de sinais, como Libras, a língua brasileira de sinais); suas unidades são discretas, o que significa que elas são isoláveis, diferentemente do *continuum* das cores, por exemplo. (FIORIN, 2013, p. 72).

[12]Ainda que, excepcionalmente, casos isolados de línguas humanas artificiais como o esperanto também possam ser parte da linguística.

Uma parte importante da teoria linguística será retomada ao longo da pesquisa, como a sociolinguística e a linguística histórica. Sem dúvida, a Linguística se mostrou a mais relevante para estes estudos. E ao final dessa análise sobre língua(gem), a Linguística se apresenta como um dos principais marcos teóricos de referência para esta pesquisa.

7.4 Elementos a serem destacados

A partir do que foi apresentado de maneira resumida nas seções anteriores sobre *Filosofia da Linguagem*, *Semiótica* e *Linguística*, é possível derivar algumas conclusões úteis para os objetivos da pesquisa.

7.4.1 Linguagem ou linguagens?

A língua(gem), tomada inicialmente em sentido amplo, pode ser utilizada por seres tão peculiares como abelhas para comunicar a localização de alimentos ou entre vários outros animais para complementar rituais de acasalamento. No âmbito humano, foco do interesse aqui, pode ser utilizada desde as idades mais tenras e algumas teorias (como o gerativismo) defendem que, pelo menos, parte da capacidade de língua(gem) é inata, ou seja, nascemos com ela.

Pode-se também elaborar uma língua(gem). A Filosofia da Linguagem, com a pretensão de maior clareza, objetividade e certeza, propõe o uso de proposições lógicas. A matemática é uma linguagem artificial largamente utilizada. E, no campo das linguagens artificiais, foram desenvolvidas e utilizadas metalinguagens; entre elas, as linguagens documentárias são de particular interesse, pois são parte da tecnologia atual para recuperação da informação. Finalmente, existe a linguagem humana natural: as línguas e seus grupos linguísticos, ramos, dialetos e variantes regionais, aos quais a Linguística dedica considerável esforço de pesquisa e estudo.

No que diz respeito às línguas humanas, apesar de passado mais de um século, Saussure foi o grande observador dos principais aspectos da língua. Ele tratou do lugar da Linguística em relação à Semiótica, da distinção entre uma linguística da língua e uma da fala e da importância da língua escrita (o que será detalhado mais adiante) e da mutabilidade da língua (também se dedicará atenção especial a este ponto). Esses são alguns dos pontos fundamentais abordados de maneira brilhante em seu *Cours*. Ele também atuou nas fronteiras da Filosofia, Semiótica e elaborou uma bem-sucedida proposta Linguística.

A distinção de Saussure entre língua e fala é particularmente importante como exemplificação do que é língua(gem). Aliás, é curioso observar que nem todos os idiomas distinguem de maneira objetiva, tal como nos conceitos de macho e fêmea, os conceitos de linguagem e língua. Saussure vai além e distingue língua e fala (*langue* e *parole*) (SAUSSURE, 2012), como visto adiante.

Para Saussure, se não fossem feitas as devidas distinções, a língua(gem) poderia ser estudada por áreas distintas como a Filosofia, a Antropologia e a Psicologia (SAUSSURE, 2012). Para ele, a língua é "um produto social da faculdade de linguagem e um conjunto de convenções necessárias, adotadas pelo corpo social para permitir o exercício dessa faculdade nos indivíduos" (SAUSSURE, 2012, p. 41). Ele advoga que a língua é "multiforme e heteróclita; cavaleiro de diferentes domínios, ao mesmo tempo física, fisiológica e psíquica, ela pertence [...] ao domínio individual e ao domínio social" (SAUSSURE, 2012, p. 41).

Um dos pais da linguística moderna vai além ao defender que "o exercício da linguagem repousa numa faculdade que nos é dada pela natureza, ao passo que a língua constitui algo adquirido e convencional" e que "não é a linguagem que é natural ao homem, mas a faculdade de constituir uma língua" (SAUSSURE 2012, p. 42). Saussure considera o ponto de vista individual em relação aos conceitos de linguagem e língua e aquilo em que o "indivíduo é sempre senhor" ele denomina de fala (*parole*) (SAUSSURE, 2012). Para ele, essa distinção da fala individual é fundamental, pois separa-se "o que é social do que é individual" ou "o que é essencial do que é acessório e mais ou menos acidental" (SAUSSURE, 2012, p. 45).

7.4.2 A língua(gem) pode ir além das línguas?

Pelo já exposto, a capacidade humana da língua(gem) não se esgota nas línguas naturais, nos idiomas falados, ou nas línguas artificiais como a de libras (surdos/mudos). Extrapolando os limites da Linguística, que está essencialmente preocupada com diferentes aspectos da língua humana natural (*langue*), a Semiótica procura compreender o restante do potencial da linguagem, desde códigos para comunicação (código Morse, sinais de fumaça) até os processos de significação de coisas e seu papel na linguagem: a significação de uma cruz para os cristãos ou o texto literário como um ícone de racismo ou intolerância são todos exemplos desses processos de significação mais sofisticados.

Não é possível, nos limites desta pesquisa, aprofundar todos os aspectos da língua(gem) para além das línguas. Na verdade, mesmo em relação específica a elas, apenas

destacam-se os pontos mais relevantes. No entanto, é importante destacar o que deve ficar evidente ao dialogar com Filosofia da Linguagem e Semiótica, ou seja, que a capacidade humana da língua(gem) extrapola os aspectos do uso linguístico propriamente.

7.4.3 A linguagem evolui?

Se a língua(gem), considerada aqui como uma *capacidade humana*, muda ou evolui no transcorrer da vida de um indivíduo ou, referente à humanidade, ao longo das gerações, não é a pergunta que cabe aqui, mas está claro que as línguas evoluem.

A percepção científica de que as línguas mudam está nos primórdios da história da Linguística moderna. Foi essa área que mais se preocupou e ainda tem se preocupado com o fenômeno da alteração linguística. O termo técnico para esse ramo de estudos é *Linguística Diacrônica*, estudo das línguas que as considera nas etapas diferentes em sua evolução. Quando se consideram os estudos num estágio específico de uma língua, trata-se de estudos *sincrônicos*. Essa distinção foi proposta por Saussure no começo do século XX e ainda é adotada.

Os estudos *diacrônicos* vão desde a segunda metade do séc. XIX com os neogramáticos, passando pelo estruturalismo e incluindo o gerativismo (SILVA, 2008). Desse fenômeno, destacam-se alguns pontos significativos.

Primeiro, a mudança ocorre continuamente e afeta partes e não o todo da língua, mas mantém uma estrutura essencial. O falante, em geral, não percebe esse processo, exceto em casos em que tem contato com pessoas de gerações muito antigas ou com a língua escrita utilizada no passado (FARACO, 2005). Aliás, o papel da língua escrita no fenômeno diacrônico é de grande importância para os objetivos deste trabalho.

> Elas costumam se desencadear na fala informal de grupos socioeconômicos intermediários; avançam pela fala informal de grupos mais altos na estrutura socioeconômica; chegam a situações formais de fala e só então começam a ocorrer na escrita. (FARACO, 2005, p. 26).

Os dois tipos de língua, a falada e a escrita, evoluem. Uma exerce força sobre a outra[13], às vezes adiando e às vezes motivando mudanças. Mas é a língua falada que realmente se altera nos contextos sociais e a língua escrita registra tais mudanças.

[13]Veja-se um exemplo com a orações relativas com preposição (FARACO, 2005).

7.4.4 A linguagem possui significados absolutos?

Essa pergunta origina-se das contribuições na área da Filosofia da Linguagem. A preocupação com os enunciados – principalmente em função do discurso científico e mesmo do próprio discurso filosófico – levou aqueles filósofos a se preocupar com a clareza, objetividade e verdade, aspectos do significado. Aqui o problema se limita à língua falada e à língua escrita.

Wittgenstein produziu muito sobre aspectos da língua(gem), mas as duas grandes obras que se destacam são de 1921 e 1945, sendo que a última critica as ideias da primeira. É por isso que seus comentadores falam em primeiro e segundo Wittgenstein. Uma amostra do primeiro Wittgenstein:

> Na linguagem corrente, acontece com muita frequência que uma mesma palavra designe de maneiras diferentes – pertença, pois, a símbolos diferentes – ou que duas palavras que designam de maneiras diferentes sejam empregadas, na proposição, superficialmente do mesmo modo. Assim, a palavra "é" aparece como cópula, como sinal de igualdade e como expressão da existência; "existir", como verbo intransitivo, tanto quanto "ir"; "idêntico", como adjetivo; falamos de algo, mas também de acontecer algo. (HUISMAN, 2000, p. 322).

Já no segundo Wittgenstein, "a significação de uma palavra é seu uso na linguagem" (HUISMAN, 2000, p. 322). Daí a ideia de "jogo de linguagem" associado a uma "forma de vida" e à cultura (HUISMAN, 2000). É preciso considerar especificamente quem diz, quando diz, onde diz e como diz cada enunciado.

7.5 Considerações finais desta seção

O objetivo dessa seção de revisão de literatura foi possibilitar o tratamento do conceito de linguagem com mais precisão. Foi possível estabelecer os limites entre linguagem, língua falada e língua escrita. Também foi possível definir os caminhos necessários para um maior aprofundamento nos seguintes aspectos:

1. a língua(gem) é, antes de tudo, uma *capacidade humana* que permite uma série de desdobramentos em relação ao pensamento e conhecimento humanos em sua interação com os demais indivíduos e cultura social;
2. a língua(gem) tem que ser compreendida em níveis ou tipos de aplicação em relação a uma pessoa e ao meio social no qual essa pessoa está inserida: linguagem no nível de

signos, linguagem humana falada (nomeada aqui simplesmente como língua), linguagem humana registrada (especificada simplesmente como escrita), linguagem humana corporal (gestos, movimentos faciais). Esses níveis não são estanques. É possível, por exemplo, utilizar a linguagem registrada com imagens fixas ou em movimento: fotografias, filmes. Pode-se registrar a linguagem falada. Um documento textual registrado (língua escrita) pode também ser um signo do tipo ícone de documento histórico ou sagrado;

3. pelo menos a língua falada e a língua escrita estão seguramente sujeitas a processos diacrônicos, ou seja, esses níveis de linguagem evoluem e sofrem mudanças ao longo do tempo. E o uso da língua registrada precisa levar em consideração o momento temporal em que houve o registro ou o uso da língua falada;
4. se a língua(gem) é uma capacidade humana, as línguas produzidas a partir dessa capacidade devem conter algo de idiossincrático. No caso da língua falada que é apreendida na sociedade, esse conflito entre o que é idiossincrático e social resulta em individualidade.

A seguir, é analisado especificamente o problema da mudança linguística, com ênfase no aspecto semântico, pois é o aspecto que mais atenção tem recebido, não apenas do ponto de vista da Linguística, mas também em Ciência da Informação.

8 MUDANÇA LINGUÍSTICA

Se se considerar a língua humana, como observado através do que foi registrado nos últimos três milênios, não há estabilidade e homogeneidade em nenhum momento, muito menos quando se comparam diferentes momentos históricos. Isso fica claro pela comparação entre os registros da língua nesses períodos. Mas, antes de tudo, é preciso distinguir os tipos de mudanças na língua falada a partir da teoria linguística.

O termo variação contempla o fato de que uma mesma língua considerada, num mesmo período de tempo possui versões diferentes se for considerado o aspecto social e geográfico. Isso significa que convivem num mesmo momento, por exemplo, versões consideradas cultas e outras populares (aspecto social) e várias versões regionais (aspecto geográfico).

Uma língua também pode ser vista do ponto de vista social e não apenas biológico (MENDES, 2013)[1]. Nessa visão social, é possível, inclusive, analisar pontos de vista sobre o uso "certo" ou "errado" de determinada variante da língua, visões essas que podem conter algum tipo de preconceito linguístico. Para os interesses imediatos deste trabalho, basta notar que, além das variações de pronúncia, as mais óbvias, "também se observam, nas mais diferentes línguas, variações na morfologia, na sintaxe e na interface entre esses subsistemas" (MENDES, 2013, p. 115). A área da Linguística que se ocupa do fenômeno da variação da língua é a Sociolinguística, área que terá revisão específica mais adiante.

Já o termo “mudança” aplica-se num sentido diferente. Contempla o fato de que uma determinada língua pode mudar num ou outro aspecto se for feita a comparação de uma mesma língua entre intervalos relativamente longos de tempo: "Assim, na história duma língua, pode haver mudanças fonético-fonológicas, morfológico, sintáticas, semânticas, lexicais, pragmáticas" (FARACO, 2005, p. 35).

De fato, uma mesma língua considerada está em evolução permanente em todos os aspectos acima. De forma que a observação através de registros mostra que pode-se considerar diferentes "estados da língua" ao longo do tempo, um termo introduzido pelo linguista Saussure:

[1]Mendes também nota que um grande expoente da visão social do estudo sociolinguístico é o estadunidense William Labov (1927-), o iniciador da Sociolinguística.

> Na prática, um estado de língua não é um ponto, mas um espaço de tempo, mais ou menos longo, durante o qual a soma de modificações ocorridas é mínima. Pode ser de 10 anos, uma geração, um século e até mais. Uma língua poderá mudar pouco durante um certo intervalo, para sofrer, em seguida, transformações consideráveis em alguns anos. (SAUSSURE, 2012, p. 146).

Há críticas ao conceito de estado da língua, como proposto por Saussure há quase 100 anos, pois de fato não parece haver nenhum tipo de estabilidade nas línguas vivas da sociedade em nenhum momento, daí a dificuldade de falar em um estado da língua pelo menos "estável". Por outro lado, parece indiscutível, através da comparação de diferentes registros escritos, que, se forem considerados períodos longos medidos em séculos, que houve alterações significativas nas línguas. E é isso que tem ocorrido desde que se comparam registros escritos há uns poucos milênios. O tema será retomado adiante.

Com base nesses dados empíricos (registros escritos), vê-se que um mesmo estado N_x de uma língua pode ser bastante diferente – em vários aspectos – se comparada com essa mesma língua num estado N_y onde y indica um estado anterior ou posterior no tempo em relação a x.

A observação dos registros escritos indica que, num mesmo estado N_x de língua falada, podem-se observar variações sociais e/ou geográficas. No final das contas, em se tratando da língua falada, nada é estático em nenhum momento. Mais uma vez, Saussure inaugura e explica dois termos essenciais quando se pensa em variação e mudança linguísticas: sincronia e diacronia:

> Para melhor assinalar essa oposição, porém, e esse cruzamento das duas ordens de fenômenos relativos ao mesmo objeto, preferimos falar da Linguística sincrônica e de Linguística diacrônica. É sincrônico tudo quanto se relacione com o aspecto estático da nossa ciência, diacrônico tudo que diz respeito às evoluções. Do mesmo modo, sincronia e diacronia designarão respectivamente um estado da língua e uma fase de evolução. (SAUSSURE, 2012, p. 122-123).

Sob a rubrica "variabilidade da língua" e pensando nas "variantes" que um profissional da informação pode deparar, a citação seguinte resume boa parte do que foi tratado:

> A língua não é um sistema imutável, homogêneo. Ela sofre alterações não só no tempo, como também no espaço (horizontalmente falando, isto é, no território físico de uma nação; e verticalmente, na estratificação social). De um ponto de vista sincrônico (voltado a um dado "estado da língua" em sua linha cronológica), os membros de uma comunidade linguística fazem usos da língua típicos de: (i) uma determinada região (variantes geográficas ou dialetos); (ii) uma classe socioeconômica (variantes sociais ou socioletos) e (iii) certas profissões (jargão, língua de especialidade ou tecnoletos). Do ponto de vista diacrônico (da evolução

> dos fatos linguísticos, considerados entre estados diferentes de desenvolvimento da língua), esta variação é mais perceptível no contraste de obras de épocas distintas, nas quais podemos localizar os correspondentes obsoletos (denominados arcaísmos) de termos atuais. (MELO; BRÄSCHER, 2011, p. 36).

A palavra “diacrônico” tem origem grega e significa através do tempo. Em Linguística, aquilo que é sincrônico preocupa-se com "a estrutura" da língua num momento considerado, enquanto que a perspectiva diacrônica preocupa-se com a "evolução da língua" (MAUTNER, 2011). Esta pesquisa adota um ponto de vista diacrônico, devido à preocupação com os resultados da evolução da língua, especificamente em sistemas de RI que contenham documentos históricos. Isso se traduz em ferramentas e procedimentos que tentam prever e planejar a mitigação de efeitos nos sistemas que serão utilizados no futuro. Um ponto de vista sincrônico, contrastando, preocupa-se com o cenário atual da língua e com o modo como os sistemas de RI devem "lidar" com suas características hoje.

8.1 Mudança linguística no aspecto semântico

Uma característica importante da Semântica de maneira geral e mais ainda da Semântica Linguística é que se trata de uma área de pesquisa ainda com problemas de delimitação de escopo e objeto. Isso significa que os conceitos de alguns termos essenciais dessa área como significado, significação, sentido e nomeação não são consensuais entre os pesquisadores, principalmente quando se consideram as relações entre Filosofia e Linguística. A obra considerada marco inicial dessa área é *Essai de sémantique: science des significations* (1897) de Michel Bréal[2] e, apesar do título, não parece haver consenso sobre se a Semântica é uma ciência e nem mesmo sobre o que é significado (GODOIS; DALPIAN, 2012). Atestando esse estatuto da Semântica como uma disciplina,

> A Semântica linguística hoje não se apresenta como uma disciplina de estatuto teórico e metodológico bem definido e unificado, e sim como um agregado de proposições e de práticas heteróclitas, que os mais recentes tratados de semântica se limitam a repertoriar sob rubricas distintas: sentido lexical, gramatical, referencial, pragmático, sistemático, contextual; semântica estrutural, funcional, gerativa, diacrônica, sincrônica etc., sempre na preocupação maior de manter atualizado esse catálogo e de assegurar a difusão de novas hipóteses, apresentando-as. (TAMBA-MECZ, 2006, p. 48).

Essa característica do estatuto atual da Semântica Linguística que também se aplica por extensão à Semântica Histórica, que é a que interessa particularmente, não chega a

[2]Edição traduzida para o inglês estadunidense, edição de 1964 (BRÉAL, 1964).

ser um problema para a pesquisa, pois serão utilizados alguns conceitos daquela área em um escopo de aplicação bem definido.

Nesse cenário, operacionalmente, adota-se esta definição para a disciplina Semântica: "a parte da Linguística (e mais especialmente da Lógica) que estuda e analisa a função significativa dos signos, os nexos entre os signos linguísticos (palavras, frases etc.) e seus significados" (ABBAGNANO, 2007, p. 1027). Porém, no mesmo verbete, o autor esclarece que, em função das inúmeras disciplinas que estudam a significação e o significado:

> No panorama dos estudos contemporâneos sobre a linguagem, o termo Semântica não designa uma teoria nem uma disciplina unitária, mas uma multiplicidade de abordagens e programas de pesquisa frequentemente bem distantes entre si e às vezes inconciliáveis. (ABBAGNANO, 2007, p. 1030).

Um dos fatores para esse panorama pode ser a relação dos estudos semânticos com a Filosofia,

> Por mais que possa parecer estranho, boa parte do trabalho mais importante em semântica, depois do final do século XIX, foi feito pelos filósofos, e foi preciso esperar muito para que os linguistas tomassem consciência dessas pesquisas e começassem a trabalhar em parceria com eles. (TRASK, 2006, p. 261).

Uma distinção relevante em Semântica estabelece limites entre os processos de significação e nomeação. No último caso, nomear a coisa é um processo distinto e independente do processo de significação e significa atribuir um nome para aquela mesma coisa, sem se preocupar em associar uma definição à coisa ou conceito nomeado[3].

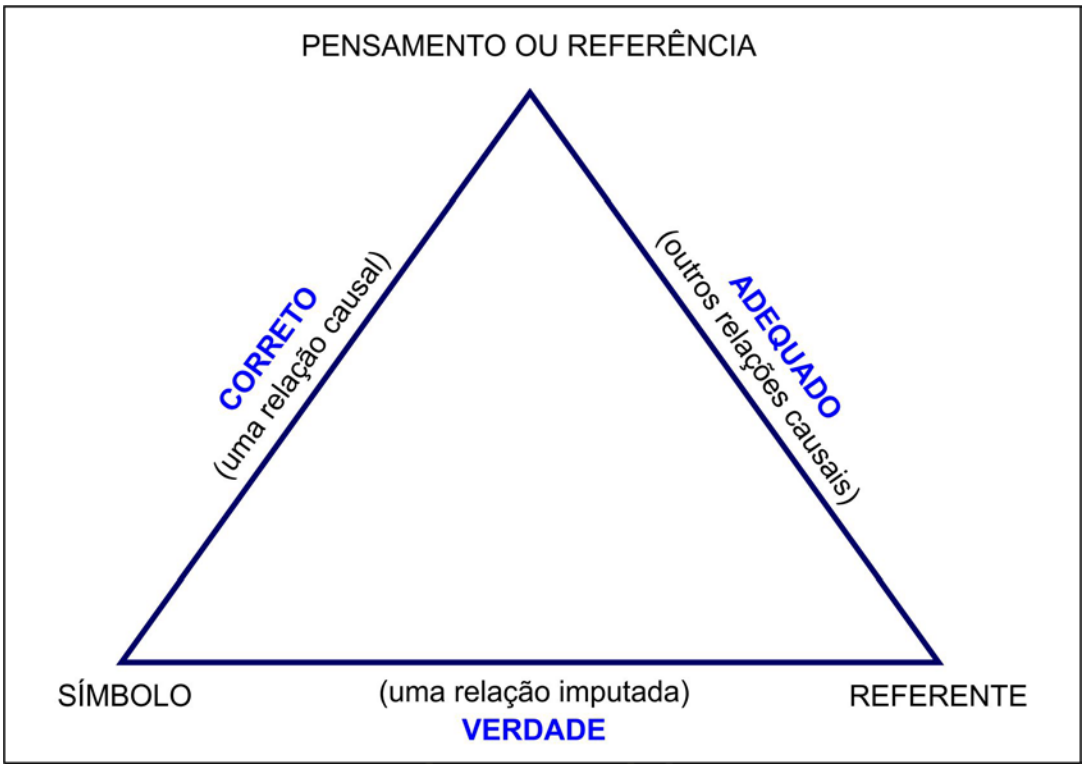
O fato de que a "ubiqüidade do sentido" (TAMBA-MECS, 2006), ou seja, a presença dele em vários ramos do conhecimento (por exemplo, em filosofia, matemática, lógica formal, crítica literária, semiótica e linguística) talvez seja uma das causas da dificuldade de sua delimitação como um campo conceitual. Talvez haja aí um paralelo com o conceito de "informação", igualmente ubíquo. O que se considera significação e significado do ponto de vista desta pesquisa? Um ponto de partida[4] é o triângulo de Ogden-Richards (OGDEN; RICHARDS, 1989), figura 5:

---

[3]A definição do processo de nomeação é excessivamente simplista se se considerar toda a atenção que a Filosofia tem dispendido a ela. Desde o Crátilo de Platão, passando por Frege (*Uber Sinn und Bedeutung*, 1892), Russell (*On Denoting*, 1905) ou Carnap (*Meaning and Necessity*, 1947) o tema tem recebido muita atenção (ABBAGNANO, 2007).

[4]Stephen Ullmann (1964) também inicia sua análise sobre significado a partir desse triângulo. A propósito, Umberto Eco redigiu a introdução de uma das edições do livro, publicado pela primeira vez em 1923: "Eu acredito que muitos acadêmicos de linguística, semântica, filosofia da linguagem e semiótica tem ao longo de suas carreiras lucrado com a leitura do livro" (OGDEN; RICHARDS, 1989, p. 5).

Figura 5 – Triângulo de Ogden-Richards



Fonte: Ogden e Richards (1989, p. 11).

Essa forma gráfica relaciona um símbolo a um referente e ao pensamento de alguém. De fato, há relações importantes entre essas três ideias que ajudam a compreender melhor o conceito de significado.

1. Relação entre um símbolo e um pensamento:

> Quando nós falamos, o simbolismo que empregamos é causado parcialmente pela referência que estamos fazendo e parcialmente por fatores sociais e psicológicos. (...). Quando nós ouvimos o que é dito, os símbolos causam em nós um ato de referência e a assumpção de uma atitude que será, de acordo com circunstâncias, mais ou menos similar ao ato e a atitude do falante. (OGDEN; RICHARDS, 1989, p. 10-11).

2. Relação entre pensamento e referente. "Mais ou menos direto ou indiretamente quando pensamos sobre ou nos referimos a Napoleão" (OGDEN; RICHARDS, 1989, p. 10-11).
3. Relação entre símbolo e referente: “Não há nenhuma relação relevante a não ser a indireta, que consiste em ser usada por alguém para significar um referente." (OGDEN; RICHARDS, 1989, p. 10-11).

Aqui interessa explorar o significado linguístico e mais especificamente ainda o significado daquilo que está escrito, pois, no final das contas, é no escrito que está o objeto da pesquisa (documentos de arquivo de guarda permanente).

8.1.1 Significado daquilo que está escrito

O triângulo de Ogden-Richards[5] abrange a ideia de significado de maneira muito ampla. Se se parte do ponto de vista do pensamento como representado lá, fatalmente se está admitindo sentido em toda sua ubiquidade, o que, inclusive, pode deixar o conceito inutilizável (ULLMANN, 1964). Nos limites dos interesses desta pesquisa, a primeira coisa a se fazer em relação ao processo de significação é limitar o escopo. Para isso, duas perguntas devem ser respondidas,

(1) Sobre o que se aplica o processo de significação? e
(2) Para que é necessário esse processo?

Assim, o ponto de vista aqui adotado sobre o problema não é o mesmo do semanticista, do filósofo de Lógica ou mesmo do linguista, é o da Ciência da Informação, com um fim bem específico: analisar os efeitos negativos da mudança linguística, também no aspecto semântico da língua, nos elementos registrados em documentos em sistemas de recuperação da informação.

Esse fim significa que a resposta da pergunta (1) acima é que a preocupação é com aquilo que está escrito (escrita), contrastando com a prioridade da Linguística para com a língua falada (língua).

O próprio triângulo de Ogden_Richards refere-se à fala e audição do que é dito (ULLMANN, 1964). Em Saussure, a ideia de significante e significado também se refere ao que é falado. "O signo linguístico une não uma coisa e uma palavra, mas um conceito e uma imagem acústica" (SAUSSURE, 2012, p. 106).

A resposta da pergunta (2) também é importante para esclarecer o ponto de vista adotado já que o processo de significação pode ser estudado por inúmeras áreas com diferentes finalidades[6]. Neste caso, interessa compreender o processo de significação do que

[5]Ideias similares de representação de signo, coisas e pensamento aparecem em várias outras construções desde o grego Aristóteles, passando pelo estadunidense C. S. Peirce.
[6]O suplemento I da edição brasileira do livro de Ogden-Richards (1972) contém um texto de Bronislaw

está escrito (em documentos) para analisar os efeitos de sua evolução em sistemas de RI. Importa esclarecer o que significa "documentos", principalmente porque isso pode incluir documentos com imagens fixas, em movimento e sonoros, ou seja, os não textuais.

No caso dos documentos textuais (mesmo aqueles que possuem ilustrações com legendas), não há maiores problemas, pois, a escrita está claramente onipresente. Mas, no caso de outros documentos cujo conteúdo não é textual, observe-se que ainda haverá texto referindo-se a seu conteúdo (por exemplo, em metadados), de maneira que o que "está escrito", mesmo em documentos não textuais, ainda ocupa um papel no processo de recuperação da informação. De fato, qualquer tipo de documento é basicamente recuperado em sistemas através de algum tipo de representação (escrita) sobre ele.

Antes de qualquer coisa, exclui-se a obrigatoriedade de discussão sobre "sentido" (que, em inglês, é traduzido como *meaning*, palavra que, em português, pode receber duas traduções: sentido e significado, a depender do contexto da obra original). Afortunadamente, a falta de uma definição consensual sobre as diferenças entre sentido e significado não terá, necessariamente, nenhum efeito prático, pois, como registra o Dicionário de Semiótica, "propriedade comum a todas as semióticas, o conceito de sentido é indefinível" (GREIMAS; COURTÉS, 2012, p. 456).

A referência para uma proposta de significado especificamente para o processo de significação daquilo que está escrito é a definição correspondente para o que é falado. Stephen Ullmann propõe e explica assim o que entende por significado de palavras:

> Ouvindo a palavra, suponhamos porta, pensará numa porta e assim compreenderá o que está a dizer aquele que fala. Para este, a sequência será exatamente a inversa: pensará, por uma razão ou por outra, numa porta, uma relação recíproca e reversível entre o nome e o sentido: se alguém ouvir a palavra, pensará na coisa, e se pensar na coisa, dirá a palavra. É a esta relação recíproca e reversível entre o som e o sentido que proponho chamar significado da palavra. (ULLMANN, 1964, p. 119).

Note-se, na explicação acima, além do uso de verbos como “ouvir” e “falar”, que ele considera dois personagens: o falante e o ouvinte. Para os objetivos desta pesquisa, há que inserir duas adaptações. Primeiro, em vez de falante e ouvinte, fala-se de autores, indexadores, editores e todo tipo de pessoa que dá origem a um documento qualquer. Esse papel será chamado genericamente de produtor (P) e, em vez de ouvinte, há naturalmente o papel de usuário (U). Segundo, no lugar de considerar o que se fala e se ouve, considera-se o que se escreve, que também será o que se lê, ainda que, dependendo de quanto tempo se lê a

---

Malinowski sobre *etnografia primitiva* onde ele declara que se ocupa da "língua primitiva, somente existente na fala atual" (OGDEN-RICHARDS, 1972, p. 295).

partir da criação do documento, o significado tenha sido alterado do ponto de vista de quem lê.

A partir dessas adaptações, adota-se a seguinte definição para significado da escrita:

(1) O resultado da relação entre aquilo que está textualmente registrado num documento ou acerca de um documento e sua respectiva definição registrada em dicionários ou equivalentes constitui seu significado.

Aqui há três elementos: (a) o texto, (b) a definição sobre aquele texto e (c) a relação entre eles.

É importante notar que, quando se trata de significado daquilo que é dito, a correspondente "definição" está na mente ou pensamento tanto de quem diz como de quem ouve e pode haver algum grau de discrepância entre um e outro. No contexto dos documentos de que se trata, isso não pode ocorrer, porque, ainda que o pensamento esteja presente tanto na pessoa que produz como também naquele que lê o documento (e que antes busca em sistemas para recuperá-lo), não é esse processo cognitivo que importa aqui, pois os sistemas de RI que armazenam documentos não armazenam pensamentos e não podem lidar com esse elemento. De maneira que o apoio estará na "definição sobre aquele texto" presente em algum elemento fisicamente registrado, o que significa outro texto ou metatexto sobre o primeiro (um tesauro ou algum tipo de terminologia, para citar exemplos em Ciência da Informação, mas também dicionários linguísticos).

Neste estudo, "definição" de significado refere-se que é textualmente registrado sobre aquilo que é escrito. As fontes possíveis são dicionários onomasiológicos ou semasiológicos, leis, documentos históricos, estudos descritivos sobre a história da instituição produtora dos documentos. A proposta (1) passa a ter com a seguinte redação:

(2) O resultado da relação entre aquilo que está textualmente registrado num documento ou acerca de um documento e suas respectivas definições para cada elemento linguístico registrada em dicionários ou equivalentes constitui o significado.

Na forma como esta definição foi apresentada, além de estar em harmonia com a tecnologia hoje disponível para sistemas de RI, ela permite a verificação empírica para cada relação escrita-significado.

Há outro aspecto semântico que deve obrigatoriamente ser abordado: trata-se de alguns fenômenos da língua falada que acabam se refletindo nos registros escritos. Os mais comuns são a polissemia, a sinonímia e a homonímia.

Na obra por muitos considerada a fundadora da Semântica (1897), Michel Bréal (1900) aborda a polissemia e sua relação com a mudança linguística – "um novo significado para uma palavra", declarando que "o mesmo termo pode ser empregado alternativamente no senso estrito ou metafórico, no senso restrito ou expandido, no senso abstrato ou concreto" (BRÉAL, 1964, p. 139). E lista vários exemplos, inclusive em línguas antigas (BRÉAL, 1964).

Stephen Ullmann, tratando do mesmo problema na metade do século XX, aborda o fenômeno dos sinônimos (sinonímia). Nota, inclusive, que esse fenômeno "bastante paradoxalmente, encontra-se onde menos seria de esperar: nas nomenclaturas técnicas." (ULLMANN, 1964, p. 292). Num capítulo que dedica a "ambiguidades", o mesmo autor elenca a polissemia e a homonímia (mesma grafia ou mesmos sons referindo-se a definições diferentes).

Finalmente, Bernard Pottier aborda o tema, incluindo também no grupo a metáfora e a paráfrase, e declara: "A correspondência um a um entre um sinal e um ’sentido’ não existe em qualquer idioma" (POTTIER, 1992b, p. 40).

8.1.2 Mudança do significado

Basicamente, trata-se da evolução semântica dos conceitos e os respectivos elementos linguísticos (P) que os designam. Também já se analisaram os fenômenos que causam relações entre termos (como a sinonímia) e entre vários termos e vários significados (homonímia) ou entre vários termos e um significado (polissemia). De maneira que já está claro que as relações entre conceitos e o texto que os designam não são relações unívocas, pelo contrário.

As fontes de que se dispõe não estão necessariamente preocupadas com a evolução da escrita, nem sempre deixam claro o objeto e consequências dessa evolução acerca da escrita. Sabe-se que a semântica linguística tem como foco a língua falada e viva no mundo, mas também se está ciente de que, pela observação de registros documentais, os significados da língua se refletem na escrita, daí serem essas causas de evolução, também causas de evolução da escrita, no sentido que é possível que elementos passem a ser registrados com novos significados, além de surgirem novos elementos para designar

significados já consolidados.

A Filologia tem analisado a mudança linguística por meio dos registros disponíveis e outros estudos[7], a Linguística Histórica também tem se preocupado com essa evolução, em todos os seus aspectos, mas é a Semântica Histórica que mais interessa nesta seção e é ela que irá fornecer os subsídios necessários ao trabalho.

É imperativo esclarecer que a área Semântica Histórica é vasta em seus interesses e aspectos da linguagem a serem analisados, como por exemplo a relação cognitiva e as mudanças semânticas, análises específicas de formação de palavras em línguas nativas, possíveis leis para mudanças semânticas[8], modelos cognitivos ou de linguagem, além de outros. Isso sem considerar a própria história da Semântica e da Semântica Histórica. Em relação à língua específica tratada, podem-se analisar partes verbais, outros componentes gramaticais como preposições e conjunções ou o vocabulário significativo propriamente (o que designa conceitos concretos ou abstratos).

Um autor clássico que influenciou bastante a Semântica e a Semântica Histórica foi Stephen Ullmann (ROTH, 1998). Duas de suas obras são consideradas fundamentais: *The Principles of Semantics* (1951) e *Semantics: an introduction to the science of meaning* (1962). No mais antigo dos livros, ele dedica dois capítulos à *Historical Semantics*. No início do quarto capítulo, ele define: "Semântica sincrônica é a ciência do significado, semântica diacrônica a ciência da mudança do significado" (ULLMANN, 1957, p. 171). O capítulo, por si só, é uma densa análise sobre alterações semânticas. Há ali também vários outros assuntos que servem para demonstrar a importância basilar daquele trabalho. O trabalho mais tardio de Ullmann parece retomar as discussões do primeiro livro, mas dedica apenas um capítulo ao assunto da mudança do significado, "De todos os elementos linguísticos arrebatados no seu curso, o significado é, provavelmente, o que menos resiste à mudança" (ULLMANN, 1964, p. 401).

Sabe-se que todos os aspectos da língua estão sujeitos à mudança, no entanto o aspecto fonético-fonológico e o semântico são os que têm recebido mais atenção, desde o início dos estudos da Linguística.

---

[7]No caso específico da Língua Portuguesa, um dos mais importantes trabalhos é a "História da Língua Portuguesa" (SILVA NETO, 1952).

[8]Uma dessas possíveis leis foi a ideia de deriva da língua, mas nenhum dicionário atual consultado mantém qualquer verbete sobre esse assunto.

8.2 Sociolinguística quantitativa

A sociolinguística, às vezes também alcunhada etnolinguística ou antropologia linguística (DUCROT; TODOROV, 2010), surge em meados da década de 1960, com destaque para os trabalhos de William Labov, que procurou adicionar componentes externos à língua (SILVA, 2008).

Com "elementos externos", quer-se dizer que a linguística estruturalista estuda a língua (langue, na terminologia saussuriana). A fala (parole), no seio da sociedade na qual é usada, é vista como algo externo à língua, "para Labov, a língua não se ’localiza’ na mente de seu falante, mas no seu uso por uma comunidade de falantes" (MENDES, 2013, p. 113).

Numa sociedade concreta e real, o uso da língua está sujeito a situações sociais. A sociolinguística fala então em "variação" do uso. Mas há uma relação com a ML, "o estudo diacrônico estuda a variação que se verifica ao longo do tempo e que pode vir a causar uma mudança na língua" (VIOTTI, 2013, p. 146).

O que caracteriza a sociolinguística é sua relação específica com o meio social em que uma língua é falada. A linguística tem como objeto a língua falada num meio social, pois os indivíduos nascem num meio social (pelo menos é uma pressuposição razoável) no qual está em uso uma língua, a qual aprendem, adotam e usam para se comunicar com outros indivíduos.

A sociolinguística não é a única área que se ocupa da língua em seu uso num contexto social qualquer, podem-se citar também outras disciplinas que estudam essa relação como a linguística histórica, a análise do discurso e a linguística aplicada (COELHO et al., 2015). Há contribuições de linguistas não considerados como sociolinguistas que produziram teorias muito antes da moderna sociolinguística, mas com ênfase na relação língua-social, como Voloschinov e Basil Bernstein (CALVET, 1975).

A sociolinguística aqui é a quantitativa (SQ), cujo iniciador foi o americano William Labov (TARALLO, 1999). A SQ, ao abordar o uso da língua no meio social, caracteriza-se pela ênfase em métodos para coleta e tratamento de dados acerca desse uso específico,

> Para desvelar tanto a estrutura linguística quanto a estrutura social, devemos, necessariamente, coletar grande quantidade de dados de muitos indivíduos; consequentemente, devemos enfrentar problemas ligados a controle de qualidade e confiabilidade, o manuseio e apresentação de dados, e a interpretação e inferência. (GUY; ZILLES, 2007, p. 19).

A SQ possui dois pressupostos teóricos que dão suporte à sua teoria da variação e mudança (TVM). O primeiro pressuposto é que a língua é um sistema organizado, o outro é que ela varia (COELHO et al., 2015). Nesses pressupostos, ser um sistema organizado não significa ser imutável e não dinâmico, a variação observada decorre de fatores linguísticos e extralinguísticos (por exemplo, fenômenos e questões sociais: música, política, guerras). Complementando e mesmo decorrendo desses pressupostos, há dois conceitos igualmente fundamentais para a compreensão da TVM: variável e variantes linguísticas.

Com relação ao primeiro pressuposto, língua vista como um sistema organizado é uma concepção relativamente recente, já se abordou a ideia correlacionada a essa de língua como um sistema complexo. No entanto, a percepção de que a língua varia ou muda, ainda que esses termos não tivessem o significado atual, está na base da história da linguística em sua fase histórico-comparativa (ELIA, 1987).

A importância de ver a língua como um sistema exige uma revisão específica sobre o tema, o que será feito em seção mais adiante. Aqui cabe notar que a SQ pode se ocupar das variações geográficas de uma língua e mesmo elaborar mapas dessas variações (COELHO et al., 2015). E também divide o trabalho de observação em relação às variações em diferentes níveis linguísticos: lexical, fonológico, morfo(fonológico, lógico e sintático), sintático e discursivo. Comparando-se esses níveis com o que muda do ponto de vista da mudança linguística (que se verá mais adiante), notam-se claras semelhanças.

No caso do nível de variação lexical, um dos mais evidentes nas observações, o quadro 3 apresenta alguns exemplos observados geograficamente na língua portuguesa, de um ponto de vista sincrônico.

Quadro 3 – Exemplos variação lexical

Item	Variações observadas
1	Abóbora, jerimum
2	Bergamota, vergamota, tangerina, laranja-cravo, mimosa
3	Mandioca, aipim, macaxeira
4	Pão francês, pão de trigo, cacetinho, filãozinho
5	Banheiro, toalete, w.c.,casinha
6	Coisa, troço, trem
7	Estojo, penal
8	Pandorga, pipa, papagaio
9	Vaso, bacio, privada

Fonte: Extraído de Coelho et al. (2015, p. 24).

Já com relação ao pressuposto de que a língua varia, é preciso ilustrar os conceitos de variável e variantes linguísticas. Com esse intuito, veja-se um exemplo na língua portuguesa para o uso do plural, extraído de Tarallo (1999). É possível identificar em falantes da língua portuguesa pelo menos três construções de sentenças nas quais o plural é flexionado de maneira diferente em cada sentença:

1. aS meninaS bonitaS
2. aS meninaS bonita[X] ,onde [X] significa omissão do componente
3. aS menina[X] bonita[X]

O "S" nas construções acima, em relação ao plural, é uma variável linguística, pois encontra-se em estado de variação em relação ao uso de diferentes indivíduos. Em SQ, fala-se que há heterogeneidade no uso social. Na verdade, um mesmo indivíduo pode fazer usos diferentes em situações sociais diferentes. Cada uma das três construções acima são chamadas de variantes linguísticas. No modelo de análise da TVM, esses conceitos são representados assim:

<S> ---------- [S]
---------- [X]

Toda a metodologia da SQ que tem sido aprimorada nas últimas cinco décadas objetiva colher dados sobre as variantes linguísticas que comprovem ou não as variáveis linguísticas propostas.

Desde os primeiros trabalhos de Labov e seus colegas que deram o impulso inicial para o que hoje é uma SQ, a coleta e análise de dados linguísticos ocorreram no âmbito sincrônico da língua. Mas as pesquisas no âmbito diacrônico, no qual se fala em mudança da língua em vez de variação da língua, também estão presentes na SQ, "a heterogeneidade da língua é observada tanto na sincronia como na diacronia, a língua não passa por períodos menos sistemáticos, ainda que esteja em constante mudança" (COELHO et al., 2015, p. 71).

Para a TVM, há uma relação entre variação e mudança linguísticas e essa relação significa que nem todo aspecto socialmente variável da língua falada significa incorporação na língua na forma de uma variante padrão, ou seja, com maior prestígio social. Mas as mudanças ocorridas devem ter passado por processos de variação linguística. Em outras palavras, admite-se que a mudança linguística (ML) depende da variação linguística, mas não o contrário.

De um ponto de vista da SQ, o fenômeno da ML deve ser analisado com base em análise empírica de dados variáveis: “fontes necessárias para se confirmar que as possibilidades de diferenciação das formas em variação estão dispostas ordenadamente na língua, isto é, que a heterogeneidade é sistemática e ordenada. Descrever dados empíricos em variação/mudança não é uma tarefa fácil” (COELHO et al., 2015, p. 76). Para dar conta destes procedimentos, a SQ preconiza questões gerais (problemas) a que um pesquisador deve responder a fim de conduzir o trabalho de análise da ML, o quadro 4 resume esses problemas.

Quadro 4 – Problemas empíricos para análise da mudança em SQ

Problema	Definição
Restrição	Qual é o conjunto de mudanças possíveis e de condições para mudanças que podem ocorrer em uma determinada estrutura?
Encaixamento	Como as mudanças estão encaixadas na estrutura linguística e social?
Transição	Como as mudanças passam de um estágio a outro?
Avaliação	Como as mudanças podem ser avaliadas em termos de seus efeitos sobre a estrutura linguística, sobre a eficiência comunicativa e sobre o amplo espectro de fatores não representacionais envolvidos no falar?
Implementação	A que fatores se pode atribuir a implementação das mudanças? Por que uma mudança ocorre em uma língua em uma época e não em outra língua e em outra época?

Fonte: Coelho et al. (2015, p. 76).

No entanto, é importante ressaltar que a busca das respostas aos problemas acima está totalmente fora do escopo desta pesquisa, que são problemas a serem analisados de um ponto de vista da linguística. E se são reproduzidas aqui é justamente com a finalidade de evidenciar isso. Nesta pesquisa, interessam os efeitos que as mudanças concretas ocorridas, independentemente dos motivos que levaram a essa ocorrência, podem e poderão ocasionar em sistemas de RI.

Nos termos da TVM algumas variantes que possuíam menos prestígio social podem desaparecer em função do tempo, mas podem também assumir um papel social mais importante. No caso da língua portuguesa brasileira, há variantes com maior prestígio social em relação a outras variantes. Se comparados diacronicamente, esses estados mostrarão variantes padrão diferentes. O grau de diferença será maior em função da distância temporal entre os estados comparados.

É possível antever o fenômeno acima descrito não apenas quando se comparam estados anteriores da língua portuguesa (considerando, por exemplo, a variante padrão), mas também em relação aos futuros estados da língua portuguesa. No último caso, trata-se do recorte que de fato interessa neste trabalho.

Na análise documental na parte de procedimentos, será possível observar vários exemplos de elementos linguísticos que estiveram sujeitos à ML, comparando-os com o estado atual da língua portuguesa. Nesse caso, no entanto, não foi feita uma análise das variantes linguísticas naquele momento quando os documentos foram registrados. Trata-se de uma análise comparativa diacrônica.

8.3 A língua como um sistema

A linguística que se preocupa com a ML é diferente do estudo da história das línguas (JANSON, 2015) e mesmo da história da escrita (GELB, 1952), mas várias áreas da linguística se ocupam da ML. Assim, há análises independentes do ponto de vista fonético-fonológico ou semântico. A grande riqueza de recursos e problemas científicos das línguas permite também várias abordagens de diferentes pontos de vista. A complexidade do problema exige segmentação do trabalho de investigação acerca do tema ML.

Já no início do séc. XIX (1821, mais especificamente), havia interesse pela ML, quando foram propostas "causas", que podem ser agrupadas em fisiológicas, psicológicas, sociais e racionais (SILVA, 2008).

A disciplina Linguística Histórica, produto do século XX, parece tentar abarcar esse fenômeno,

> A realidade empírica central da linguística histórica é o fato de que as línguas humanas mudam com o passar do tempo. Em outras palavras, as línguas humanas não constituem realidades estáticas; ao contrário, sua configuração estrutural se altera continuamente no tempo. (FARACO, 2005, p. 14).

No caso brasileiro, já havia estudos históricos da língua portuguesa no final do séc. XIX sob a rubrica Filologia e houve uma retomada no final do séc. XX (FARACO, 2005). Contudo analisar qual disciplina exatamente aborda a ML é menos importante para os interesses desta pesquisa do que esclarecer melhor o que muda numa língua qualquer em função do tempo. E essa resposta, por sua vez, deve se apoiar em uma teoria linguística. Tal exame tem que passar por, pelo menos, duas linhas linguísticas hoje tradicionais: o estruturalismo saussuriano e o gerativismo.

O estruturalismo de base saussuriana é a linha mais antiga e ainda influente na atualidade. Inclui várias correntes que dele se originaram ao longo do séc. XX e algumas contestações sobre suas teorias. De fato, foi essa linha que estabeleceu a dicotomia

sincrônico-diacrônico – e a precedência do estudo sincrônico (FARACO, 2005). Num primeiro momento da proposta teórica (início do séc. XX), deu-se atenção para o estudo da mudança fonética e etimologia como em Saussure (2012).

O gerativismo, por sua vez, teve suas bases estabelecidas por Noam Chomsky na década de 1950. Sua grande inovação foi a concepção de um modelo que considera a capacidade inata para aprender uma língua qualquer. Nas palavras[9] do próprio Chomsky: "A linguagem humana parece estar biologicamente isolada em suas propriedades essenciais e ser um desenvolvimento na verdade recente sob uma perspectiva evolucionista" (CHOMSKY, 1998, p. 17). Uma criança submetida a uma língua domina-a a partir de um aparato biológico que ela já traz consigo ao nascer, tanto quanto outros órgãos do corpo. Isso resulta na ideia de competência (regras de uma língua) e desempenho (uso efetivo da língua) (WEEDWOOD, 2002).

É importante acrescentar que, dentro do que pode ser agrupado como estruturalismo e gerativismo, há diferentes e importantes escolas que abordaram o problema da ML, como a Escola de Praga e a abordagem sistêmica da língua (SILVA, 2008).

É possível comparar o estruturalismo e o gerativismo a partir de suas dicotomias fundamentais acerca de suas respectivas concepções de língua (VIOTTI, 2013). Ainda segundo Viotti (2013), no estruturalismo a dicotomia ocorre entre a face social da língua (langue) e a face individual (parole). No gerativismo, por sua vez, a dicotomia ocorre entre a língua-e (eventos reais da fala) e a língua-i (parte mental da língua). Nas duas dicotomias, opõe-se o individual (parole e língua-i) do social ou coletivo (langue e língua-e). Dessa interação complexa, vários pesquisadores têm indicado a necessidade do conceito de língua como sistema e um dos fenômenos observáveis nesse sistema é a ML. De fato, uma análise da história da linguística e da ML evidencia, em vários momentos, a necessidade de observar e analisar a língua como um sistema. A própria sociolinguística precisa ver a língua (inter)relacionada ao sistema social. Em trabalhos estruturalistas ainda na década de 1930, já aparece um "princípio da abordagem sistêmica da diacronia" (FARACO, 2005, p. 158).

Atualmente, "vários pesquisadores têm, explícita ou implicitamente, sustentado a hipótese de que a língua humana é mais bem caracterizada como um sistema complexo, dinâmico e adaptativo" (VIOTTI, 2013, p. 149). A concepção de língua como sistema possibilita até mesmo uma crítica à dicotomia sincrônico-diacrônico:

[9]Palestra proferida por Chomsky na Universidade de Brasília (Unb) em novembro de 1996.

> Todos sabemos que la sincronía no existe en la realidad: aparece como una etapa necesaria de la investigación. El tiempo, que nos domina siempre, hace que cualquier estudio tiene que incorporarse en una visión diacrónica, y esto vale para el significante (la fónica y la gráfica históricas), para los signos léxicos o para el funcionamiento semántico-sintáctico (el **sistema** nominal, el **sistema** verbal, el **sistema** relacional), sin olvidar las estructuras textuales. (POTTIER, 1992a, p. 111, grifos nossos).

A atualidade da proposta de considerar a língua como um sistema (complexo) pode ser notada, por exemplo, em uma publicação recente que organizou vários autores e seus pontos de vista especificamente para discutir a língua como um sistema complexo:

> Estritamente de um ponto de vista linguístico, a língua (language) é um sistema complexo no qual fonética, morfologia, sintáxe, semântica, léxico, pragmática... interagem de maneira a produzir sentenças aceitáveis. Mas, língua é também um sistema complexo do ponto de vista da natureza. Língua é um objeto fisiológico, neurológico e psicológico. E, é claro, uma entidade sociológica. (BEL-ENGUIX; JIMÉNEZ-LOPEZ, 2010, prefácio).

Analisar o problema da ML a partir de uma visão da língua como um sistema complexo, considerando o aspecto individual e social ao mesmo tempo e com o mesmo peso, significa também reconhecer a complexidade dos elementos da língua que mudam com o tempo. Esse é o assunto de nossa próxima seção.

8.4 O que muda na língua

É preciso analisar e descrever os componentes da língua para identificar, com mais precisão, o que exatamente muda ou pode mudar numa língua em função do tempo.

Tradicionalmente, o exame de uma língua qualquer pode ser dividido em três partes, considerando aquilo que é mais exterior e o que está mais próximo à significação: meios materiais, gramática e dicionário (DUCROT; TODOROV, 2010).

Por meios materiais, entendem-se os sons (pronúncia, audição) e a escrita (formas de escrita, alfabeto, texto). A gramática se divide em morfologia, que aborda os elementos linguísticos independentemente de suas (inter)relações na frase ou período, e sintaxe, que trata dos elementos linguísticos (inter)relacionados de acordo com os critérios de oposição, complementação etc. E finalmente o dicionário refere-se aos elementos linguísticos ou léxico em relação aos sentidos que possuem e os vários problemas de sentido (sinonímia, homonímia etc.). A Linguística, ao longo do séc. XX e atualmente, critica essa divisão tradicional. (DUCROT; TODOROV, 2010).

Em linguística contemporânea, parece haver unanimidade sobre a existência de pelo menos seis aspectos de estudo e abordagem da língua: fonético-fonológico, morfológico, sintático, semântico, lexical e pragmático, porém isso não significa que exista a mesma unanimidade sobre o que são esses aspectos ou mesmo se há limites claros entre eles. Sobre o aspecto lexical, por exemplo, de um ponto de vista lexicográfico:

> Léxico está situado em uma espécie de intersecção que absorve informações provindas de caminhos diversos: dos sons (fonética e fonologia), dos significados (semântica), dos morfemas (morfologia), das combinações sintagmáticas (sintaxe) ou do uso linguístico e das situações comunicativas (pragmática). (ALVES, 2007, p. 77).

Mas de um ponto de vista semiótico,

> Léxico é a lista exaustiva de todas as lexias de um estado de língua natural. O valor desse conceito, de caráter operatório, deve ser apreciado em função do de lexia, de sua capacidade, principalmente, de ser tomada como unidade de base para a análise semântica. (GREIMAS; COURTÉS, 2012, p. 285).

Fica claro da análise do ponto de vista lexicográfico e do semiótico que tratar de um aspecto (no exemplo, o lexical) pode envolver – e frequentemente o faz –também outros aspectos. Donde se conclui que não há limites fixos entre esses aspectos. Todos eles – de uma forma ou de outra – estão (inter)relacionados.

Dois motivos, pelo menos, são os responsáveis pela impossibilidade de apresentar esses aspectos de estudo de uma língua qualquer sem suas (inter)relações: (1) a própria indefinição sobre o que é linguística e (2) descompasso no desenvolvimento de suas áreas de estudo.

Sobre a indefinição do que é linguística como uma área científica, convém destacar que se trata de uma disciplina que começou a ser ministrada em universidades muito recentemente. No Reino Unido, por exemplo, os primeiros cursos começaram apenas em 1964 (CRYSTAL, 1981). No Brasil, o curso de Linguística foi implantado nos cursos de Letras em 1961 e a pós-gradução, apenas uma década depois disso. Antes disso, os primeiros cursos de Letras desde a década de 1930 tratavam de filologia e de uma perspectiva histórica normativa da língua portuguesa (FIORIN, 2006).

A história do estudo da língua portuguesa no Brasil remonta à segunda metade do século XIX (VAREJÃO, 2009). Todavia, o estudo da língua mediante um método científico e de maneira ampla, incluindo a própria visão de sua própria história de maneira sistematizada, é um empreendimento de meados do século XX e, em alguns aspectos ainda está iniciando, como se verá adiante.

Sobre o descompasso no desenvolvimento das áreas que estudam uma língua, é preciso ressaltar que inicialmente havia ênfase apenas no aspecto fonético-fonológico e nas gramáticas que englobavam a morfologia e a sintaxe (sob uma ótica tradicional, como visto antes). A semântica teve seu desenvolvimento científico em meados do séc. XX, apesar de o termo ter sido cunhado no final do séc. XIX (LEROY, 1967). Com a pragmática, o descompasso é ainda mais marcante, manuais comuns nos estudos linguísticos no Brasil no final da década de 1980 como Borba (1986) ou Lyons (1987) sequer elencam a pragmática em seus sumários. E praticamente no século XXI, "a pragmática ainda é vista por muitos estudiosos, não sem razão, como um verdadeiro saco de gatos" (RAJAGOPALAN, 1999, p. 1).

O cenário, portanto, para alcançar o propósito de definir com maior precisão o que muda no processo de ML precisa ser analisado com as ressalvas necessárias em relação aos aspectos históricos do estudo da língua. Epistemologicamente, esses diferentes aspectos estão em construção, alguns que são objeto de estudo mais antigo como o fonético-fonológico estão mais consolidados, outros ainda não. Isso se coaduna com a conclusão da seção anterior sobre a necessidade de analisar uma língua como um sistema complexo. Operacionalmente, serão consideradas as definições para os diferentes aspectos a partir da Linguística Geral, *quadro 5*, mas enfatizando as (inter)relações[10] necessárias entre esses entes.

Quadro 5 – Aspectos passíveis de mudança linguística

Aspecto	Definição básica
Fonético/ Fonológico	O aspecto fonético refere-se "aos sons da língua em sua realização concreta, independentemente de sua função linguística" que cabe à fonologia a qual por sua vez se ocupa "dos sons da língua do ponto de vista de sua função no sistema de comunicação linguística (pronúncia, sotaque, acentos)" que pode compreender tanto os sons comuns em várias línguas como específicos para cada caso em particular (fonologia do inglês). Fonética e fonologia possuem a mesma etimologia e inicialmente eram tratadas como uma única ciência. A fonologia se firmou como independente apenas na década de 1930. Mas evidentemente são extremamente relacionadas, ambas estudam os sons humanos.
Morfológico	O aspecto morfológico refere-se "as regras que regem a estrutura interna das palavras, isto é, as regras de combinação entre os morfemas-raízes para constituir palavras e a descrição das formas diversas que tomam essas conforme a categoria de número, gênero, tempo, pessoa e conforme o caso de flexão".
Sintático	O aspecto sintático refere-se à descrição "das regras pelas quais se combinam as unidades significativas em frases". A expressão análise sintática significa justamente a verificação das funções das unidades em orações, frases e períodos.

continua...

[10]É preciso não confundir a clara distinção entre esses aspectos no tocante às especializações de estudo. Assim, o trabalho de um profissional especializado em fonética distingue-se claramente do trabalho de um linguista especializado em semântica. Mas, em termos de relações, os elementos linguísticos precisam ser vistos em sua totalidade: som e significado, por exemplo.

continuação

Semântico	O aspecto semântico refere-se às relações dos elementos linguísticos com seu significado. Os maiores problemas nesse aspecto referem-se à própria definição de significado (não se confunde com referência que é a relação denotativa entre um objeto e a palavra utilizada para se referir ao objeto). Pode extrapolar o elemento linguístico individual e abordar também frases, orações ou mesmo discursos (textos).
Lexical	O aspecto lexical refere-se ao "conjunto das unidades que formam a língua de uma comunidade, de uma atividade humana, de um locutor etc." Uma oposição comum é entre léxico e vocabulário (unidades do discurso: vocabulário).
Pragmático	O aspecto pragmático refere-se "às características de sua utilização (motivações psicológicas dos falantes, reações dos interlocutores, tipos socializados da fala, objeto da fala etc.)".

Fonte: Extraído de Dubois et al. (1998).

8.4.1 Exemplos do que muda

Empreende-se a seguir uma análise de casos concretos e reais de ocorrências na fala que podem afetar ou ocasionar a ML. Concomitantemente, os exemplos ajudarão a ilustrar a dificuldade de separação desses aspectos[11]. Exceto quando notado, todos os exemplos foram extraídos do trabalho de análise sobre ML em Viotti (2013).

Na língua portuguesa atual, há estruturas padrão que podem ser agrupadas em intransitivas ou transitivas (essa última com ou sem complemento preposicionado).

Sentenças intransitivas:

(1) O menino correu.

(2) A criança dormiu.

Sentenças transitivas:

(3) O menino chutou a bola.

(4) A mãe beijou o nenê.

Sentenças transitivas com complemento preposicionado:

(5) A polícia atirou no ladrão.

(6) O motorista bateu no muro.

[11]É importante notar aqui que as análises ocorrem na fala (real) de pessoas em grupos sociais. Para essa análise não pode haver qualquer tipo de "preconceito linguístico" (BAGNO, 1999) por comparação com o Língua vernácula oficial (considerada muitas vezes como a culta). O uso na fala explica em parte a ML da Língua Oficial "Os estudiosos compreenderam, mais claramente que antes, que as mudanças na língua dos textos escritos correspondentes a diversos períodos - mudanças do tipo da que com os séculos transformou o latim em francês, italiano ou espanhol, por exemplo - poderiam ser explicadas em termos de mudanças que haveriam ocorrido na língua falada correspondente" (LYONS, 1987, p. 9).

Esses padrões podem, no entanto, ser utilizados com novos itens candidatos a fazer parte do léxico oficial da língua.

(7) Eu *deletei* o arquivo.
(8) O professor *escaneou* o artigo.

Verbos novos como deletar, escanear, escaipar ou xerocar se adequam, tanto fonologicamente quanto morfossintaticamente, à língua, sempre com base em estruturas padrão (fonológicas, morfológicas e sintáticas).

No surgimento de novos usos para elementos já utilizados no léxico, também pode ocorrer esse processo a partir de um novo uso baseado numa concepção padrão já existente, como o termo "barraco".

Barraco possui o significado comum de casa pequena → precária → malfeita → bagunçada, daí utilizar o termo barraco também no sentido de confusão e precariedade.
(9) Aprontei um grande *barraco* na festa (exemplo nosso).

Outro exemplo é o uso de novos particípios no português, que têm aparecido nas formas coloquiais de uso.

(10) O João já tinha *chego* quando a Maria saiu. (para *chegar-chegado*)
(11) Eu também tinha *falo* a mesma coisa pro Pedro. (para *falar-falado*)

Ou para verbos terminados em (-er) e (-ir):

(12) Eu teria trago tudo o que você precisava pra festa. (para trazer-trazido)
(13) O Pedro tinha peço pra a gente comprar mais refrigerante. (para pedir-pedido).

Outros aspectos da língua que têm sido observados são a alteração de estruturas padrão, como aquelas exemplificadas no começo – uso de complemento com intransitividade. Os verbos chegar e sair são usados normalmente como intransitivos (não pedem complemento).

(14) O Pedro *chegou* o sofá até a janela.

(15) O homem *saiu* o carro da garagem.

Da mesma forma, também se verificam usos de verbos normalmente transitivos em situações normalmente intransitivas (um único argumento).

(16) Meu jardim *destruiu* todo.

(17) O programa que eu queria não *instalou*.

O que se pretende demonstrar com os exemplos acima é que, em todos os casos, os diversos aspectos da língua (fonético-fonológico, morfológico, sintático, semânticos, lexical e pragmático) concorrem no uso e consequentemente como candidatos no processo de mudança dentro do sistema,

> Absolutamente todas as inovações são absorvidas pelo sistema. A mudança é a vida do sistema. Portanto, a noção tradicional de que a mudança é estranha ao sistema não se mantém na perspectiva da língua como um sistema complexo, dinâmico e adaptativo. (VIOTTI, 2013, p. 171).

Finalmente, é importante esclarecer que – de um ponto de vista da Ciência da Informação e orientado pelos objetivos desta pesquisa – não interessa aqui abordar questões que são objeto de estudo específico da Linguística, por exemplo: "Quais são as causas da mudança?" "Como ocorre a mudança no meio social?", "Há ubiquidade na mudança linguística?[12]" "É possível haver previsibilidade na mudança linguística nos diferentes aspectos?[13]".

8.5 Considerações finais desta seção

Esta seção tratou do fenômeno da ML, um aspecto fundamental para o desenvolvimento desta pesquisa. Procurou-se delimitar o que é variação e mudança linguística, com base em teorias linguísticas. Há relações entre ambos os fenômenos, mas esta pesquisa refere-se aos efeitos da ML, especificamente. Também foi esclarecido o que exatamente muda numa língua em função da ML. Assim, fica claro que o aspecto semântico é

[12]Todas as línguas registradas e submetidas à análise, até agora, evidenciaram algum tipo de mudança ao longo do tempo.

[13]No aspecto fonético as *Leis de Grimm* pretendem prever a evolução de uma língua. No aspecto semântico, os primeiros trabalhos de *Bréal* no final do séc. XIX pretendiam prever a evolução de significado dos termos.

um dos mais estudados em termos de efeitos da ML. Finalmente, como apoio aos objetivos e demais procedimentos empíricos nesta pesquisa, analisar a língua como um sistema complexo é importante, pois nenhum aspecto linguístico está sujeito à ML independentemente dos demais.

Na seção seguinte, é analisada, com mais detalhes, a escrita em contraste à língua.

9 A LÍNGUA ESCRITA

Essa seção trata do conceito de língua escrita ou simplesmente escrita. Trata-se de uma abordagem importante no contexto desta pesquisa, principalmente em sua diferenciação em relação ao conceito de língua falada. Antes de tratar especificamente do tema, cabe uma revisão sobre o papel da memória (e sua preservação) para a ciência da informação, destacando sua relação com a própria escrita.

9.1 Memória e ciência da informação

O conceito de memória, no sentido de preservação cultural, parece estar ganhando importância como tema de pesquisa em Ciência da Informação. Há uma década e meia, Birgen Hjorland já abordava as relações entre documentos, instituições de memória e ciência da informação, ainda que não tenha deixado claro qual exatamente era o conceito de memória a que se referia (HJORLAND, 2000). Naquele trabalho, as bibliotecas são incluídas no rol das instituições de "memória", ao lado das que já são tradicionais em sua relação com a memória cultural: arquivos e museus (HJORLAND, 2000).

Uma década após a abordagem daquele autor e após análise do mesmo tema, conclui-se que a "pesquisa sobre patrimônio cultural é conduzida dentro e nas fronteiras da ciência da informação e biblioteconomia, primariamente em seus campos periféricos (arquivo, patrimônio e estudos museológicos)" (DALBELLO; VAMANU, 2010, p. 2).

Contemporânea a essa citação, uma análise da literatura nacional sobre o conceito de memória em Ciência da Informação também conclui que ela possui um caráter periférico e que a produção científica é pouco expressiva (OLIVEIRA; RODRIGUES, 2011).

No âmbito da Ciência da Informação, pelo menos a produzida nacionalmente, o tema preservação é o mais associado ao tema memória (MONTEIRO; CARELLI; PICKLER, 2006). Ainda que as autoras citadas mencionem as grandes dificuldades para a preservação de conteúdo na internet, contrastando com essa visão, há várias iniciativas que indicam a real possibilidade de tal preservação. Veja-se, por exemplo, o arquivo da web portuguesa (GOMES et al., 2008).

O objetivo aqui é também explorar um dos vértices do tema memória numa perspectiva da Ciência da Informação, especificamente, a relação com o conceito de escrita.

O conceito de documento digital também será abordado em sua relação com a problemática tratada. O conceito documento ganha uma relevância nova em relação ao

conceito informação. De fato, fala-se em preservação de documentos digitais, todavia não se utiliza, em nenhum caso conhecido, o termo preservação da informação. Hjorland também vê essa inclinação da ciência da informação para a ideia de documento em sua relação com o tema memória (HJORLAND, 2000).

9.2 Fatos notáveis sobre a histórica da escrita

Antes de tudo, é importante distinguir entre a evolução histórica da habilidade humana para falar (surgimento da fala) e a evolução da língua e a da escrita, objeto específico desta análise.

A teoria evolucionista prescreve que um dos princípios de evolução é o surgimento de uma mutação em um indivíduo e essa mutação passa para os demais (JANSON, 2015). Ninguém sabe, no entanto, quando surgiu o primeiro "falante". Estima-se, com base em estudos anatômicos e genéticos, que a capacidade humana de falar surgiu entre 90 mil e 50 mil anos atrás (LIEBERMAN, 2007), mas certamente passou por vários estágios de evolução, pois "isso não pode ter acontecido de uma vez: as línguas, provavelmente, se desenvolveram de forma gradual ao longo de vários e vários milhares de anos" (JANSON, 2015, p. 20).

Aqui, a preocupação se limita à evolução da escrita, uma invenção muito mais recente (aproximadamente há 6.000 anos) que o surgimento da capacidade da fala.

Entretanto, é importante notar que o estudo da história das línguas faladas – do ponto de vista da linguística histórica – vem sendo feito após a existência da escrita e com base nos documentos que ela possibilitou que fossem produzidos.

Para alguns autores, a linguística histórica é a história da língua escrita, mas sem a fala não se escreve, pode-se entrever ou entreouvir a voz através dos textos: tarefa difícil e apenas aproximativa, 'ouvir o inaudível' (SILVA, 2008).

A cronologia histórica do desenvolvimento da escrita e do letramento (uso da escrita por indivíduos em grupos sociais) permite uma melhor compreensão dessa que pode ser considerada um elemento de valor essencial para a configuração contemporânea de nossa sociedade e cultura.

É preciso definir o que pode ser considerado como o início do desenvolvimento da escrita até chegar à forma e característica atuais. Para estabelecer um ponto de partida, tome-se esta definição de escrita:

> Uma escrita é um sistema de signos gráficos que remetem aos signos orais emitidos na palavra. Entre os signos gráficos da escrita e os signos orais falados, reina uma correspondência biunívuca que permite, de um lado, representar pela escrita - escrever - todo discurso gerado pela palavra e, de outro, reencontrar de forma idêntica o discurso falado na sua representação escrita – ler. (VANDERMEERSCH, 1995, p. 47).

Note-se o trecho "todo discurso gerado pela palavra". A escrita possui limites e não pode representar todas as ideias, conhecimento e sentimentos humanos, ainda que existam tentativas literárias nesse sentido. A escrita possível, como bem delimita o autor da definição, é a que corresponde ao discurso gerado pela língua.

Os primórdios da escrita surgiram com as primeiras civilizações agregadas em cidades há mais ou menos 6.000 anos, na Mesopotâmia. E aquela escrita tinha uma forma muito diferente da atual.

Um ponto que às vezes gera certa polêmica entre os pesquisadores do assunto é o que exatamente constitui o início da invenção da escrita. Não há nada de divino nela, apesar de essa crença ainda existir entre algumas comunidades atuais (FISCHER, 2003). A essência do conceito atual de escrita é o registro da língua falada, de seus sons, daí às vezes se utilizar o termo *fonografismo*. Já se identificaram exemplos, anteriores às formas identificadas na região onde a invenção parece ter sido criada, na Mesopotâmia, de tentativas de registro de conhecimento por meio de outras formas. Vários grupos humanos fizeram registros em pedras, cavernas, ossos e outros objetos: "antes da escrita completa, a humanidade fez uso de uma variedade de símbolos gráficos e mnemônicos (ferramentas de memória) de vários tipos para armazenar informações" (FISCHER, 2003, p. 14). Até que ponto essas práticas influenciaram na invenção da escrita talvez seja impossível determinar hoje.

Há uma distinção importante em relação às formas humanas de se expressar em geral em contraposição às formas de se comunicar de maneira perene (GELB, 1952). Nessa distinção, os seres humanos são capazes de se expressar naturalmente, sem outros recursos, na forma visual (como gestos), através do tato (tapas nas costas) e sonoramente (língua). Mas essas formas estão limitadas no tempo e espaço. Aquilo que é dito, por exemplo, é dito num determinado local e momento temporal. Mas é possível que as pessoas tenham sentido necessidade de transpor esses limites e, para isso, passaram a registrar uma parte de seu conhecimento, daí o surgimento dos registros humanos.

A tese que defende a necessidade de registro para comunicação perene parece razoável. Contudo, em se tratando de registros humanos e de um ponto de vista documental, importa distinguir entre o que é intencional e não intencional.

Ao manufaturar um machado de pedra, a intenção dessa tarefa deve ter sido utilizar tal ferramenta para algum tipo de corte. Ao mesmo tempo, ao fazer isso, não intencionalmente, registravam-se informações sobre como aquele ser vivia e, de certa forma, sobre o conhecimento técnico já acumulado naquele momento. Esse segundo "uso" da ferramenta é resultado de nossa perspectiva cultural atual. Da mesma forma, ao desenhar um animal qualquer numa parede de caverna, é razoável considerar que o objetivo tenha sido de culto ou algum tipo de superstição, mas é menos provável que tenha sido para, intencionalmente, registrar outro ser que coabitava naquela região e muito menos com o intuito de comunicar isso às pessoas no futuro.

A escrita, no sentido atual e desde seu primeiro uso, pressupõe a intenção de registrar elementos do conhecimento humano, pelo menos os expressos pela língua, de maneira a ultrapassar os limites espaço-temporais da língua falada: um auxílio à memória humana comum. Até que outros sistemas de escrita semelhantes àqueles da Mesopotâmia sejam arqueologicamente descobertos em outros sítios, se é que serão um dia, o *fonografismo* e a intenção de registro reforçam que foi naquela região que se deu o início da escrita.

O maior aprimoramento na antiguidade que levou ao desenvolvimento pleno da escrita foi a invenção rudimentar e parcial da fonetização, também na Mesopotâmia (FISCHER, 2003). A fonetização significa a técnica de registrar o som da língua num material qualquer, ainda que no início figuras e grafismos coexistissem com os registros fonéticos. Naquele momento histórico, foram utilizadas tabuletas de argila como suportes de escrita, material abundante na região, as quais podem ser observadas até hoje em museus.

Alguns autores preferem usar a nomenclatura glotofonia (SAMPSON, 1996) e subdividir os sistemas em semasiográficos (indicam ideias de maneira direta) e os glotográficos (representação da língua). Esses últimos, por sua vez, são subdivididos em fonográficos propriamente (registro de sílabas) e logográficos (registros de palavras). Mas trata-se de uma subdivisão que reflete apenas uma classificação preferencial de alguns autores.

O ponto realmente importante é o *princípio fonográfico* que leva ao *fonografismo e fonetização*. Desde seu surgimento, sucessivos aprimoramentos foram sendo introduzidos até os sistemas atuais. No começo, havia sistemas mistos que registravam sons da língua e figuras representando coisas concretas. Depois, passou-se ao registro apenas das sílabas utilizadas no vernáculo local. A grande melhora foi introduzida na Grécia (900 AC, há menos de 3.000 anos, portanto) com o registro gráfico das vogais e a invenção dos primeiros alfabetos, antecessores diretos do atual modelo ocidental, "se, pela palavra alfabeto nós

entendemos uma escrita que expressa os sons individuais da língua, então o primeiro foi formado pelos gregos" (GELB, 1952, p. 197).

Nesse ponto do processo histórico, cabe notar outro aspecto da escrita em sua relação com a sociedade. Desde seu surgimento mais rudimentar, se se comparar com os sistemas atuais de escrita, sua prática era delegada a profissionais, chamados de escribas. Esses "técnicos da escrita" atuaram nas esferas do governo, religião e o que se chama hoje de "escola" (COULMAS, 2014).

No ocidente, essa situação só começa a se alterar a partir do Renascimento europeu e o surgimento do movimento humanista que começou a se opor aos escolásticos que dominavam as universidades (BURKE, 2003).

Mais adiante, com o surgimento da ciência moderna, fica clara a importância da escrita.

> Nenhuma cultura de tradição oral jamais conseguiu, até hoje, desenvolver ciência verdadeira: os saberes de alto padrão derivam todos de ambientes dotados da escrita e capazes, graças a ele, de construir sistemas de conhecimento extensos, precisos, controlados e sistematizados e, além disso, ampliáveis e aperfeiçoáveis por uma classe mais ou menos prolongada de competências. (BOTTÈRO, 1995, p. 22-23).

Já com relação ao uso pleno da escrita pelas pessoas, mesmo na contemporaneidade, o letramento, ainda que amplamente difundido, não é pleno para todas as classes sociais:

> O letramento é, em si mesmo, uma característica definidora de classe social. O letramento é um instrumento de poder social. As pessoas se tornam parte de uma cultura ao aprender a interpretar e usar seus signos e símbolos particulares. Usam a língua em relações sociais que elevam seu conhecimento e desenvolvem seu potencial. Habilidades letradas reduzidas podem excluir as pessoas dos grupos sociais dominantes e das oportunidades numa sociedade. (COULMAS, 2014, p. 89, citando o *Movement for Canadian Literacy*).

A seguir, são abordadas as relações e diferenças da escrita com a língua.

9.3 A escrita e a língua

As línguas faladas e as línguas escritas[1] são tão interconectadas que se corre o risco de não as diferenciar. Nesta pesquisa, no entanto, é importante destacar as diferenças

[1]Alguns especialistas em linguística preferem o termo escrituras, mas não parece haver unanimidade sobre isso.

entre língua e escrita[2] e elencar os pontos que permitirão uma clara distinção entre ambas.

Um dos primeiros trabalhos dedicados à língua escrita (GELB, 1952) enxerga três características contrastantes entre a língua e a escrita. A primeira é que a escrita é mais conservadora, ou seja, suas regras e formas resistem às mudanças linguísticas bem mais que a língua falada. A segunda é que a escrita acaba se tornando um elemento restritivo às mudanças nas línguas faladas, freando sua evolução. Aquele autor cita exemplos de línguas ágrafas, que não só tiveram mudanças consideráveis entre duas gerações seguintes de falantes, mas também que originaram dialetos e novas línguas (GELB, 1952). No entanto, mesmo nas sociedades que dominam a escrita, a língua sofre mudanças. A terceira é que a escrita preserva as formas antigas de uma língua qualquer na forma de "documentos históricos".

Não é difícil confundir a escrita com a língua, pois a primeira se esforça para reproduzir a última.

Para Coulmas, “nenhum trecho de escrita é uma versão fiel da fala. Ela desconsidera o ritmo, a variação de volume, a entonação, a altura do som e outros aspectos articulatórios, para não mencionar a mímica e os gestos” (COULMAS, 2014, p. 72).

Contrastando com a ideia de que a escrita é simplesmente um reflexo da língua falada, algumas culturas têm, na versão escrita de seu vernáculo, diferenças gritantes. O fenômeno da diglossia[3] refere-se ao fato de que, em algumas culturas, inclusive na contemporaneidade (a língua árabe, por exemplo), as diferenças entre a língua falada e a escrita são tão diferentes que parecem duas línguas distintas (LEIKIN; IBRAHIM; EGHBARITH, 2014).

Chamam a atenção também casos como o de Moçambique no continente africano, país onde a língua oficial é estrangeira: o português, considerado língua materna por apenas 10% da população. Nessa língua oficial, são publicados (escritos) os documentos legais e administrativos (COULMAS, 2014), donde se pode concluir, paradoxalmente, que os documentos de Moçambique não são redigidos na língua (falada) do país, que possui vários dialetos.

---

[2]Agradeço à Profa. Dulce Baptista por chamar a atenção para a oposição língua e escrita, que temos feito ao longo desta tese em detrimento de fala e escrita. A origem de nossa opção terminológica se deve ao uso linguístico desses termos, pois fala é um termo limitante em alcance e refere-se, do ponto de vista linguístico, aos processos mecânicos de articulação e produção de som. A alternativa língua em oposição a escrita se mostra menos problemática, ainda que para alguns o termo língua incorpore fala e língua escrita.

[3]DIGLOSSIA é uma situação relativamente estável da linguagem na qual, em adição aos dialetos primários da linguagem (que podem incluir padrão ou padrões regionais), há uma variedade divergente, altamente codificada (frequentemente gramaticalmente mais complexa) sobreposta. É um veículo da larga e respeitada literatura escrita, tanto de um período anterior ou de outra comunidade de fala a qual é aprendida largamente pela educação formal e usada para a maioria das aplicações escritas e faladas, mas não é utilizada por nenhuma seção da comunidade para a conversação ordinária (BRUNELLE, 2008).

Outro aspecto de comparação entre a língua e a escrita, do ponto de vista da linguística, é a importância relativa de ambas como objetos de estudo. A situação parece ter começado a se modificar nas últimas décadas, mas a linguística parece simplesmente ignorar a escrita, como pode ser verificado nesta citação de quase meio século atrás: "a língua é basicamente expressão oral, e a escrita não apresenta qualquer interesse teórico" (BLOOMFIELD, 1969, p. 886 apud SAMPSON, 1996, p. 7).

Sampson discute as razões dessa visão da linguística, que, para ele, é equivocada e propõe o estudo da escrita do ponto de vista histórico, tipológico e psicológico (SAMPSON, 1996). O filósofo francês Derrida também se interessou pelo estudo da escrita (que na tradução brasileira vem como escritura), chegando a propor uma ciência para a escrita (Gramatologia),

> A ciência da escritura deveria, portanto, ir buscar seu objeto na raiz da cientificidade. A história da escritura deveria voltar-se para a origem da historicidade. Ciência da possibilidade da ciência? Ciência da ciência que não mais teria a forma da lógica mas sim da gramática? (DERRIDA, 2011, p. 35).

Mais adiante, discute-se a possibilidade dos estudos sobre a escrita do ponto de vista da relação memória x ciência da informação.

9.4 Escrita e documento de arquivo

A escrita é uma ferramenta com múltiplas aplicações que vão desde áreas como a literatura artística e científica até manuais profissionais ou não. De qualquer forma, é um elemento fundamental para o auxílio ao registro da memória individual e coletiva,

> Nós nos armamos contra a transitoriedade implícita na mortalidade da memória por meio da criação de memórias artificiais. O mais antigo auxílio à memória é a escrita; na antiguidade em argila ou placas de cera, na idade média em pergaminho e velino e, mais tarde, em papel. (DRAAISMA, 2005, p. 21).

Nessa seção, são exploradas suas relações com o registro do conhecimento humano para a manutenção da memória social a longo prazo através do conceito de documento de arquivo.

Com relação ao tipo fundamental de conteúdo, os documentos em geral e também aqueles considerados de arquivo podem ser classificados entre imagens fixas ou em movimento, som e textos. Há várias combinações possíveis a partir dessa classificação,

principalmente se for considerado o universo dos documentos digitais. Em qualquer caso, os documentos textuais ou basicamente textuais, cuja mensagem pode ser transmitida por uma língua na forma escrita, ainda são os mais comuns. E isso é facilmente observável no mundo real.

Em terminologia arquivística um "documento textual" é um gênero documental "reunião de espécies documentais que se assemelham por seus caracteres essenciais, particularmente o suporte e o formato, e que exigem processamento técnico específico e, por vezes, mediação técnica para acesso" (ARQUIVO NACIONAL, 2005, p. 99). Documentos incluídos no gênero "documento textual", ainda de acordo com a terminologia arquivística, são "manuscritos, datilografados ou impressos, como atas de reunião, cartas, decretos, livros de registro, panfletos e relatórios" (ARQUIVO NACIONAL, 2005, p.79).

Os documentos textuais são os mais comuns quando o objetivo é utilizá-los para análises históricas "A moderna ciência histórica, baseada na crítica factual do documento escrito, surgiu, justamente, como resultado da ação de classicistas e estabeleceu os termos da análise textual tradicional" (FUNARI, 2003, p. 15). A supremacia dos documentos textuais históricos hoje é função do fato de que os documentos com imagens (fotografia), cinema e o som começaram a aparecer apenas no final do século XIX, na esteira do desenvolvimento tecnológico ocorrido naquele período.

Normalmente, os documentos textuais históricos são objeto de estudo da disciplina história, com a arquivologia atuando na tarefa de sua organização, preservação e disponibilização, mas a memória associada aos documentos históricos é abordada por várias disciplinas: "com efeito, essas três áreas (arquivologia, biblioteconomia e museologia) valem-se da memória no sentido de armazenagem e preservação dos saberes (conservação), para a posterior recordação por parte da sociedade" (MONTEIRO; CARELLI; PICLKER, 2006, p. 115).

A língua escrita é limitada em relação à língua, todavia registra elementos suficientes para análises da cultura e vida social de uma sociedade. Na medida em que se tornou mais sofisticada e utilizada na sociedade (letramento), a escrita cada vez é mais útil nessa tarefa de reconstituição da história:

> Todas as línguas mudam e se desenvolvem no curso do tempo, mas só as línguas escritas carregam os registros de seu próprio passado, registros que podem ser inspecionados, referidos como exemplos, idealizados, citados literalmente, falsificados, canonizados, condenados como "tirania", traduzidos. (COULMAS, 2014, p. 41).

O documento de arquivo, especificamente em relação aos demais tipos de documentos, possui uma característica única quanto à preservação da memória[4]. É sua relação com os demais documentos de arquivo que compõem um determinado arquivo. De fato, a definição de documento de arquivo pode simplesmente remeter ao conceito arquivo[5], de tal maneira que o resultado total para a preservação da memória é uma sinergia de um conjunto documental (arquivo). Nessa linha de pensamento,

> O documento arquivístico é um artefato humano com pressupostos e características específicas. O ambiente e o conteúdo são delimitados e definidos pelo sujeito acumulador, que pode ser uma pessoa física ou jurídica. Então quando falamos de arquivo, estamos nos referindo a um conjunto finito de documentos acumulados, que tem suas fronteiras demarcadas pela missão do criador, no caso das instituições, e pela área de atuação, no caso das pessoas físicas. Ao contrário daqueles encontrados em bibliotecas, por exemplo, os documentos arquivísticos não constituem um conjunto formado em vista de uma finalidade específica: eles representam, mais que tudo, o produto da atividade do sujeito criador. (SOUSA, 2007, p. 113).

Cabe aqui uma distinção importante para os documentos de arquivo. O que se denomina no Brasil documento de arquivo corrente a literatura em inglês chama de *records*. Da mesma forma, o documento de arquivo de guarda permanente é conhecido em inglês por *archive document*. Essas distinções são importantes, pois, em qualquer idioma, a teoria identifica fases do documento de arquivo, desde sua produção (fase corrente) até sua guarda definitiva (fase permanente).

Apesar de polêmico, do ponto de vista de alguns autores que consideram que TODO documento pode conter ou ajudar a manter a memória cultural, considera-se que os documentos de guarda permanente são os "mais importantes" para a função de preservar a memória. Ao longo desta pesquisa e de seu escopo, quando se trata de documentos de arquivo, consideram-se especificamente aqueles de guarda permanente ou *Archive Document*.

A escrita não é usada apenas nos documentos textuais. Ela também é utilizada para registrar informações sobre documentos, textuais ou não. Mais adiante, será abordada em detalhes esta situação: a representação das informações registradas em documentos. Por ora, é importante chamar a atenção para esse fato, uma vez que a escrita acerca das informações registradas será fundamental para o problema de pesquisa.

---

[4]No caso brasileiro e também no de vários outros países, a função de ajudar a preservar a memória cultural é tão evidente para documentos de arquivo que há legislação nacional apoiando esse papel. Vide coletânea de legislação de arquivo disponível no sítio do Arquivo Nacional brasileiro.

[5]De fato, é assim que ocorre, por exemplo, no Dicionário Brasileiro de Terminologia Arquivística (ASSOCIAÇÃO DOS ARQUIVISTAS BRASILEIROS, 1996).

No caso do documento de arquivo, a representação dos documentos e de seus "conjuntos documentais" ocorre através de catálogos, inventários e guias (AGUIAR, 2008). Talvez a característica mais diferenciadora desses "produtos de representação" seja o objetivo de contextualizar os documentos em relação às instituições que os produziram. Sobre isto:

> A descrição arquivística é o processo em que o arquivista cria representações de um determinado acervo arquivístico, apresentando seu contexto e conteúdo. É uma atividade intelectual que demanda competências de interpretação de texto, conhecimento histórico e habilidade para redigir descrições dos acervos. O objetivo é o controle dos documentos arquivísticos, tendo em vista a promoção do acesso. (SILVA; ORRICO, 2013, p. 211).

Atualmente, as atividades de descrição arquivística são regradas por normas internacionais, as quais no caso brasileiro possuem equivalentes publicados por nosso Arquivo Nacional[6]. A norma de descrição arquivística mais conhecida é a Norma Geral Internacional de Descrição Arquivística, mais conhecida pela sigla original (ISAD-G). Uma característica desta norma é a descrição multinível, que significa a descrição do todo de um acervo de documentos de arquivo em relação a uma instituição específica (fundo) e abaixo deste nível outros tantos quanto se exigir e for possível exaurir a descrição desse referido fundo. Há várias regras, obrigatórias ou não, áreas e campos a serem seguidos a fim de obter padronização internacional e melhores condições de acesso aos documentos de arquivo descritos,

> Elementos de informação específicos sobre documentos de arquivo são registrados em cada fase de sua gestão (por exemplo, criação, avaliação, registro de entrada, conservação, arranjo) se tais documentos devem, por um lado, ser preservados e controlados com segurança e, por outro, ser acessíveis no tempo oportuno a todos que tenham o direito de consultá-los. A descrição arquivística no sentido mais amplo do termo abrange todo elemento de informação, não importando em que estágio de gestão ele é identificado ou estabelecido. (CONSELHO INTERNACIONAL DE ARQUIVOS, 2000, p. 11).

No entanto, a descrição arquivística que interessa é aquela aplicada a documentos de guarda permanente, que, do ponto de vista das normas de descrição ISAD(G), é a mesma descrição aplicada a qualquer documento de arquivo, independentemente de sua avaliação como permanente ou não. Outras normas também complementam a ISAD(G) com aquelas aplicáveis à descrição de pessoas e autoridades.

---

[6]No endereço eletrônico específico para publicações do Arquivo Nacional, pode-se ter acesso gratuito a todas as normas internacionais (em português) para descrição arquivística <http://www.conarq.arquivonacional.gov.br/publicacoes_.html#>.

A descrição de documentos de arquivo também pode se harmonizar e ser tratada juntamente com outros tipos de documentos (de outras áreas) como biblioteca e museu. O projeto de organização da história da energia elétrica no Estado de São Paulo (com documentos do final do século XIX a meados do século XX) ilustra esta possibilidade. O projeto também utilizou princípios das normas internacionais de descrição arquivística (LIMA; VITORIANO; BARBANTI, 2015).

Ainda se falará sobre o documento de arquivo e sua descrição na seção que trata de representação da informação, na revisão de literatura.

9.5 Escrita e documento digital

A escrita não está presente em um documento digital textual de maneira superficialmente diferente em relação a um documento em suporte tradicional, mas a complexidade em um documento digital é maior em vários aspectos,

> O conceito tradicional de documento - todo o suporte material da informação - deve ser revisto, uma vez que não encontraremos seu sentido e seu significado tomando, apenas, sua forma e seu potencial informativo, sem considerar a interlocução e, mais especificamente, a intenção de preservação no âmbito da memória social. (DODEBEI, 2011, online).

No documento tradicional textual, o conhecimento humano é registrado com o emprego de uma língua. De certa forma, a língua utilizada e as informações registradas se confundem. No entanto, no caso do documento digital, além da língua utilizada, também há a linguagem binária. Esta é uma língua artificial, compreensível por máquinas. O que é realmente gravado o é nessa língua.

A língua humana em um documento digital é obtida através de recursos de *software* e *hardware*. No fundo, a língua escrita em um documento digital não é fisicamente gravada. Esse fato pode levar a informações interessantes se se considerar que a essência de uma língua escrita é seu registro material em termos de fonetização. Além disso, o que é realmente gravado (a linguagem binária) é comum para qualquer tipo de conteúdo documental além do texto: imagem, som e outras elementos como páginas da internet, planilhas ou bancos de dados. De certa forma, a linguagem binária é um novo "alfabeto" em que "qualquer outro alfabeto pode ser representado e processado e no qual somos capazes de representar conhecimento expresso em qualquer formato utilizado antes na história das sociedades modernas" (FINNEMANN, 1999, p. 14).

De fato, o documento digital e sua complexidade frente aos documentos tradicionais significam uma realidade tecnológica completamente distinta e nova. Uma metáfora interessante é a que se refere a uma nova galáxia, à galáxia de Gutenberg (a cultura impressa), a antiga, versus a galáxia de Turing e a mídia digital (FINNEMANN, 1999).

O uso cada vez mais intenso do documento digital foi potencializado por outra invenção tão ou mais revolucionária que ele: a internet. As noções de escrita e texto chegaram mesmo a receber novas nomenclaturas. Assim, fala-se em hypertexto ou cybertexto (GUNDER, 2001). Não parece se tratar apenas de novos nomes, mas também de novas características, como a fixidez de informação:

> O computador está questionando a ideia de fixidez: no lugar do texto impresso fixo e estável, o computador nos oferece um texto fluido e interativo. O computador promete, portanto, reverter as qualidades que Eisentein identificou na revolução da imprensa. (BOLTER apud DALGAARD, 2001, p. 2).

Em relação à representação de informações, os registros digitais também trouxeram novas tecnologias e desafios. O termo metadados se refere a algo já bastante tradicional em Ciência da Informação, ou seja: o que se refere a outros documentos. No mundo digital, seus equivalentes ganham novos contornos ao se tratar de padrões e aplicações específicas em documentos digitais. É possível fazer uma crítica sobre a "reinvenção da roda" para muitas das "novas tecnologias" (LANCASTER, 2004). Ao lado de muitas inovações, há também muitas repetições:

> (...) o que pensamos que sejam inovações muitas vezes são meras repetições (...) nossa profissão pode desenvolver-se de modo mais rápido e melhor por meio de inovações cumulativas, construindo sobre os alicerces de seu passado ao invés de ignorá-lo. (HOLMES, 2001 apud LANCASTER, 2004, p. 11).

Com isso, encerra-se a breve revisão sobre a escrita e sua relação com o problema de pesquisa. A seguir, trata-se da representação, recuperação e sistemas de informação.

9.6 Considerações finais desta seção

Esta seção tratou principalmente do conceito de escrita, em função de sua importância no objeto desta pesquisa. É essencial sua demarcação em relação ao conceito de língua falada. Estabeleceu-se também sua relação com o problema da memória, do documento de arquivo e do documento digital, pois são essas as relações específicas tratadas ao longo do trabalho.

10 REPRESENTAÇÃO E RECUPERAÇÃO DA INFORMAÇÃO

Língua(gem), linguística e CI possuem relações importantes. Isso se revela claramente na produção brasileira de CI. O relatório[1] "O Grupo Temma na ECA-USP, 2001-2011" é muito esclarecedor sobre essa relação, se analisarmos a produção científica ali registrada.

Nos dez anos examinados naquele relatório, entre livros, capítulos de livros, artigos e trabalhos apresentados pelas pesquisadoras, aparecem vários aspectos diretamente ligados à língua(gem). Abundam termos como "linguagem verbal", "linguística aplicada", "linguística", "terminologia", "linguagem", "função comunicativa", "vocabulário" e "língua portuguesa". Além disso, frequentemente, esses termos são associados à representação e recuperação da informação. Destaco que o tema memória e recuperação da informação também aparece em produtos daquele grupo, como em *Documentation: la mémoire et les systèmes de recherche d'information* (SMIT; TÁLAMO, 2006).

Na verdade, parece impossível que se desenvolvam pesquisas em Ciência da Informação sem considerar algum aspecto do conceito língua(gem), visto que a língua (escrita ou falada) está necessariamente presente em todo documento registrado ou no processo de produção e é um componente essencial no que é comunicado através desses documentos. Essa presença também se mostra evidente pela produção dos diferentes PPGCIs em todo o Brasil[2].

Nesse cenário, a Linguística parece desempenhar um papel de destaque já que, desde o início da década de 80 no Brasil, já há trabalhos relacionando Ciência da Informação e Linguística, como em CINTRA, 1983. Além disso, na introdução de um livro sobre Linguística e o profissional da informação, encontra-se esta citação esclarecedora:

> A Ciência da Informação preocupa-se com os aspectos da geração, comunicação e uso da informação. Considerando-se a linguagem como o principal veículo de comunicação e sendo a Linguística a ciência que estuda a linguagem enquanto sistema de comunicação e auto expressão, a relação entre as duas ciências torna-se clara e indiscutível, como destacam vários autores da área. (MELO; BRÄSCHER, 2011, p. 13).

Nesta pesquisa, é explorado o recorte conceitual sobre a representação e recuperação da informação a partir de três pilares: o *documento*, a *língua(gem)* e a

[1]Veja no diretório de grupos do CNPQ: <dgp.cnpq.br/dgp/espelhogrupo/6321322781228728>.

[2]A título de indicação, efetuamos uma busca simples nos ANAIS do XV Enancib (2014). A busca pelos termos língua e linguística apresentou dezenas de resultados quando pesquisamos os ANAIS do GT1 - Estudos Históricos e Epistemológicos da Ciência da Informação, GT2 - Organização e Representação do Conhecimento e GT10- Informação e Memória.

informação. De fato, esse tripé parece fundamental para uma parte considerável dos problemas em Ciência da Informação. Ainda que este trabalho se interesse especificamente pela (inter)relação entre eles e a representação e recuperação das informações registradas em documentos.

Documento é o elemento mais concreto nesse tripé. Trata-se de algo que pode ser materialmente explorado. Tem peso, dimensões e características físico-químicas. Isso tanto é válido para documentos digitais quanto para documentos em suportes tradicionais (como o papel, papiro ou vinil) ainda que o problema da materialidade do documento digital seja mais complexo do que o dos documentos em suportes tradicionais. Para este último, a possibilidade de exibição em vários suportes de maneira permanente ou não e a necessidade de plataformas de *hardware* e *software* adicionam muitas variáveis importantes (ROSS, 2012).

Informação, por sua vez, é um aspecto bem mais abstrato e difícil, se comparado ao aspecto "documento". Retomando a teoria da informação, nos moldes definidos originalmente por Shannon[3], há um transmissor A, um receptor B, um canal e o que é transmitido, a coisa transmitida (SHANNON, 2001). Essa coisa não é o "documento" em si, ainda que possa conter uma cópia do conteúdo de um documento ou referências a um documento, em qualquer caso agregado com outros dados específicos naquela transmissão. Também não é a *língua(gem) ou Língua* utilizada naqueles documentos. Desta forma, temos de falar em outra instância presente nessa transmissão, para além do "Documento". Enfim, é isso o que chamamos aqui de forma ampla de "Informação".

Língua(gem), finalmente, é o último aspecto dessa tríade. Em se tratando de Documentos com conteúdo passíveis de processos de significação (inter)humanos, a Língua estará presente necessariamente. Por isso, podemos afirmar que a Língua e a Escrita são os dois aspectos da língua(gem) humana, como abordados conceitualmente nas seções anteriores que podem ser considerados como elementos essenciais tanto nos processos de representação (como efetuados em CI), como também em recuperação da informação (nos moldes da Ciência da Computação).

Considerando o conceito de íngua, escrita e mudança linguística nos respectivos aspectos conceituais destacados e discutidos na revisão conceitual levada a cabo nas seções anteriores, pergunta-se:

[3]A teoria da informação nos moldes propostos por Shannon em 1948 é uma abordagem matemática do sinal transmitido. Ainda hoje é abordado justamente em engenharia de comunicações. Aqui é útil para exemplificar os contrastes entre documento, informação e língua(gem).

Como a representação e a recuperação da informação registrada em documentos de arquivo de guarda permanente relacionam-se com aqueles conceitos?

Além disso, como (do ponto de vista linguístico) as pessoas se relacionam com aqueles processos, quais papéis assumem (produtor de documentos e informações, pesquisador de documentos e informações e gestor (indexador, classificador, catalogador) de documentos e informações) e como se inserem na problemática geral? Para responder a essas perguntas, é preciso analisar mais detalhadamente cada um dos processos, representação e recuperação da informação.

10.1 Representação da informação

A perspectiva da Ciência da Informação trata de vários processos relacionados à informação registrada que incluem produção documental e organização até o acesso às informações (ALVARENGA, 2003).

Interessa aqui, especificamente, o processo de representar as informações registradas em documentos: "A principal característica do processo de representação da informação é a substituição de uma entidade linguística longa e complexa – o texto do documento – por sua descrição abreviada" (NOVELLINO, 1996, p. 38). Ainda que a citação se refira ao "texto do documento", esse processo também se aplica à imagem, som e combinações possíveis, dentro de suas especificidades e limites.

Em termos mais concretos, o processo de representação da informação se materializa em instrumentos de representação associados a cada tipo de processo. Como ilustração, para o caso de um vocabulário controlado: “controle de vocabulário igual a processo, um objetivo que se deseja atingir, vocabulário controlado igual a instrumento para nomear as atividades/funções, gerando confiança no sistema” (SMIT; KOBASHI, 2003, p. 20).

A ideia de "processo" é fundamental. Ela perpassa definições de vários autores e pode ser estendida para a representação da informação de maneira generalizada, assim como foi feito para o caso dos vocabulários controlados na citação acima. Processo pode ser definido como: "1. Procedimento, maneira de operar ou de agir, para indicar o método que consiste em ir das causas ao efeito, ou do efeito às causas" ou "2. Devir ou desenvolvimento (...) por exemplo para designar a formação do mundo" ou "Concatenação qualquer de eventos, por exemplo, o processo digestivo ou o processo químico" (ABBAGNANO, 2007, p. 936). É

principalmente no último sentido que o termo é aplicado: processo de representação da informação e "concatenação de eventos".

É importante notar que o termo "processo" também serve para denominar várias etapas na representação da informação. Assim, pode-se falar no processo de indexação, no processo de elaboração de vocabulário controlado. Do prisma do processo de representação da informação, esses processos podem ser considerados subprocessos.

O efeito principal no processo de representação da informação é obter o "estar em lugar de, isto é, estar numa tal relação com outro que, para certos propósitos, é compreendido por alguma mente como se fosse a outra coisa" (PEIRCE, 1977, p. 61 apud CERVANTES, 2009, p. 28). Para obter esse resultado, do ponto de vista da Ciência da Informação, é fundamental o uso de "instrumentos", as ferramentas que permitem executar o processo de maneira satisfatória.

Entre os instrumentos mais comuns atualmente usados estão o tesauro, o vocabulário controlado, as terminologias, as socioterminologias e as ontologias. É possível chamar esses "instrumentos" genericamente de linguagens documentárias: "Esses instrumentos são denominados, de uma forma geral, linguagens documentárias, como o tesauro[...]" (CAMPOS, 2001, p. 21). Em outros termos, "a representação documentária de um texto é formulada em uma linguagem que não se confunde com a linguagem do texto, mesmo que os termos tenham aparentemente a mesma forma" (KOBASHI, 2007, p. 7).

Além dos instrumentos utilizados no processo, também há o papel dos produtos específicos no mesmo processo. Esse é um ponto que parece poder suscitar alguma confusão, pois as diferenças entre um "instrumento" e um "produto" podem ser sutis: “os níveis básicos da descrição arquivística, configuram-se através dos produtos de representação da informação arquivística, conhecidos como catálogos, inventários, guias e inventários” (AGUIAR, 2008, p. 169, grifo nosso).

Se um vocabulário controlado qualquer (impresso ou digital) é um instrumento da representação, por que um catálogo de arquivo (impresso ou digital) também não o é? Já que ambos – um catálogo e um vocabulário controlado – foram obtidos a partir de um processo de trabalho com o objetivo de representar informações registradas?

Uma tentativa de diferenciação é aqui assinalada:

> a) processos: compreendem a análise, a condensação e a representação; b) produtos: a transformação dos documentos em produtos que facilitam a consulta aos originais, em áreas especializadas do conhecimento, como os índices, os resumos, catálogos impressos e catálogos de acesso público online - OPAC's, por exemplo; e c) instrumentos: ferramentas de linguagem documental para a representação

> padronizada do conteúdo temático de documentos, fazendo uso acentuado das tecnologias de informação e comunicação, tais como classificações, cabeçalhos de assuntos, tesauros, terminologias, ontologias, etc. (GUIMARÃES, 2008, p. 84 apud MARTINS, 2014b, p. 100).

O entendimento, com o intuito de estabelecer um referencial conceitual, pelo menos operativo, nesta pesquisa é que a representação da informação é um processo e compreende um conjunto de etapas de trabalho (entre as quais estão os instrumentos de representação) que se inicia a partir dos documentos a serem tratados e tem o objetivo principal de entregar produtos que representam esses mesmos documentos substituindo-os e facilitando sua busca e recuperação no momento oportuno por usuários.

No caso, interessa destacar e explorar o processo de representação da informação, em seus instrumentos e produtos, o qual sempre tem um elemento comum e necessariamente presente: a escrita. E, para os objetivos de pesquisa, essa é a característica importante a ser explorada mais detidamente.

A escrita é essencial nesta pesquisa por várias razões. Primeiro, porque sempre está presente quer no conteúdo dos documentos (em princípio apenas nos textuais, pois os imagéticos e sonoros normalmente não possuem escrita, pelo menos não no mesmo sentido dos documentos textuais) quer nos produtos e instrumentos que representam os documentos (e nesse caso há escrita para todos os tipos de documentos: textuais, imagéticos, sonoros e suas combinações). Para estabelecer diferenças e não criar confusões, chamar-se-á de $E_{conteúdo}$ a escrita presente no conteúdo de documentos e de $E_{representa}$ a escrita presente nos instrumentos e produtos da representação.

Assim, um acervo de documentos que tenha passado por um processo de representação de suas informações sempre possui manifestações de uma língua qualquer em sua forma escrita, pelo menos em um dos dois casos acima.

$$\text{Acervo de documentos possui} \rightarrow E_{representa} \text{ e/ou } E_{conteúdo}$$

Exemplo:

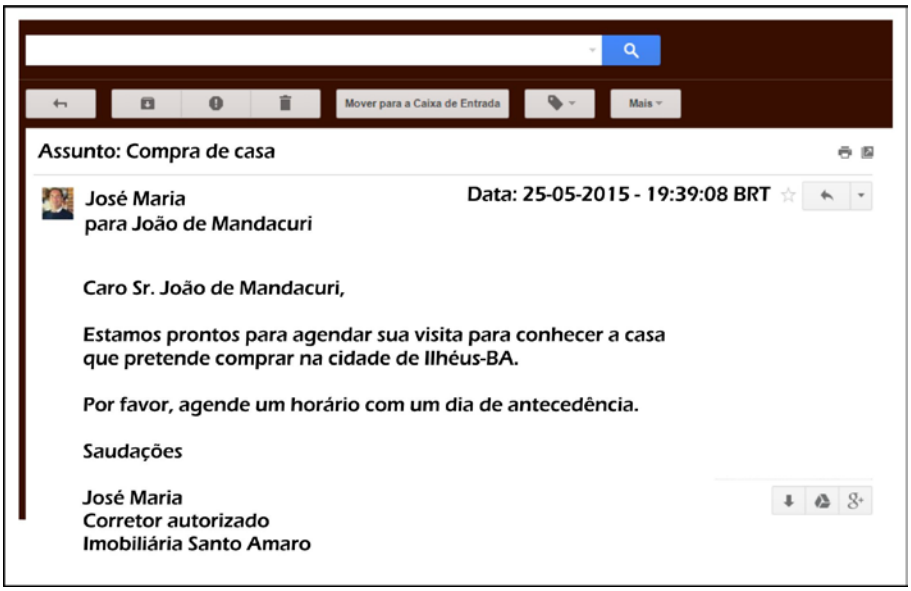
Para ilustração, suponha-se um documento textual (e-mail) com a configuração da *figura 6*. Nesse documento a $E_{conteúdo}$ corresponde a elementos como "compra de casa", "visita", "Ilhéus", "imobiliária", "corretor", coisas concretas ou abstratas. Verbos: "pretender", "agendar", "conhecer" e outros elementos como um formato de data "25-05-2015", artigos, preposições, nomes próprios. Enfim, vários elementos do léxico da língua portuguesa.

A escrita no conteúdo de um documento é constituída pela somatória dos

elementos linguísticos presentes no documento considerado:

$\mathbf{E_{conteúdo}} = \sum \mathbf{e_1}$ **a** $\mathbf{e_n}$, onde de 1 a n são os elementos linguísticos registrados no documento considerado.

Figura 6 – Documento digital e-mail, adaptado de imagem real (dados fictícios)



Fonte: Elaboração própria a partir da edição de tela do aplicativo G-Mail.

Esse mesmo documento pode ser representado por meio de vários instrumentos e produtos do processo de representação. Um trecho de uma ontologia que pode representar algumas informações sobre esse mesmo documento (em linguagem de marcação XML) poderia se apresentar como na *figura 7*.

Figura 7 – Exemplo hipotético de trecho em linguagem XML de uma ontologia

```
<owl: Class rdf:ID="e-mail"
<rdfs:label xml:lang="pt">"Mensagem Eletrônica"</rdfs:label>
<rdfs:subClassOf rdf:resource="Documento_digital"/>
</owl:Class>
<owl: Class rdf:ID="subject"
<rdfs:label xml:lang="pt">"compra de casa"</rdfs:label>
<rdfs:subClassOf rdf:resource="e-mail"/>
</owl:Class>
```

Fonte: Elaboração própria.

Uma linguagem artificial estruturada como a XML é adequada para uso

computacional, pois todos os elementos são colocados em posições muito específicas e permitem o reconhecimento de seu papel no todo. Por outro lado, há muito mais elementos apenas para permitir essa estruturação da linguagem. De qualquer forma, pode-se notar (em vermelho na *figura 7*) que o assunto da mensagem eletrônica em questão, o elemento linguístico "compra de casa", está presente.

A escrita, na representação de um documento, é, então, constituída pela somatória dos elementos linguísticos presentes nos instrumentos ou produtos utilizados para representar o documento considerado:

$\mathbf{E_{representa} = \sum e_1 \text{ a } e_n}$, onde de 1 a n são os elementos linguísticos registrados no instrumento e/ou produto considerado.

Se fosse tomada como exemplo uma fotografia sem qualquer tipo de legenda ou anotações (sem nenhum texto, portanto), só seria possível a existência de elementos em $E_{representa}$ para esse caso específico. Isso após algum processo de representação daquele documento fotográfico.

Dependendo do instrumento ou produto do processo de representação utilizado, os elementos derivados do processo poderiam ser tratados em suas relações semânticas, por exemplo, sinonímia ou homonímia, no caso de um tesauro típico.

Para alguns contextos de Tecnologia da Informação (TI) e representação, os termos nos documentos presentes no acervo podem ser todos agrupados e recebem uma pontuação de importância que será utilizada na etapa de "recuperação da informação". Para esses casos, também conhecidos como *bag-of-words*, "a lista de termos é extraída do conjunto total de documentos. Os documentos são expressos na forma vetorial onde cada coordenada está associada a um termo que representa uma característica do documento" (BOTTÉRO, 1995, p. 19). Esses termos também podem corresponder aos elementos linguísticos de $E_{representa}$, como deve ficar mais claro na posterior análise sobre recuperação.

Considerando que esta pesquisa lida com documentos de guarda permanente, é importante notar algumas questões conceituais sobre representação de informações registradas em documentos de arquivo.

Vários fatores acarretaram a falta de integração das disciplinas básicas de tratamento documental (biblioteconomia, documentação e arquivologia), talvez o principal deles seja a comunicação entre as áreas (ESTEBAN NAVARRO, 1995). O fato é que cada

área e seus documentos típicos possuem suas peculiaridades, mas, essencialmente, o princípio de representar os documentos e gerar produtos representantes permanece tanto para documentos de arquivo como para os de biblioteca, por exemplo, ainda que não se utilize sempre a mesma nomenclatura para as mesmas operações,

> Tais operações incluem a descrição de dados objetivos e intelectuais do documento. No âmbito da biblioteconomia, dá-se o nome de catalogação e indexação, respectivamente. No âmbito dos arquivos, estas operações vêm sendo denominadas de descrição arquivística de uma forma geral. (CAMPOS, 2006, p. 18).

Na próxima seção, verificar como a $E_{representa}$ e a $E_{conteúdo}$ se relacionam com outro processo: a recuperação da informação registrada em documentos.

10.1.1 A Linguística nos processos de representação e recuperação

Segundo Baranow (1983), a FID patrocinou um abrangente trabalho sobre a relação entre Linguística e Ciência da Informação no início da década de 1970[4] sob o título *Linguistics and Information Science* (JONES; KAY, 1973). Além desse trabalho e quase concomitantemente a ele, foram publicados outros – ainda que não tão abrangentes – os quais também salientaram essa relação como em Montgomery (1972) e Gardin (1973).

Em meados da década de 1960, podia-se identificar pessimismo sobre as vantagens dessa relação, mas tanto a linguística como a ciência da informação ainda eram campos novos: "linguística está ainda num estágio muito primitivo de desenvolvimento para prover uma base firme para os tipos de atividades práticas que os cientistas da informação estão envolvidos" (JONES; KAY, 1973, p. 4), mas, desde mais de quatro décadas atrás, ambas as disciplinas amadureceram bastante. No Brasil, desde o início da década de 1980, é possível identificar trabalhos que exploram especificamente essa relação como em Baranow (1983) ou Cintra (1983). E o tema ainda desperta interesse, como exemplificam dois livros recentes publicados sobre o assunto: Melo e Brascher (2011) ou Almeida (2011).

A pergunta sobre se a linguística pode ser útil para avanços na solução de problemas em ciência da informação deveria ser reformulada nestes termos: Quais teorias linguísticas são aplicáveis como apoio a soluções (pelo menos parciais) para os problemas

[4]Tivemos a oportunidade de adquirir a primeira edição dessa obra e corroboramos com a afirmação de Baranow. O livro ainda é, provavelmente, o mais abrangente sobre essa relação (apesar de desatualizado em alguns pontos) e focaliza na verdade a recuperação da informação, trata de linguística, estabelecendo relações com sintaxe e semântica. Apesar de publicado em 1973, aborda pesquisas e trabalhos desde meados da década de 1960.

enfrentados pela ciência da informação? Nesses termos, a pergunta pode ser respondida com mais objetividade.

Um ponto que merece ser esclarecido desde o início da discussão é sobre o elemento comum entre as duas áreas: a língua. Ainda que isso possa ser capcioso sem alguns esclarecimentos adicionais. A língua é claramente o objeto de estudo da linguística, a qual dá prioridade para sua forma verbalmente expressa e utilizada socialmente, de um ponto de vista sincrônico, como já mencionado na revisão de literatura anterior.

A ciência da informação, por sua vez, lida com a correspondente forma escrita daquele objeto prioritário linguístico, seja a escrita em textos registrados (conteúdos) ou a escrita acerca de conteúdo (representações, inclusive sobre documentos não textuais). Como também já citado antes, há diferenças importantes entre língua e escrita.

É claro que, possuindo a escrita uma relação estreita com a língua, as teorias linguísticas sobre a língua podem, potencialmente, ser aplicáveis quando se trata da prática em ciência da informação. É o que ocorre no caso da teoria linguística sobre o significado: a semântica. São emprestados dessa teoria o conceito de homonímia e o de polissemia, aplicáveis em instrumentos como tesauros e ontologias (KOBASHI, 2007).

O mesmo ocorre para a construção de terminologias, donde uma teoria sociolinguística sobre a variação da língua relaciona-se com a elaboração de sócio-terminologias (MAIMONE; TÁLAMO, 2011).

Num estudo publicado em 2000, foi possível identificar intersecções entre a linguística e a ciência da informação em sete pontos distintos: teórico, quantitativo, temático, aplicativo, ensino, tecnológico e nominativo (MENDONÇA, 2000).

Uma área da CI que possui forte relação com teorias linguísticas é a organização da informação, em especial a representação e recuperação e respectivos sistemas,

> Acessar informação é uma atividade primariamente linguística e os documentos disponíveis para recuperação através de sistemas de informação hoje são na grande maioria textos. Linguistas sabem sobre textos e deveriam saber sobre discurso e diálogo. A pesquisa sobre acesso à informação deveria precisar de linguistas. (KARLGREN, 2000, p. 41).

Alguns problemas, no entanto, permanecem em aberto com relação aos limites da teoria linguística para os processos de representação e de recuperação da informação. As teorias linguísticas abordam diferentes aspectos da língua, mas são certamente mais desenvolvidas ao nível do léxico (elementos linguísticos isolados) até a sintaxe (combinações entre o léxico), ou seja, sentenças, frases e orações. Mas, ao se tratar de textos e discursos, não

há teorias tão completas e consolidadas, o que se reflete nos limites e deficiências atuais da indexação automática ou geração de resumos.

Isso pode ser facilmente demonstrado se for feita uma análise morfológica de um elemento linguístico, seja o termo "homem" (substantivo, gênero masculino, pertencente ao léxico do vernáculo português brasileiro) ou uma análise sintática desse elemento numa frase: "Sócrates é um homem" (Sócrates = sujeito, é = verbo de ligação e "um homem" = predicativo do sujeito). Todo elemento lexical ou frase estão sujeitos a esse tipo de análise, que é completa conforme as respectivas teorias linguísticas, mas isso não ocorre para a análise de textos (ou discursos). Ainda que alguns textos sigam certas regras, como os textos científicos, a grande maioria não as segue, para não mencionar os textos de caráter poético. Os textos parecem fugir a regras universais de análise de maneira que a linguística ainda não deu conta de teorias sobre isso.

10.2 Recuperação da informação

Do ponto de vista de alguém que nasceu no mundo dos computadores pessoais e da internet, ou seja, nos últimos vinte anos, quando se fala em recuperação da informação, esse alguém pode fazer associações com buscas em algum aplicativo comercial. Contudo, a recuperação da informação começou muito antes da existência da internet (SANDERSON; CROFT, 2012).

Na verdade, a organização de documentos que pressupõe a possível recuperação é algo já encontrado na antiguidade, por exemplo, na Biblioteca de Alexandria. De maneira que a história de várias disciplinas que lidam com o documento hoje se confunde com o que se pode chamar, grosso modo, recuperação da informação (RI). Mas pode-se limitar a análise ao início do uso do termo com essa nomeação específica: *Information Retrieval*, como originalmente usado por Calvin Mooers no início da década de cinquenta nos EEUU: “a recuperação de informação trata dos aspectos intelectuais da descrição da informação e sua especificação para busca e também de qualquer sistema, técnicas ou máquinas que são empregadas para realizar esta operação” (MOOERS, 1951 apud FERNEDA, 2003, p. 11).

Mooers, o criador do termo *Information Retrieval* (IR), era um pesquisador da área de Tecnologia da Informação. E de fato, a Ciência da Computação é a área que vem introduzindo os maiores desenvolvimentos sobre recuperação da informação. Isso pode explicar, em parte, por que a CI – apesar de naturalmente interessada e envolvida com a RI, como demonstram várias pesquisas (FERNEDA, 2003; LIMA, 2004; AGUIAR, 2008;

LEITE, 2009; PAVÃO, 2014)[5] – não detém a exclusividade ou mesmo a predominância de pesquisas na área.

Uma alternativa é entender que as duas ciências (informação e computação) trabalham de maneira complementar, de maneira que a computação fornece "meios" e a CI, métodos e técnicas para tratamento:

> A Ciência da Computação fornece o meio, as ferramentas tecnológicas para o desenvolvimento do ambiente de recuperação da informação; já a CI fornece os métodos e técnicas de tratamento informacional (organização e representação) dos documentos disponíveis e faz uso das aplicações tecnológicas em seu fazer. (ALVES et al., 2007, p. 38).

Do ponto de vista da Ciência da Computação, é frequente reconhecer, nas técnicas da biblioteconomia, muitas das bases da RI, pelo menos em determinado sentido e escopo:

> A recuperação da informação (RI) está encontrando materiais (usualmente documentos) de natureza não estruturada (usualmente texto) que satisfaz uma necessidade de informação dentro de largas coleções (normalmente armazenado em computadores). Definido dessa maneira, recuperação da informação costumava ser uma atividade que apenas algumas pessoas se envolviam: bibliotecários de referência, auxiliares de advogados e profissionais similares de pesquisa. (MANNING; RAGHAVAN; SCHÜTZE, 2009, p. 1).

Uma coisa é tratar documentos numa biblioteca (mesmo que digital) e outra bem diferente é tentar recuperar conteúdo na rede internet. No caso da biblioteca, além do reduzido número de documentos, em geral se pode contar com um mínimo de tratamento de representação em seus documentos ao passo que, na internet, quase nada é tratado, se comparado ao tamanho total da rede.

Seja como for, essa discussão não é essencial para o desenvolvimento desta pesquisa e, se se enveredou por esse caminho, é com o objetivo de caracterizar um pouco a história de desenvolvimento e características da RI. Em resumo, sobre essa questão, pode-se afirmar que tanto a Ciência da Informação como a Ciência da Computação se ocupam da RI, às vezes, utilizando ferramentas e técnicas diferentes ou com diferenças no alcance de seus objetivos e, às vezes, com base teórica comum.

Em Ciência da Computação, há uma abordagem que procura automatizar todas as etapas de trabalho necessárias. No entanto, parece claro que alguma tentativa de representar as informações nos documentos abordados sempre ocorrerá, ainda que de maneiras bem diversas do que ocorre com o tratamento biblioteconômico ou arquivístico. O processo de

[5]Apenas para citar algumas Teses que utilizam o termo Recuperação da Informação em seus títulos.

representação, embora utilize técnicas de computação, ainda ocorre, pois se lida com informação produzida por humanos por meio da língua necessariamente. Como fica claro a partir dessa conclusão de análise de modelos de recuperação, de um ponto de vista da computação:

> O processo de recuperação de informação é inerentemente impreciso devido a fatores que talvez nunca serão totalmente equacionados. A modelagem matemática desse processo só é possível através de simplificações teóricas e da adequação de conceitos tipicamente subjetivos como "informação" e "relevância". Estas simplificações refletem em limitações qualitativas que se relacionam, por um lado, com a representação da complexidade semântica dos textos, e por outro, com a interação do usuário com os sistemas de recuperação de informação. (FERNEDA, 2003, p. 53-54).

Veja-se um exemplo de representação matemática (matriz) de termos em documentos na *tabela 3*. Essa matriz pode ser utilizada num sistema de recuperação para indicar se determinados documentos possuem ou não os termos utilizados numa busca.

Tabela 3 – Modelo binário para sistemas de recuperação da informação

Documento	Termo 1	Termo 2	Termo 3	Termo N
Doc_0001	0	1	0	0
Doc_0002	0	0	1	1
Doc_0003	1	1	1	0
Doc_0004	0	1	0	0
Doc_0005	0	1	0	1
Doc_0006	1	0	0	0
Doc_NNNN	1	1	0	0

Fonte: Elaboração própria.

No modelo binário, os termos (do 1 ao N) são associados aos documentos considerados para recuperação. No exemplo, o termo 2 está presente no doc_0003. É por isso que a matriz indica 1. Já o termo 3 não está presente no doc_0005 e, por isso, está marcado como 0. Um usuário que busque documentos sobre o termo 2 receberá do sistema de RI os documentos doc_0001, doc_0003, Doc_0005 e todos os outros até N que tenham o *status* "1", mas não receberá o doc_0002 ou doc_0006 como resposta.

Seja o termo 1 = "lavanderia" e o termo 2 = "medicina", pode-se afirmar que o doc_0006 é representado, nessa matriz, por "lavanderia" e não é representado por "medicina".

O ponto que se quer ressaltar com esse exemplo é que essa matriz também utiliza elementos linguísticos: os termos 1 a N, da mesma forma que qualquer instrumento tradicional utilizado na representação da informação de maneira que o tratamento

matemático-computacional não pode prescindir dos mesmos elementos linguísticos analisados na seção anterior sobre o P. de representação da informação, ou seja, ainda é válida a proposição em se tratando de processamento automático:

$\mathbf{E_{representa} = \sum e_1 \text{ a } e_n}$, onde de 1 a n são os elementos linguísticos registrados no instrumento e/ou produto considerado.

Substituindo e_1 a e_n pelos termos 1 a N no exemplo da matriz. Aqui, "instrumento e/ou produto" equivale(m) à matriz binária.

Importante também destacar que, da mesma forma como a representação da informação é vista como um processo, o mesmo ocorre com a recuperação da informação. De maneira que se fala aqui em dois processos independentes. Além disso, a relação que interessa entre esses dois processos são os *elementos linguísticos comuns*[6], além é claro dos documentos, cuja recuperação origina ambos os processos.

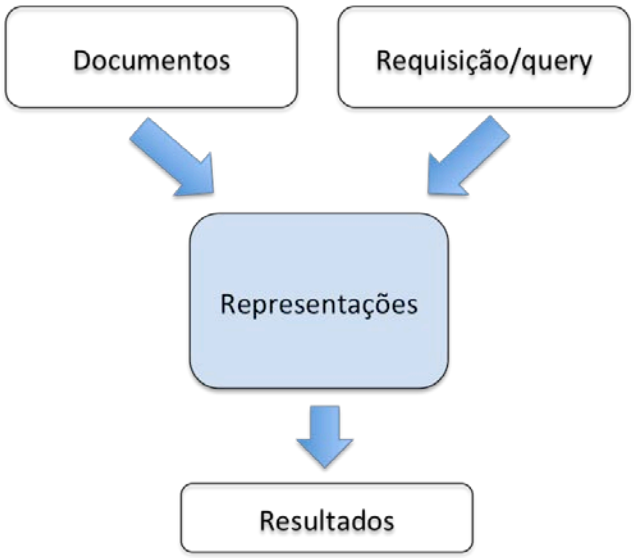
10.2.1 Modelos para RI e sua relação com a ML

Desde a década de 1960, é possível agrupar os modelos para recuperação da informação em três grandes grupos: (1) orientados para sistemas, (2) orientados para usuários e (3) os modelos cognitivos para recuperação da informação (INGWERSEN; JARVELIN, 2005).

Os modelos orientados para sistema são os mais básicos (*figura 8*) e excluem a participação do papel das pessoas nos processos representados. Consideram-se na modelagem orientada a sistemas os documentos, os quais registram a informação a ser representada ou recuperada, a própria representação correspondente e a requisição de pesquisa (*query*). As representações são armazenadas em bases de dados a partir da análise de documentos, utilizando-se algum algoritmo e as requisições efetuam consultas nessas mesmas bases de dados.

[6]Podem-se usar os termos linguísticos léxico e ou vocabulário, optou-se por essa forma porque em alguns casos o uso desses dois termos é restritivo, como em vocabulário de medicina, parte do léxico da ciência.

Figura 8 – Modelo básico orientado a sistemas



Fonte: Elaboração própria.

É importante ressaltar, no modelo básico acima, a ausência do papel das pessoas no processo representado por esse grupo de modelos, já que o objetivo deles é a avaliação da eficiência das *queries* e dos resultados nas buscas efetuadas nas bases de dados. No entanto, é difícil imaginar a aplicação da teoria linguística sobre mudança da língua num modelo desses, já que essa teoria depende fortemente das pessoas (e suas línguas) envolvidas naqueles processos.

O segundo grupo de modelos para representação e recuperação é aquele voltado para o usuário, também conhecido como modelos do comportamento informacional,

> A maioria dos modelos de comportamento informacional são definições, frequentemente na forma de diagramas, que tentam descrever uma atividade de busca de informação, as causas e consequências desta atividade, ou ainda os relacionamentos entre os estágios ou fases no comportamento de busca de informação. (WILSON, 1999 apud GARCIA, 2007, p. 77).

Os principais modelos no grupo dois que têm sido relatados na literatura são o modelo geral de comportamento de informação de Wilson, o modelo ISP (*Information Search Process*), o modelo de Dervin (*Sense-Making*) e o modelo de Ellis (comportamental de busca de informações). No entanto, uma pesquisa específica de análise desses modelos não evidencia nenhuma relação explícita com a teoria da ML (GARCIA, 2007). Isso apesar de enfatizar o papel do usuário e de seu contexto social real.

> Observa-se, então que uma abordagem que procure colocar o usuário no centro das preocupações na Organização da Informação deve recorrer, necessariamente, à experiência acumulada nos chamados estudos de usuários. Ou seja, os estudos sobre os comportamentos informacionais dos usuários, sobretudo os de busca e recuperação da informação os quais incluam as variáveis que interferem nestes comportamentos, podem contribuir para o desenvolvimento de novos sistemas e ferramentas para a Organização e Representação da Informação, assim como para o uso da informação, seja nos contextos acadêmicos/científicos ou, ainda, em outros tipos de organizações. (GARCIA, 2007, p. 126).

Confirmando isso, outra análise sobre a produção científica em ciência da informação e recuperação da informação que elenca as principais abordagens sociais também não inclui teorias linguísticas,

> As principais abordagens sociais que constituíram correlações com a temática Recuperação da Informação foram *Epistemologia Social*, *Psicologia Social*, *Hermenêutica*, *Construtivismo Social* e *Análise de Domínio*. Destacaram-se também outras temáticas que firmaram fortes laços de relacionamentos com essas abordagens e, sobretudo, com o tema Recuperação da Informação, entre elas: *Teoria da Informação* e *Organização do Conhecimento.* (MARTINS; LIMA, 2013, p. 12, grifos nossos)

Analisando o modelo *Sense-Making* (DERVIN, 1998), chamou a atenção o fato de a metodologia naquele modelo considerar também o fator tempo (algo essencial na teoria da mudança linguística), "Os conceitos fundacionais na metodologia *sense-making* são **tempo**, espaço, movimento, questões (*gap*), passos (*step-taking*), situação, ponte e resultado" (DERVIN, 1998, p. 39, grifo nosso). Mas a análise mais detalhada da descrição desse modelo (DERVIN, 1998) não estabelece relações com a teoria da ML. Com base num estudo de onze revisões de literatura, observa-se sobre estudos de usuários uma "carência de bases teóricas nos estudos" (ROLIM; CENDÓN, 2013, online).

O terceiro grupo de modelos para recuperação da informação, o cognitivo, parece ter suas bases na década de 1970, "por volta dos anos 70 o paradigma da informação deslocou-se em direção a uma contextualização mais ampla, tendo como foco principal o usuário e seu conhecimento individual, dando origem assim ao Paradigma Cognitivo" (ALMEIDA et al., 2007, p. 20).

De maneira geral, esse grupo de modelos baseia-se em cinco dimensões inter-relacionadas, conforme Saracevic (1996):

1. A interação de recuperação da informação é um conjunto de processos cognitivos, ocorrendo em potencialmente todos os elementos processuais da recuperação;

2. Usuários interagem não apenas com sistemas de recuperação da informação, mas também com objetos de informação – textos – que são estruturas cognitivas, consideradas como um espaço informacional;
3. O espaço cognitivo do usuário é um conjunto de elementos causais estruturados; contexto cognitivo e situacional são predominantes;
4. Interações ocorrem em diferentes níveis e consequentemente são de diferentes tipos;
5. O processo é altamente dinâmico. Uma poli representação é aplicada simultaneamente a ambos, espaço cognitivo do usuário e espaço informacional dos sistemas de informação.

Novamente, o exame da proposta inicial do modelo cognitivo por Ingwersen (1996), apesar de prever um "nível linguístico", não estabelece nenhuma relação com a teoria da ML, explicitamente ou não.

A conclusão à qual se chega, analisando – ainda que de maneira não exaustiva – os três grupos de modelos para representação e recuperação da informação é que os grupos (2) orientados ao usuário e (3) com orientação cognitiva, ambos ao considerar as pessoas nos processos de representação e recuperação e às vezes o fator "tempo" e "mesmo questões linguísticas", não especificam efeitos da ML, pelo menos de maneira explícita.

A linguagem natural e seu processamento não são estranhos nos problemas de recuperação da informação, "linguagem natural [língua] é variável, rica, flexível e em constante evolução. Ela contém muitas subculturas ou discursos baseados na idade, classe, raça, profissão, contexto ou uso" (INGWERSEN; JARVELIN, 2005, p. 151). Mesmo considerando esse cenário, os efeitos da ML em SRIs não fazem parte do rol de aspectos da língua mais relevantes.

10.2.2 O Problema do vocabulário e as pessoas

Em CI, têm um papel relevante as pessoas envolvidas em todos os processos de organização da informação desde a produção até a busca e recuperação, passando também pela representação.

É frequente que, nos temas discutidos em CI, as pessoas sejam analisadas no papel de usuários. Por exemplo, quando se abordam os SRIs, fala-se da necessidade do usuário e a incerteza entre o que o usuário quer, de que acha que precisa e o que expressa

linguisticamente: “um problema importante para a recuperação da informação e (para a representação do conhecimento em geral) é o perigo de descompasso entre o vocabulário expresso na busca do usuário e o vocabulário utilizado nos documentos relevantes” (KRAAIJ, 2004, p. 6).

Outros estudos enveredam pelo caminho do problema da interatividade entre o usuário e os sistemas de recuperação de informações. É nessa linha que são abordadas, também, as interfaces de usuário:

> Também nos modelos de Saracevic (1996) e Spink (1997), encontram-se referências importantes sobre a interação do usuário com o sistema de recuperação, utilizando conceitos relativos à interação homem-computador e que auxiliaram na definição deste modelo. Para Saracevic (1996), os usuários passam por três níveis de interação, com o sistema, com o texto e com o contexto, partindo da hipótese de que os usuários interagem com os sistemas para utilizar as informações e que o uso está relacionado com a aplicação situacional. (PAVÃO, 2014, p. 193).

O aspecto das pessoas que é particularmente importante tem a ver com as relações das línguas (e suas diferentes versões no tempo: estados da língua) utilizadas pelas pessoas no papel de produtor da informação registrada, representador dessas informações e recuperador.

Destacando-se o aspecto linguístico, o papel de recuperador aplica-se ao já tradicional usuário – como nos modelos de Saracevic, Ingwersen e Belkin (SARACEVIC, 1996), mas há outros papéis que também são relevantes, pelo menos a partir do prisma examinado. Esses papéis podem ser resumidos no *quadro 6*.

Em geral, eles papéis são assumidos por pessoas diferentes. Aqui se utilizarão nomes fictícios apenas com o propósito de ilustrar e destacar esse fato.

Quadro 6 – Pessoas e seus papéis em sistemas

Pessoas		
Papel de produtor (criador)	Papel de quem representa (indexador)	Papel de recuperador (usuário)
$E_{conteúdo}$	**$E_{representa}$**	**$E_{recupera}$**
elementos linguísticos no documento original (normalmente não presentes para documentos não textuais)	elementos linguísticos nos instrumentos e/ou produtos do P.de representação	elementos linguísticos das sentenças de busca

Fonte: Elaboração própria.

Sandra produz um documento memorando e expressa nele os elementos linguísticos de sua língua, no contexto sociocultural em que vive ($E_{conteúdo}$). Mário, na

profissão de indexador, representa aqueles elementos linguísticos presentes no memorando que considera mais relevantes ($E_{representa}$) para substituir o documento original de Sandra. Esses elementos linguísticos ($E_{representa}$) são utilizados para um determinado sistema a fim de recuperar, entre outros, aquele memorando. Em outro momento futuro (oito décadas na frente), Manuel expressa seus elementos linguísticos ($E_{recupera}$) para buscar e, eventualmente, recuperar documentos que devem incluir também o memorando que está sendo considerado no exemplo (de Sandra).

O fator complicador nesse cenário é o tempo de longo prazo, suficiente para alterar a língua original utilizada por usuários como Sandra, Mário e Manuel em função da mudança linguística.

Para documentos de guarda permanente, a pessoa que criou o documento sempre será a mesma, mas, ao longo do tempo, várias pessoas podem atuar como indexadores, além das pessoas que originalmente executaram essa atividade. De fato, normalmente várias pessoas atuam em equipe nos processos de representação em um mesmo momento. Da mesma forma, além dos vários usuários que buscam informações no sistema de RI considerado, outros usuários no futuro também poderão tentar utilizar os novos sistemas que estarão disponíveis.

No futuro, as pessoas no papel de representadores e de buscadores/recuperadores estarão utilizando outras versões da língua originalmente utilizada, de maneira que se pode falar em língua[1], língua[2], etc. (estados de língua) indefinidamente. Sabe-se que isso ocorre em função da teoria linguística sobre ML já abordada na revisão de literatura. Nessas condições, o problema original de possível falta de descompasso entre o vocabulário utilizado pelas pessoas em diferentes papéis tende a ficar cada vez mais relevante, em função da passagem do tempo.

10.3 Sistemas de recuperação da informação

Há uma relação próxima entre a representação da informação, como abordada em Ciência da Informação, e sistemas de recuperação da informação. De fato,

> Em síntese, estamos convencidos de que a informação organizada em sistemas requer mecanismos de mediação. As Linguagens Documentárias são, nesses dispositivos, instrumentos privilegiados de mediação que apresentam dupla função: a) representar o conhecimento inscrito e b) promover interação entre usuário e dispositivo. (KOBASHI, 2007, p. 2).

A rigor, sistemas de recuperação da informação (SRIs) podem recuperar, de fato, representações de documentos que contêm informação potencialmente útil para usuários. Além disso, o processo de alcançar o documento em si é independente da busca (LANCASTER, 1979 apud STYLTSVIG, 2006). Há um exemplo no decorrer das pesquisas nesta tese. Utilizou-se o Portal de Periódicos CAPES/MEC como fonte de pesquisa e, ao definir condições de buscas (booleana), foi recebida uma relação de descrições de vários documentos. Ao ler aqueles que estão em uma das línguas vernáculas do autor, seu conteúdo se transforma em informação útil. Antes disso, apenas se tentou acesso aos documentos originais, cujo acesso, em alguns casos, é negado.

É possível identificar uma relação entre produtores, especialistas em assuntos, usuários e os demais elementos para recuperação da informação (LANCASTER, 2004). O aspecto importante nessas relações é que "programas de computador" e "instrumentos de apoio intelectual", como ontologias são elementos de um "ecossistema" complexo, para usar os mesmos termos de Lancaster.

A mediação entre a informação desejada e as "pessoas em busca de informação" pode ocorrer por meio de diferentes tecnologias da informação (TICs). Não se refere somente a implementações no nível de linguagens, ferramentas de *software* ou tipos comerciais de bases de dados, ou seja, tecnologias da informação, pelo fato de que certos instrumentos tradicionais na Ciência da Informação (tesauros e ontologias) podem ou não ser utilizados como apoio nos SRIs. Na verdade, às vezes, nenhum instrumento tradicional em Ciência da Informação é efetivamente utilizado, como no caso do modelo binário e do uso de uma matriz.

Retomando a primeira citação desta seção, os sistemas em questão precisam ser estudados tanto em relação às tecnologias para representação do que é armazenado (exemplo *bag of words*) como do ponto de vista das tecnologias utilizadas para efetivar as buscas pelas descrições do que foi armazenado.

Um problema central que norteia o desenvolvimento tecnológico dos sistemas de recuperação da informação, em relação à RI, é sua relação com os aspectos subjetivos das línguas dos documentos armazenados. Esses aspectos conduzem a problemas que provocam incerteza entre o que o usuário busca, o que está armazenado no sistema (documentos e suas respectivas representações) e o que é ou pode ser recuperado. Para ilustrar essa subjetividade, imagine-se que um problema de busca de informação fosse hipoteticamente o seguinte:

"Um usuário expressa o desejo de localizar uma sílaba, digamos "pro" num documento textual com vinte páginas."

Nesse caso hipotético, o usuário pode expressar, clara e inequivocamente, o que deseja (sua *query*). Do ponto de vista do sistema, o problema também é trivial, basta que ele procure sequências de caracteres no conteúdo textual do documento que correspondam ao que é solicitado. Nesse cenário, tudo é claramente expresso e mensurável, mas, em situações reais, as *queries* podem ser bem mais complexas.

Mesmo considerando que um usuário saiba o que procura – por exemplo, mais informações sobre "cavalos" –, é bem provável que não saiba exatamente o que quer saber. Pode justamente estar tentando entender esse universo: alimentação, tratamentos, raças, locais de venda, preços, regulamentação sobre criação, registros de propriedade. Existem muitas possibilidades. Um sistema não tem acesso à mente do usuário e depende, necessariamente, da expressão linguística (na forma escrita)[7] sobre "pistas" do que o usuário necessita e o que busca.

Da mesma forma, normalmente há quantidades enormes de documentos com informações sobre qualquer assunto, mesmo para um caso simples. Pode-se considerar uma busca mais complexa como procurar "precedentes de julgamentos sobre perdão de dívidas tributárias relativos a impostos federais". Os sistemas não têm como verificar de forma totalmente exata e objetiva se o que há armazenado confere com o que o usuário expressa: "O desafio de base na recuperação é que a necessidade do usuário e o conteúdo do documento são ambos não observáveis e da mesma forma é a relevância entre eles" (SPARCK JONES; WILLETT, 1997 apud KRAAIJ, 2004, p. 4).

Essa incerteza entre a necessidade do usuário e o que um sistema recupera é um problema que se tenta equacionar antes da própria existência dos sistemas de recuperação. É em grande medida graças a esse problema que se criaram linguagens documentárias, indexação e técnicas de resumos e tantos outros mecanismos de tratamento intelectual da informação armazenada.

Uma alternativa utilizada pelos sistemas é a tentativa de automatizar esses processos originalmente biblioteconômicos, muitas vezes chamada de indexação automática: "A indexação automática é definida como a identificação dos termos significativos pelo computador a partir do texto integral ou do resumo dos documentos" (ORTEGA, 2002, p. 111). No entanto, a indexação automática ainda não funciona tão bem como a indexação humana, ainda que muitos avanços possam ser notados (LANCASTER, 2004).

[7]Vários sistemas atuais já utilizam interfaces de voz para receber as instruções de pesquisa. Meu celular iphone, por exemplo, tem a interface SIRI. Mas é importante notar que a voz aqui é apenas uma facilidade para o usuário, esses comandos de voz são transformados em elementos textuais e utilizados pelos sistemas como se tivessem sido digitados.

Uma alternativa de que os sistemas têm lançado mão é a recuperação pelo texto completo (*full text indexing*). Aqui não há necessariamente tratamentos semânticos nos termos utilizados na recuperação do sistema:

> A indexação de texto completo utiliza as representações textuais da *query* e dos documentos e trata cada palavra como um termo de indexação. Essa representação também é conhecida como representação *bag-of-words*, porque toda a ordem das palavras é perdida. A indexação de texto integral é fundamentalmente diferente da indexação controlada porque a relação dos termos indexados com um significado (relativamente) sem ambuiguidade é dispensada. (KRAAIJ, 2004, p. 5).

Há vários modelos de tecnologias para implementação da recuperação por texto integral como o modelo booleano, vetorial, probabilístico, por lógica *fuzzy*, dinâmicos e outros (FERNEDA, 2003).

O problema da incerteza em função da subjetividade e até de ambiguidades das línguas – principalmente da língua escrita e as tentativas de contornar esse problema –, tanto de forma manual (indexação por profissionais de ciência da informação) ou automática (várias tecnologias, inclusive tentativas de indexação automática), exige a necessidade de tentar mensurar o grau de qualidade da recuperação.

Certamente, os indicadores mais citados sobre o assunto são a indicação de precisão (*precision*) e revocação (*recall*), sendo que o coeficiente de precisão é obtido pela divisão do número de itens relevantes recuperados pelo número de itens recuperados pelo sistema. O coeficiente de revocação é obtido pela divisão do número de itens relevantes e recuperados pelo número de itens relevantes existentes no sistema (BOCCATO, 2009).

Há, todavia, vários fatores e indicadores possíveis para mensurar a qualidade de sistemas de recuperação da informação e há projetos específicos apenas para atestar a qualidade de algoritmos utilizados em processamento automático, como nos testes do TREC (*Text Retrieval Conference*): "um fórum internacional que possui coleções de documentos para avaliação de sistemas de recuperação de informação" (LEITE, 2009, p. 39).

10.3.1 Processamento da linguagem natural e mudança linguística

O processamento da linguagem natural (PLN), ou NLP na sigla em inglês, é uma subárea da linguística computacional que procura desenvolver ferramentas para tratamento da língua em aplicações de SRIs. No entanto, as técnicas de PLN normalmente são tratadas como pré-processamento e não são incluídas nos modelos de RI em si (HIEMSTRA, 1998). De maneira que sua análise deve ser feita em separado à análise dos modelos de RI.

Técnicas de PLN têm sido utilizadas em processos de representação e recuperação da informação em relação ao problema da variação da língua. No entanto, o termo "variação da língua" possui um significado especial no âmbito das pesquisas com PLN, mas não está relacionado à variação sociolinguística ou à teoria da mudança linguística.

Em PLN, "variação linguística" pode ocorrer no sentido de variação morfológica, variação lexical, variação semântica ou variação sintática (ARAMPATZIS et al., 2000), seguem os exemplos nesta obra (serão utilizados exemplos originais da publicação em inglês):

- Variação morfológica: wolf e wolves; man e man's
- Variação lexical: film e movie (sinônimos para filme)
- Variação semântica: bands e radio frequency bands (ambos referem-se a bandas)
- Variação sintática: "near to the river, air pollution is a major problem" não é uma frase que se refere a "river pollution"

A literatura sobre o assunto também relaciona como "variabilidade" a "subespecificação" e a "redução" (LEWIS; JONES, 1996). Nesse caso, "subespecificação" refere-se a utilizar termos vagos numa requisição de informação (*query*). Ao utilizar somente o termo *cheap* = produção econômica, esse termo também pode se referir a *cheap* = baixa qualidade. E "redução" significa excesso de simplificação na especificação da requisição de informação como utilizar o termo "*construction*" em vez de "*design and construction*".

Há várias soluções tecnológicas para corrigir a "variação linguística" nos sentidos exemplificados acima. E mesmo a variação sociolinguística ou a mudança linguística também podem ser tratadas tecnologicamente, se adequadamente formatadas em teorias e se houver subsídios de elementos linguísticos que indiquem tanto a variação como a mudança, em termos de elementos linguísticos (equivalências, sinônimos, formas equivalentes etc.).

Outro aspecto fundamental sobre as técnicas de PLN é que são justamente elas que permitem a integração de instrumentos tradicionais da ciência da informação como tesauros ou vocabulários controlados e os modelos de *full-text-indexing*, "o longo debate sobre controlado versus indexação de linguagem natural tem-se tornado menos importante já que muitas bases de dados comerciais utilizam ambos" (LEWIS; JONES, 1996, p. 95). Note-se que se trata de uma citação de mais de uma década atrás.

Encerra-se esta seção neste ponto. A seguir, é apresentada uma análise sobre como a CI, pelo menos brasileira, tem encaminhado a condução do problema dos efeitos da ML em SRIs.

10.3.2 PLN e humanidades digitais

Humanidades digitais é um termo que reúne na atualidade várias disciplinas de pesquisa sobre registros culturais, atual ou não (interessam particularmente os registros culturais produzidos nos séculos anteriores), seus conteúdos na forma digital e ferramentas de leitura e análise desse conteúdo. Acrescenta-se esse tema à revisão de literatura em função da análise documental levada a efeito mais adiante. Essa área fornece as teorias e conceitos que serão aplicados naquela análise.

No Brasil, a revista Texto Digital[8] reúne trabalhos de vários pesquisadores brasileiros ou não com resultados de pesquisas em humanidades digitais. Há vários outros periódicos no mundo que trata de aspectos ou áreas específicas, conforme resumo da área preparado pelo grupo de pesquisas sobre humanidades digitais da Universidade de São Paulo[9].

Um elemento específico no universo do que é chamado de humanidades digitais, de um ponto de vista de uma docente em linguística, que é particularmente importante, é o conceito de "texto digital" em contraste com o conceito de texto tradicional, tanto manuscrito como impresso mecanicamente:

> No caso do "texto digital", estaremos diante de algo inteiramente diverso Neste caso, não apenas a forma de levar a informação codificada é singular, mas – fundamentalmente – o processamento da informação a ser codificada e decodificada é outro, uma vez que envolve, além da correspondência "lógico-sensorial" humana, etapas de correspondência lógica artificial. (SOUZA, 2009, p. 162).

Um texto digital, no sentido acima definido, não se limita ao que é produzido originalmente em meio digital, mas também inclui o conteúdo cultural produzido no passado que – após os devidos tratamentos técnicos – pode ser pesquisado e analisado como se fosse originalmente produzido na forma digital. Um dos projetos mais importantes nesse sentido foi o tratamento de textos clássicos gregos, incluindo sua disponibilização para acesso público, através do projeto *Thesaurus Linguae Graecae*[10] (TLG), iniciado na década de 1970 na

[8]Vide sítio da revista: <https://periodicos.ufsc.br/index.php/textodigital/about>.
[9]Vide sítio do grupo: <http://humanidadesdigitais.org/>.
[10]Vide sítio do projeto: <http://stephanus.tlg.uci.edu/history.php>.

universidade estadunidense da Califórnia. No Brasil, o projeto Brasiliana USP[11] com a digitalização e tratamento da biblioteca pessoal de José Mindlin, também é um exemplo de projeto nesse sentido. Ambos os projetos tratam importantes documentos e seus conteúdos, amostras relevantes de elementos de nossa cultura e os disponibilizam.

Tais projetos ilustram tratamentos de textos antigos. As pesquisas sobre humanidades digitais não se limitam a esse tipo de texto, mas é ele que aqui interessa em função da já mencionada pesquisa documental. Nesse contexto, o processamento da linguagem natural (PLN) tem atuado conjuntamente com as pesquisas em humanidades digitais, inclusive com o que tem sido chamado de filologia digital, fornecendo várias ferramentas de *software* que permitem desenvolver pesquisas com mais eficiência, inclusive pesquisa no sentido de busca e recuperação da informação, já que um texto antigo apresenta vários problemas para a pesquisa em SRIs contemporâneos e, como é defendido nesta tese, também nos sistemas que serão utilizados nos próximos séculos.

De fato, as ferramentas de apoio às humanidades digitais permitem projetos ambiciosos como a centralização de vários tipos de documentos, em várias línguas e registros temporais, como no exemplo de tratamento dos registros num museu (KOOLEN et al., [200?]). Em função das especificidades de cada língua, ferramentas de *software* precisam ser produzidas, testadas e aprimoradas para cada vernáculo. Para o caso da língua portuguesa, tanto de Portugal como do Brasil, há iniciativas de projetos nesse sentido. Um projeto da Biblioteca de Évora em Portugal resgata uma amostra de registros do português antigo.

O plano de investigação, as tarefas, a metodologia e os resultados previstos visavam, pois, alcançar um maior conhecimento das fontes metalinguísticas do português, contribuir para o seu espaço na rede e para o avanço da investigação sobre a língua portuguesa e a sua memória (GONÇALVES; BANZA, 2013).

Para o estudo de textos na forma digital, o conceito de *corpus* é fundamental e se refere ao conjunto documental objeto de tratamento – através de ferramentas de *software* baseadas em PLN – para posterior pesquisa, sobre as etapas para construção:

> A vida útil de um córpus pode ser dividida em quatro etapas: projeto, compilação, anotação e uso. A etapa de projeto consiste na definição dos objetivos do córpus e na tomada de decisões a respeito de sua constituição. A etapa de compilação envolve a estratégia de coleta de textos, conversão para o formato digital (caso ainda não estejam) e pré-processamento desses textos. Na etapa de anotação (opcional), os metadados dos textos (por exemplo, informações estruturais de parágrafos e capítulos ou informações lingüísticas nos níveis morfossintático e sintático) são identificados e anotados para uso em ferramentas de processamento de córpus. Por

[11]Vide sítio do projeto: <http://www.brasiliana.usp.br/>.

> fim, o córpus é então usado para as pesquisas para as quais foi originalmente concebido. (CANDIDO JUNIOR, 2008, p. 18).

A partir da escolha e tratamento, obtém-se a versão digital do corpus, que não se limita apenas à uma versão digitalizada, mas sim na forma de texto digital, editável e tratado com editores comuns de texto ou também disponíveis em linguagens de marcação como HTML ou XML. Mas essas novas versões em texto digital também precisam de ferramentas de apoio para o tratamento adequado do vernáculo e de seu léxico (elementos linguísticos), como dicionários específicos adaptados a esse contexto. Sobre dicionários nesse contexto:

> O léxico computacional, ou dicionário, é uma estrutura fundamental para a maioria dos sistemas e aplicações de PLN. Trata-se de uma estrutura de dados contendo os itens lexicais de uma língua e as informações correspondentes a estes itens. Esses itens podem ser palavras isoladas (como lua, mel, casa, modo) ou composições de palavras com um significado específico (por exemplo, lua de mel ou Casa de Cultura ou *a grosso modo*). (MUNIZ, 2004, p. 5).

O português brasileiro já possui dicionários para aplicações em PLN, um deles está integrado ao aplicativo UNITEX[12] no padrão DELA (*Dictionnaires ´electroniques du d'informatique documentaire et linguistique*) da Universidade de Paris 7, na França. O processo de criação desse dicionário compreendeu três etapas, " projeto e implementação dos dicionários DELAS e DELAF, o projeto e implementação do dicionário DELACF e o desenvolvimento da biblioteca para acesso e manipulação ao UNITEX-PB" (MUNIZ; NUNES; LAPORTE, 2008, p. 5).

Tanto a ferramenta de *software* UNITEX na versão português brasileiro (UNITEX-PB) como os dicionários na citação anterior serão utilizados na análise documental desta tese.

10.4 Considerações finais desta seção

Esta seção tratou do processo de representação da informação, do ponto de vista da ciência da informação e do processo de recuperação da informação do ponto de vista da ciência da computação. Tratou-se com mais profundidade de modelos para recuperação da informação, pois essa análise permite a compreensão adequada das relações com a teoria linguística da mudança linguística (ML) que será aplicada ao longo da pesquisa. Concluiu-se com uma breve análise das relações entre teorias linguísticas de maneira geral e os dois

[12]<http://www-igm.univ-mlv.fr/~unitex/>.

processos independentes e inter-relacionados aqui analisados: representação e recuperação da informação. Também foram incluídos nas análises desta seção técnicas de processamento de linguagem natural. Tanto o processo de representação, como também o processo de recuperação da informação apoiam a elaboração do objetivo específico um, o quadro sinóptico. As considerações sobre o processamento da linguagem natural serão aplicadas no objetivo específico da análise documental.

É importante esclarecer que ao abordar de maneira ampla o processo de recuperação da informação não pretendemos fazer um recorte sobre tecnologias específicas como a RI na Internet ou em sistemas do tipo *desktop*. Estamos pesquisando os efeitos da mudança linguística em futuros sistemas de RI e não é possível antever qual tecnologia atual será a mais utilizada no futuro, ou mesmo se qualquer uma delas, a exemplo da duas aqui citadas, ainda será utilizada. Assim, ao tratar do processo de recuperação da informação, objetivamos extrair noções conceituais gerais sobre este processo. Mas, principalmente, tanto ao tratar de recuperação da informação como também da representação da informação nosso objetivo foi explorar as relações entre esses processos e a língua, notadamente a relação de dependência com a língua. É com este intuito que estão nesta seção as subseções sobre linguística nos processos de representação e recuperação e a análise dos modelos de RI em sua relação com a língua e mudança linguística.

11 QUADRO SINÓPTICO PRINCIPAIS CONCEITOS

A seguir, apresenta-se um quadro sinóptico dos conceitos básicos diretamente relacionados ao problema desta pesquisa. As explicações e comentários a seguir são baseados nos itens da revisão de literatura anterior, portanto resultado coletivo. E, se não foram citadas novamente todas as referências, é porque isso seria redundante no contexto de resumo apresentado para cada conceito. Mas são indicadas as seções em que os conceitos são tratados na revisão de literatura anterior.

A figura 9 representa este quadro. Note-se que a complexidade dos conceitos no quadro sinóptico é reduzida a um bloco representado por retângulos ou a figura de documentos ou de uma pessoa. As relações entre os blocos são estabelecidas com setas. O quadro está dividido entre parte superior e inferior. Na superior, representa-se uma perspectiva sincrônica e na inferior a perspectiva (privilegiada nesta pesquisa) diacrônica

Linguagem

Linguagem é um termo que aponta para um significado ou até significados bastante abrangentes. O significado mais próximo dos interesses desta tese está associado às formas de comunicação e transmissão de conhecimento humano. Há também pesquisadores que utilizam o termo para se referir às formas de comunicação entre não humanos como abelhas e baleias. Mesmo apenas no escopo do ser humano, as possibilidades de comunicação e linguagem são muitas. Assim, pode-se falar de linguagem em relação às expressões faciais, contato físico (aperto de mão, aceno de adeus, tapinha nas costas, abraço), sinais de fumaça, comunicação verbal pela fala, comunicação artística pelas artes plásticas, comunicação pela escrita. Não bastassem todas essas possibilidades, muitas traduções do inglês para o termo *language* não distinguem entre língua (no sentido de língua falada) e linguagem num sentido mais amplo. A ponto de, nesta tese, ter-se optado pela grafia “língua(gem)”, como o fazem outros autores, para tentar obter maior precisão linguística. O conceito linguagem foi tratado no item 7 da revisão de literatura.

No contexto desta tese, esse conceito serve como parâmetro de referência e contraste com os termos que realmente são aplicados nesta pesquisa: língua (falada) e escrita (língua escrita). Na figura 9 (parte superior e inferior), a linha pontilhada que segue para baixo sugere justamente outros significados para o termo, como a comunicação não verbal. A seta indica sua relação com a língua.

Língua falada

Língua falada, termo que, nesta tese, é nomeado simplesmente como língua, é objeto de estudo de várias disciplinas. A linguística a tem como objeto prioritário de estudo. É a capacidade humana de se comunicar verbalmente nos limites do que pode ser expresso verbalmente. Materialmente, a língua se manifesta principalmente por meio da capacidade de fala, mas, complementando concomitantemente a fala, estão também expressões faciais e corporais. Uma característica importante desse conceito para a pesquisa é sua não perenidade, ou seja, por si só, o que é dito não se mantém (tanto para quem diz como para quem ouve) com precisão, necessitando de outros recursos para ser memorizada ou preservada de maneira inequívoca. A escrita é o mecanismo privilegiado nesta tese como mecanismo de perenidade da língua, mas, de fato, atualmente, a língua pode ser gravada com tecnologias sonoras, inclusive associadas a vídeo. Deve ser entendida como um sistema complexo com vários aspectos interdependentes que precisam ser analisados em conjunto. O conceito de língua falada foi tratado na seção 7, particularmente 7.4.1. Ele é retomado também na seção 8.2 (a língua como um sistema).

A linguística estuda esse sistema complexo em diferentes aspectos: fonético/fonológico, lexical, sintático, morfológico, semântico e pragmático. Na figura 9 (parte superior e inferior), o conceito aparece relacionado ao conceito de linguagem e escrita. As relações na figura também procuram evidenciar que o uso social de uma língua é uma função da variação linguística e dos níveis de língua associados, como se verá adiante.

Língua escrita

Nesta tese, língua escrita é simplesmente escrita. Também é objeto de estudo de várias disciplinas, como a paleografia e a própria linguística, embora para esta não seja objeto prioritário de estudos, mas sim a língua (falada). É uma ferramenta que permite o registro de uma língua nos limites do que pode ser verbalmente expresso, mas com menos recursos que uma língua na forma falada, como pausas, entonações e associação com expressões faciais. Esses outros elementos no uso da língua são difíceis de reproduzir na escrita. É um elemento essencial em documentos textuais, pois é através desta ferramenta que informações são registradas em suportes. Pela importância, o conceito foi tratado numa seção exclusiva, seção 9 (a língua escrita).

A escrita textual, no sentido contemporâneo, está diretamente associada a uma língua. De fato, os estudos sobre uma língua são geralmente feitos pelas representações escritas. As evoluções numa língua, por exemplo, são perceptíveis e analisadas pela

comparação de textos escritos. A escrita registra perenemente as características do estado atual de uma língua em seus aspectos lexical, morfológico, sintático, semântico e pragmático. Na figura 9 (parte superior e inferior), as setas sinalizam sua relação direta com o conceito língua e seu uso na produção de documentos textuais.

Variação linguística

O conceito de variação linguística tem empréstimo teórico, principalmente, da sociolinguística. Esta é uma subárea da linguística que procura compreender o uso de uma língua no contexto social, embora pareça irrelevante fazer tal afirmação, pois toda língua só faz sentido num contexto social. Com uso no contexto social quer-se dizer que uma determinada língua vernácula não é homogênea, ou seja, não se pode falar numa língua qualquer como algo fixo e bem determinado. Uma mesma língua varia no uso social, desde as camadas menos instruídas da sociedade até as camadas mais cultas. E, mesmo numa camada social como a culta, há variação entre o uso familiar e profissional. Varia também geograficamente com o uso regional de determinado vocabulário ou mesmo variações fonético-fonológicas (português do sul e do nordeste, por exemplo). O conceito de variação linguística foi tratado na seção 8 e com mais detalhes na seção 8.1 (sociolinguística quantitativa).

Essas variações são observadas concomitantemente num momento temporal (sincrônico), de maneira que se pode falar no conceito de “níveis de uma língua”, abordado adiante. As relações na figura 9 (parte superior e inferior) procuram evidenciar que os níveis de língua são o resultado da variação linguística em uma língua vernácula. A variação linguística define, em grande medida, como uma língua se manifesta numa região e situação social.

Os níveis de língua

É um conceito diretamente relacionado ao conceito de variação linguística. Os níveis de língua são o resultado da adequação ao uso social e regional. Eles implicam formas específicas para o uso da língua (como o vocabulário e uso fonológico). Uma mesma pessoa pode e geralmente utiliza diferentes níveis de sua língua nativa, como em seu núcleo familiar ou no ambiente profissional e acadêmico. Há classificações para níveis de uma língua, os termos mais comuns são língua culta ou coloquial, baixo calão, linguagem técnica. A escrita reflete o nível de língua utilizado. Em ambientes profissionais, geralmente, o registro da escrita segue a norma com maior prestígio social e/ou oficial do país. O conceito de nível de língua foi tratado na seção 8.1 (sociolinguística).

Na figura 9 (parte superior e inferior), as setas no quadro sinóptico indicam sua relação entre o conceito “variação linguística” e “língua”.

Documento Textual Histórico de Arquivo

Os registros de informações nestes documentos ocorrem de maneira natural e espontânea, pois são produzidos para cumprir funções administrativas, legais e/ou fiscais, mas, ao cumprir essas funções, acabam por registrar diferentes aspectos da vida cultural de um país. Assim, documentos textuais com caráter histórico de arquivo fazem parte do que é considerado patrimônio cultural de um país ou legado cultural. Um documento textual é produzido por uma pessoa mediante o uso de uma determinada língua e sua respectiva escrita que é, portanto, parte essencial do processo de produção do documento. Este conceito foi tratado junto ao conceito de escrita na seção 9.4 (escrita e documento de arquivo).

Na figura 9 (parte superior e inferior), as relações na cadeia “linguagem” + “língua” + “escrita” e “variação linguística” + “níveis de língua” + “língua” + “escrita” são evidenciadas pelas setas indicadas.

Representação da informação

Trata-se de uma área prioritariamente pesquisada em ciência da informação e organização da informação. Refere-se a vários processos e instrumentos que procuram “representar” as informações registradas em um documento, de certa forma e até certos limites, substituindo o documento original. Os produtos que substituem o documento original, como um resumo, palavras-chave ou termos indexadores, podem então ser utilizados para analisar o documento original antes de seu acesso efetivo. Interessa particularmente o fato de que as informações que representam o documento original podem também ser utilizadas nas operações de busca e recuperação de documentos. As representações dos documentos textuais (mas de fato não apenas do tipo textual) são informações registradas por meio da escrita associada à língua das pessoas que executam as atividades de representação. Há procedimentos para automatizar a representação como a “indexação automática”, mas ainda não são totalmente seguros na qualidade dos resultados para todo tipo de texto e documento. Este conceito foi tratado na seção 10.1 (representação da informação).

Na figura 9 (parte superior e inferior), a seta procura evidenciar sua relação com documentos textuais.

Recuperação da informação

Trata-se de uma área prioritariamente estudada em ciência da computação e

refere-se aos processos informatizados para tanto buscar como recuperar efetivamente a informação digital registrada. Normalmente envolve bancos de dados e modelos específicos para busca e recuperação. Atualmente, destacam-se seus avanços no âmbito da rede internet, na qual buscadores comerciais como yahoo, bing e google têm apontado uma revolução na relação entre as pessoas e as informações disponíveis. A busca e a recuperação podem ser feitas diretamente em documentos, mas são quase sempre executadas sobre algum tipo de representante dos documentos originais (não necessariamente no mesmo sentido utilizado em ciência da informação) como uma *bag-of-words*, elemento utilizado na chamada recuperação em *fulltext indexing*. Para executar uma busca por informações registradas em documentos (e na verdade em qualquer recurso digital como *sites* ou bancos de dados), é necessário que uma pessoa especifique sua proposição de busca (o que deseja procurar) na forma de texto (uma *query*). Aqui, é relevante destacar que essa pessoa – em princípio – formula o texto na query utilizando sua própria língua (ou outro vernáculo que domine) como a conhece naquela época. Ainda que possa formular utilizando formas antigas do léxico por exemplo, desde que conheça tal versão de sua língua. Este conceito foi tratado na seção 10.2 (recuperação da informação).

Na figura 9 (parte superior), o bloco referente a este conceito está esmaecido (mais apagado), pois a parte superior refere-se à perspectiva sincrônica (todos os processos ocorrendo contemporaneamente). Na parte inferior do quadro sinóptico, o bloco encontra-se na cor normal, pois está inserido num esquema diacrônico, ou seja, ocorre em um estado posterior da língua considerada em relação ao estado no qual as informações foram registradas.

Sistemas de recuperação da informação

Este conceito pode ser utilizado em sentidos amplos, mas, no escopo desta tese, trata-se de sistemas informatizados associados à implementação da busca e recuperação da informação registrada em documentos textuais. Normalmente, baseiam-se em modelos específicos, alguns deles analisados na revisão de literatura. Há várias tecnologias disponíveis para sua implementação, destacando-se aquelas voltadas para a busca e recuperação na rede internet. Uma característica também relevante no escopo desta tese é a dependência em relação ao registro escrito. Tanto as solicitações de busca (*queries*) como os resultados são apresentados na forma escrita, ainda que alguns sistemas implementem tecnologias para conversão de voz em texto (e vice-versa) a fim de simplificar a interface humana. Este conceito foi tratado na seção 10.3 (sistemas de recuperação da informação).

Da mesma forma que para o bloco “recuperação da informação”, este conceito é apresentado esmaecido na parte superior do quadro sinóptico (perspectiva sincrônica) para demonstrar que interessa o uso deste conceito e suas ferramentas de *software* e *hardware* quando utilizadas em estados posteriores da língua (no futuro) em relação ao estado atual (parte inferior da figura).

Problema do Vocabulário

Este é o nome dado, principalmente no âmbito da ciência da computação, aos diferentes usos que cada pessoa faz de uma língua. Mesmo quando se considera um grupo de pessoas que utiliza nativamente uma mesma língua vernácula, cada pessoa (individualmente) pode fazer uso de recursos específicos, ou seja, o problema se caracteriza em função da riqueza expressiva de cada língua, incluindo regionalismos e variações no uso social (diferentes extratos sociais), mas não se limitando à variação linguística. Do ponto de vista da busca e recuperação da informação, se diferentes pessoas registram e propõem *queries* sobre um mesmo conceito ou objeto, mas utilizando diferentes elementos linguísticos (no nível lexical ou sintático-morfológico, por exemplo), podem surgir problemas nos processos de recuperação da informação. A mudança linguística (analisada no próximo item) pode potencializar o problema do vocabulário na medida em que se consideram diferentes pessoas utilizando diferentes estados de uma língua nas etapas de registro, busca e recuperação. Este conceito foi tratado juntamente com o conceito de recuperação da informação na seção 10.2.1 (o problema do vocabulário e as pessoas).

O bloco correspondente ao conceito “problema do vocabulário” no quadro sinóptico da figura 9 (parte superior) também aparece esmaecido, pois ele ocorre na perspectiva sincrônica e diacrônica, mas aqui interessa sua ocorrência entre estados de uma língua (perspectiva diacrônica) apresentada na parte inferior do quadro sinóptico. E influenciado pelo problema da mudança linguística.

Estado de uma língua

A vantagem didática no uso de um quadro sinóptico, na forma gráfica, como apresentado na figura 9 é a exibição de todos os conceitos e de suas relações concomitantemente. Optou-se por tratar por último do conceito de estado de uma língua e mudança linguística (no próximo item), pois ambos os conceitos e suas relações com os demais parecem ser os elementos chave para compreensão do cenário tratado no quadro.

Estado de uma língua é o nome que se dá para o sistema língua (que é complexo) quando analisado num determinado momento e comparado a outros momentos. Há evidências

de que uma língua muda contínua e permanentemente, mas, se comparada entre lapsos temporais relativamente longos, como séculos, observam-se diferenças em todos os aspectos da língua considerada. As mudanças que ocorrem continuamente em intervalos curtos de tempo são normalmente muito pequenas e às vezes não são incorporadas em novos estados subsequentes de uma língua. É preciso então distinguir entre as diferenças observadas quando se comparam longos períodos de tempo (estados de uma língua) e as pequenas mudanças (muitas vezes não perenes) detectadas em curtos intervalos de tempo. Na prática, pode-se falar em dois estados de uma língua entre dois anos seguidos, mas dificilmente se podem comparar diferenças significativas neste caso. É também importante considerar aqui o fato de que uma língua não é homogênea, pois existem vários níveis de uma mesma língua. Isso implica que esta comparação só pode ser feita entre níveis de língua específicos. Por exemplo, podemos é possível comparar uma língua entre dois estados consecutivos em relação ao nível culto utilizado na escrita. Este conceito foi abordado juntamente com o conceito de mudança linguística na seção 8 (mudança linguística).

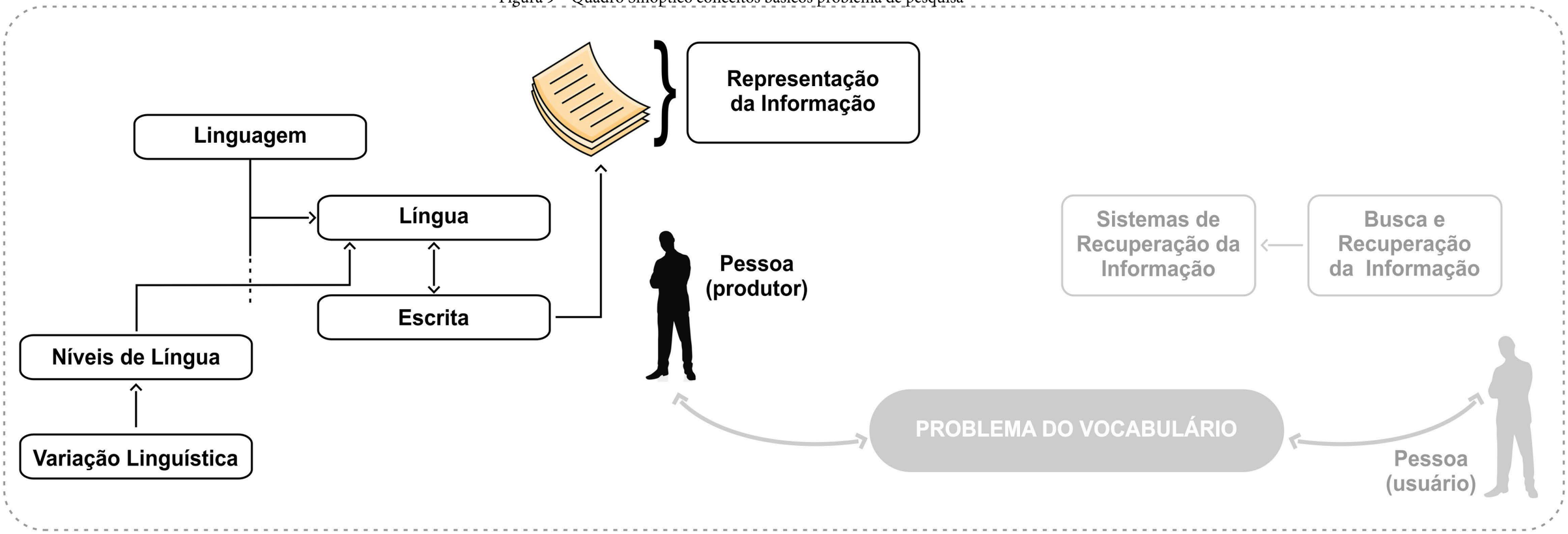
No quadro sinóptico os estados de uma língua estão representados pelos tracejados na parte inferior. No lado esquerdo os processos contemporâneos (estado de língua: anterior) e no lado direito os processos que ocorrerão no futuro (estado de língua: posterior).

Mudança linguística

É um fenômeno que ocorre na língua, observado nos registros escritos, ainda que, desde o advento do registro sonoro, também pode ser observado por intermédio deste último. Desde que as línguas puderam ser registradas e assim se pôde compará-las em diferentes momentos históricos, tal comparação – através da escrita – permitiu verificar mudanças em vários aspectos de uma língua. Não há motivos para acreditar que as línguas, como é o caso da língua portuguesa, deixará de mudar hoje e no futuro. Até porque se observa o fenômeno da variação linguística na atualidade. Admite-se que a variação linguística é uma das fontes das mudanças observadas entre estados de uma língua. Por sua importância, este conceito foi tratado numa seção exclusiva (8).

Na representação do quadro sinóptico, um bloco negro na vertical entre os tracejados do estado anterior e estado posterior representa, abstratamente, as mudanças ocorridas e comparadas entre os estados considerados.

Figura 9 – Quadro Sinóptico conceitos básicos problema de pesquisa



Fonte: Elaboração própria.

12 METODOLOGIA

Esta seção apresenta diferentes aspectos da metodologia de pesquisa utilizada nesta tese. Apresenta-se a caracterização da pesquisa com fundamento em autores da disciplina metodologia de pesquisa. Na sequência, o escopo de pesquisa é definido e são apresentados os procedimentos específicos que foram realizados. Ao final, apresentamos as conclusões obtidas.

12.1 Caracterização da metodologia da pesquisa

Classificações de coisas, teorias ou conceitos são organizações lógicas a partir do ponto de vista de uma pessoa ou grupo. É assim também com a classificação sobre metodologia de pesquisa. A depender do autor e de sua escola, diferentes combinações sobre métodos e princípios de pesquisa podem ser propostos, muitas vezes apenas com alterações nas nomenclaturas entre as diferentes obras.

Desde a formulação do projeto de pesquisa que deu origem a esta tese, foram consultadas diversas obras disponíveis como Cervo e Bevian (2002), Richardson (2010), Gil (2006), Laville e Dionne (1999), Severino (1992), Eco (2006) e, no prisma estadunidense, Booth, Colomb e Williams (2003). Todas estas obras foram consultadas a fim de estabelecer uma referência adequada e sólida para sustentar a metodologia. Optamos pela adoção da classificação sobre metodologia científica do Prof. Antonio Carlos Gil. Os motivos são, primeiro, o fato de sua obra ser planejada especificamente para as ciências sociais e, segundo, o fato de tratar-se de uma obra consagrada. Sua primeira edição é de 1982, foi utilizada aqui a 5.ed (7. Reimpressão) de 2006 (a sexta edição é de 2008). Além disso, a classificação adotada por este autor possui muitos elementos comuns às demais obras consultadas.

De acordo com o referencial do Prof. Gil, “para que um conhecimento possa ser considerado científico, torna-se necessário identificar as operações mentais e técnicas que possibilitaram a sua verificação” (GIL, 2006, p. 26). O autor organiza os métodos entre aqueles “que proporcionam as bases lógicas da investigação científica e o dos que esclarecem acerca dos procedimentos técnicos que poderão ser utilizados” (GIL, 2006, p. 26-27).

Neste caso, em relação às bases lógicas da investigação, nosso trabalho adotou, predominantemente, ao método indutivo “De acordo com o raciocínio indutivo, a generalização não deve ser buscada aprioristicamente, mas constatada a partir da observação de casos concretos suficientemente confirmadores dessa realidade” (GIL, 2006, p. 28). O

autor também defende tratar-se do “método mais adequado para investigação nas ciências sociais” (GIL, 2006, p. 29).

Com relação aos meios técnicos da investigação, esta pesquisa apoiou-se, predominantemente, no método estatístico (com estatística básica), “este método fundamenta-se na aplicação da teoria estatística da probabilidade e constitui importante auxílio para a investigação em ciências sociais” (GIL, 2006, p. 35).

O problema de pesquisa abordado nesta tese mostrou-se complexo, pois necessita da inter-relação de teorias e conceitos de diferentes áreas: principalmente a linguística, a ciência da computação e finalmente a ciência da informação, área na qual se encontra o ponto de vista deste trabalho. Essa complexidade também pode ser verificada e confirmada nos diferentes assuntos, temas e conceitos tratados na revisão de literatura.

Tal complexidade levou a considerar como nível de investigação social o estudo exploratório para o problema de pesquisa proposto: ”as pesquisas exploratórias têm como principal finalidade desenvolver, esclarecer e modificar conceitos e ideias, tendo em vista, a formulação de problemas mais precisos ou hipóteses pesquisáveis para estudos posteriores” (GIL, 2006, p. 43).

Com relação à dicotomia metodológica entre métodos quantitativos e qualitativos, pode não haver total clareza de diferenciação entre estes métodos,

> A pesquisa moderna deve rejeitar como uma falsa dicotomia a separação entre estudos qualitativos e quantitativos, ou entre ponto de vista estatístico e não estatístico. Além disso, não importa quão precisas sejam as medidas, o que é medido continua a ser uma qualidade. (GOODE; HATT, 1973 apud RICHARDSON, 2010, p. 79).

Nesta parte da pesquisa, ainda que tenham sido utilizados essencialmente métodos quantitativos com estatística básica nos procedimentos (mais adiante apresentados), as características sociais do problema explorado (língua e língua escrita) também exigem tratamento qualitativo,

> O aspecto qualitativo de uma investigação pode estar presente até mesmo nas informações colhidas por estudos essencialmente quantitativos, não obstante perderem seu caráter qualitativo quando são transformadas em dados quantificáveis, na tentativa de se assegurar a exatidão no plano dos resultados. (RICHARDSON, 2010, p. 79).

Caracterizada a pesquisa num primeiro nível, abordamos na sequência a questão do escopo de pesquisa e os procedimentos específicos adotados.

12.2 Escopo da pesquisa

Primeiro, a perspectiva é em relação aos documentos históricos produzidos contemporaneamente que deverão ser recuperados pelas próximas gerações que farão uso de novos estados da língua portuguesa, conforme já defendemos no início desta tese. Por questões práticas, como explicamos adiante, fez parte de nosso *corpus* de pesquisa documentos não contemporâneos, mas sim produzidos na época do império brasileiro (1830).

A referida recuperação é aquela que ocorrerá mediante sistemas de recuperação da informação (SRIs) do tipo informatizado.

Entre os vários tipos de documento, o escopo se limita aos documentos de arquivo avaliados como de guarda permanente. Esta escolha é a mais adequada, pois se tratam dos documentos mais comuns a serem mantidos indefinidamente para recuperação futura, reconhecidos como patrimônio documental.

12.3 Análise documental

Os dados coletados e analisados foram obtidos por intermédio de pesquisa documental num *corpus* constituído de documentos legislativos (decretos) do período imperial brasileiro, na regência de D. Pedro II. Sobre as vantagens do uso de dados documentais nessas condições:

> Nos levantamentos, quando se indaga acerca do comportamento passado, o que se obtém, na realidade, é a percepção do respondente a esse respeito. Já os dados documentais, por terem sido elaborados no período que se pretende estudar, são capazes de oferecer um conhecimento mais objetivo da realidade. (GIL, 2006, p. 166).

O problema específico a ser investigado através do método de pesquisa documental nesta seção foi compreender como a língua portuguesa – registrada em determinado extrato da sociedade brasileira – sofreu os efeitos do fenômeno da mudança linguística, se comparado ao estado atual da mesma língua. Isso também significa compreender em que aspectos da língua se deram as possíveis mudanças. No entanto, não se tratou de uma investigação estritamente linguística, pois não se procuraram possíveis causas da mudança ou um levantamento minucioso de algum fenômeno linguístico específico. Trata-se de uma atividade na qual se procurou caracterizar os elementos que sofreram mudanças (em todos os aspectos possíveis) a fim de estimar os efeitos futuros dessa mudança em SRIs.

Metodologicamente, as condições dos dados utilizados caracterizam essa pesquisa documental como um tipo de pesquisa histórica que possui etapas típicas que podem ser sintetizadas em (I) determinação e adequação dos dados a serem analisados; (II) coleta de dados; (III) descrição dos dados e (IV) análise dos dados (RICHARDSON, 2010).

Ainda que a perspectiva seja explorar os efeitos da ML nos sistemas futuros em relação a RI, na prática, é possível apenas analisar acervos documentais do passado, já concretamente produzidos e disponíveis para este tipo de exame. Os dados obtidos ajudaram a esclarecer a exploração do problema de pesquisa nesta tese.

12.4 Determinação e adequação dos dados

Universo ou população de pesquisa. O universo de pesquisa que atende aos interesses do problema de pesquisa é composto por registros escritos em estados anteriores da língua portuguesa no Brasil. Mesmo considerando que a produção documental no Brasil, se comparada a hoje, era pequena em função das condições sociais e de estrutura dos períodos colonial, império e república velha, há milhares de documentos (talvez milhões) que podem ser analisados em diversos arquivos estaduais e federais. Além do Arquivo Nacional brasileiro, alguns órgãos de maior porte mantêm seus próprios arquivos históricos, como o Senado Federal[1], a Câmara dos Deputados[2] e o Supremo Tribunal Federal[3], além de outros.

Amostragem. Tendo sempre a língua como norte em sua relação com os efeitos da mudança linguística, o mais adequado é utilizar documentos históricos que contenham os mais antigos registros da língua portuguesa, por possuírem a fixação de estados da língua mais afastados em relação ao estado atual. No entanto, é preciso considerar alguns fatores práticos limitadores.

Primeiro, na fase colonial do Brasil, principalmente nos primeiros séculos, havia uma forte influência do português de Portugal por conta do domínio desse país. Segundo, a produção documental nesse período depende de procedimentos especiais de paleografia e diplomática para a correta leitura e interpretação dos textos. A paleografia dispõe de técnicas para leitura de caligrafias antigas e a diplomática dispõe de técnicas para verificar a autenticidade de documentos, notadamente os históricos. O uso dessas técnicas extrapola os limites de recursos e não permitiria uma quantidade suficiente de documentos para análise dentro do prazo disponível.

[1]<http://www12.senado.gov.br/institucional/arquivo/sobre-o-arquivo/historico-do-arquivo-do-senado>.
[2]<http://www2.camara.leg.br/documentos-e-pesquisa/biblarq/o-arquivo>.
[3]<http://www.stf.jus.br/portal/cms/verTexto.asp?servico=sobreStfAcervoArquivo>.

Terceiro, a questão da língua brasileira, de acordo com Silvio Elia, a questão da língua portuguesa no Brasil surgiu no século XIX:

> Para isso concorreram dois fatores: a) independência, que, liberando o país da submissão oficial ao cânone português, permitiu que os brasileiros passassem a cuidar por si mesmos dos problemas relativos à língua herdada; b) o movimento romântico que buscava na alma do povo as bases da cultura nacional. (ELIA, 2003, p. 139).

O Brasil independente de Portugal, início do século XIX, significa um país com mais de três séculos e vida política consolidada (1822 com D. Pedro I e 1889 com a proclamação da República). Esse é um período propício para análise em termos de representatividade da amostra documental, ou seja, comparação da língua portuguesa brasileira. Mas, mesmo nesse período, os documentos típicos de arquivo nas instituições ainda eram manuscritos[4]. No entanto, há um trabalho de (re)compilação da leis do império impresso em tipografia em edição original de 1872 publicado pela Imprensa Nacional[5], o que contribui com outra característica importante na representatividade da amostra aqui considerada: sua autenticidade. Nessa publicação da Imprensa Nacional, há decretos, cartas régias e leis desde 1808[6]. Além disso, esse material está atualmente digitalizado, o que significa acesso simplificado no trabalho de campo. Se se considerar a diferença entre os dois anos delimitadores do período e sua média, obtém-se [(1889 - 1822) / 2 = **33**]. Assim, o ano de 1833 se mostra como um bom representante amostral. Entre os diversos documentos disponíveis, os decretos tratam de vários aspectos da vida social brasileira naquele período. Assim, os documentos de arquivo *Decretos do Império* no ano de 1833 são a escolha de *amostra documental*.

12.5 Detalhes da amostra

O volume original considerado “Collecção Leis do Imperio 1833” contém outros documentos legais, além dos decretos da amostra. Assim, para simplificação e maior objetividade nos trabalhos, preparamos outro arquivo[7] com os decretos considerados na amostra.

---

[4]Como exemplo, o fundo dos documentos de arquivo da Constituição de 1824 é constituído por várias séries de documentos manuscritos, vide: <https://arquivohistorico.camara.leg.br/index.php/assembleia-geral-consituinte-e-legislativa-do-imperio-do-brasil-1823>

[5]Esse documento digitalizado corresponde ao arquivo Leis 1833 parte1.pdf, gravado no disco anexo.

[6]Vide: <http://www2.camara.leg.br/atividade-legislativa/legislacao/publicacoes/doimperio>.

[7]Arquivo Decretos 1833.pdf, gravado no disco anexo.

A numeração desses decretos inicia em 1(um) e continua até 67(sessenta e sete), no entanto a prática de numeração na época era manter a mesma sequência de numeração tanto para decretos, como leis e outros atos legais. Além disso, alguns decretos não estão numerados. Para obter maior homogeneidade na amostra (todos os decretos com as mesmas características), foram omitidos os demais atos legais nesta numeração e também os decretos não numerados. Dessa forma, há na verdade 53 decretos na amostra, os quais serão referenciados como documento 1 a 53. A tabela 4 relaciona esses documentos.

De acordo com a tabela 4, pode-se observar que não existe um decreto de número 21 a 24, 31, 46, 48, 52 a 53, 57 a 59, 62 e 66, quatorze decretos no total (67 – 14 = 53 documentos).

Para algumas análises, foi útil, como se verá adiante, trabalhar com versões em texto dos documentos digitalizados. Assim, foram produzidos arquivos[8] nos formatos .doc e .txt com versões em texto do arquivo em pdf. Esses conteúdos textuais foram produzidos a partir da imagem digitalizada original e passaram por correções para adequação que implica manter a ortografia (e acentuação/pontuação) dos originais.

Tabela 4 – Decretos analisados

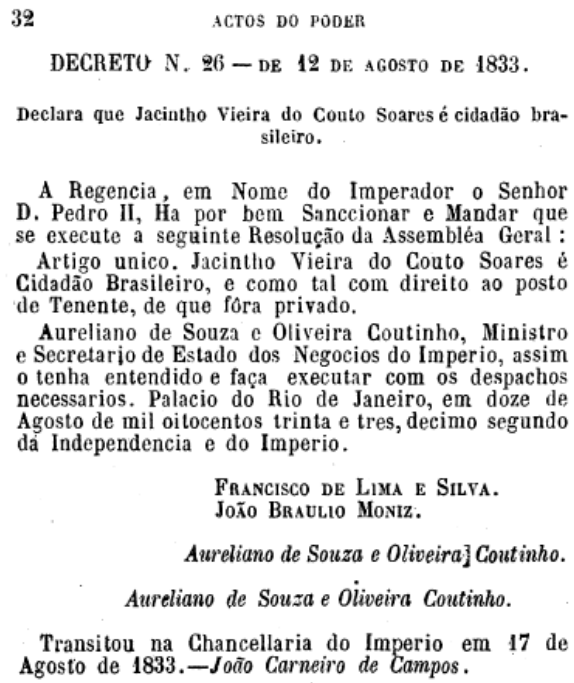
doc	Dec	doc	dec	doc	dec	doc	dec
1	1	16	16	31	36	46	55
2	2	17	17	32	37	47	56
3	3	18	18	33	38	48	60
4	4	19	19	34	39	49	61
5	5	20	20	35	40	50	63
6	6	21	25	36	41	51	64
7	7	22	26	37	42	52	65
8	8	23	27	38	43	53	67
9	9	24	28	39	44		
10	10	25	29	40	45		
11	11	26	30	41	47		
12	12	27	32	42	49		
13	13	28	33	43	50		
14	14	29	34	44	51		
15	15	30	35	45	54		

Fonte: Elaboração própria.

[8]Adiante, serão indicados os arquivos específicos utilizados.

A *figura 10* ilustra a página 32 do documento original com a imagem digitalizada. Nos arquivos em texto, a numeração da página (no exemplo 32), o cabeçalho “ACTOS DO PODER” e a frase de encerramento que começa em “Transitou na Chancellaria...” foram omitidas a fim de padronizar, porque esses dados são comuns a todos os documentos. De maneira que os arquivos textuais começam em “DECRETO N. [...] e finalizam no último nome que corresponde a quem assinou por último o documento.

Figura 10 – Amostra de decreto



32 ACTOS DO PODER

DECRETO N. 26 — DE 12 DE AGOSTO DE 1833.

Declara que Jacintho Vieira do Couto Soares é cidadão brasileiro.

A Regencia, em Nome do Imperador o Senhor D. Pedro II, Ha por bem Sanccionar e Mandar que se execute a seguinte Resolução da Assembléa Geral:

Artigo unico. Jacintho Vieira do Couto Soares é Cidadão Brasileiro, e como tal com direito ao posto de Tenente, de que fôra privado.

Aureliano de Souza e Oliveira Coutinho, Ministro e Secretario de Estado dos Negocios do Imperio, assim o tenha entendido e faça executar com os despachos necessarios. Palacio do Rio de Janeiro, em doze de Agosto de mil oitocentos trinta e tres, decimo segundo da Independencia e do Imperio.

FRANCISCO DE LIMA E SILVA.
JOÃO BRAULIO MONIZ.

Aureliano de Souza e Oliveira] Coutinho.

Aureliano de Souza e Oliveira Coutinho.

Transitou na Chancellaria do Imperio em 17 de Agosto de 1833.—*João Carneiro de Campos.*

Fonte: Elaboração própria a partir da publicação Leis do Império.

12.6 Recursos computacionais

É possível utilizar recursos computacionais para análise de documentos com texto, automatizando, por exemplo, a análise lexical, sintática e semântica. O uso desses recursos permite análises mais detalhadas e precisas. Na amostra documental, serão utilizados recursos computacionais como auxílio das análises que serão efetuadas.

Tais recursos computacionais são produtos da linguística computacional e também da área de processamento de linguagem natural (PLN). Um aspecto importante nessas ferramentas de *software* é a adequação a determinada língua vernácula. A análise lexical e morfológica para uma determinada palavra ou lexema, na terminologia daquelas áreas, só pode apresentar resultados satisfatórios se as características específicas da língua inglesa, italiana ou portuguesa, para citar três exemplos, forem consideradas. Assim, o “o”

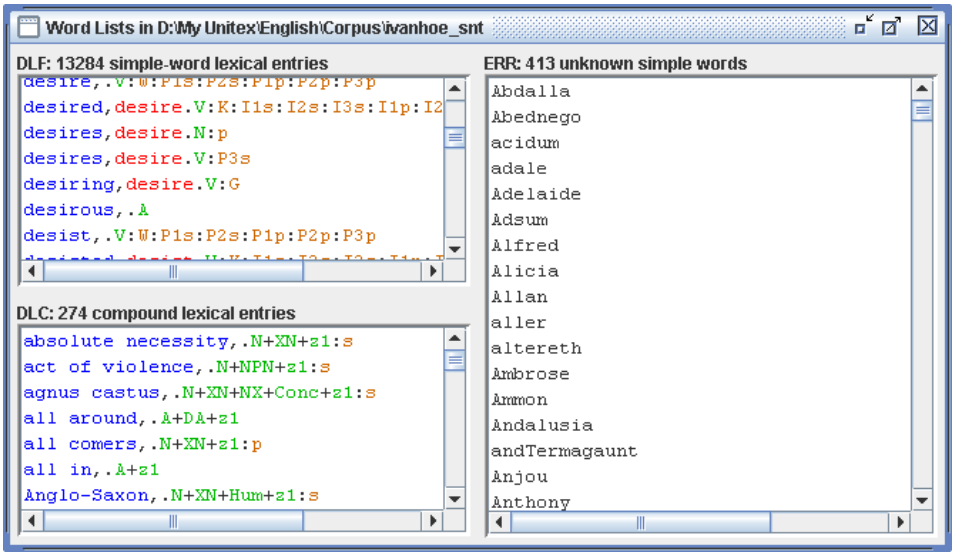
tanto pode ser um artigo em português como possuir o sentido do “ou” português em outras línguas. Infelizmente, não há muitos recursos computacionais para processamento de documentos textuais disponíveis em língua portuguesa, menos ainda se forem consideradas as desenvolvidas no Brasil. O Núcleo Interinstitucional de Linguística Computacional (NILC) da Universidade de São Paulo (USP) mantém uma relação de recursos e ferramentas de *software* com suporte para a língua portuguesa, inclusive na versão brasileira.

Os recursos computacionais podem ser classificados em dois grupos: (I) recursos de conhecimento linguístico e (II) recursos para processar uma língua (MUNIZ, 2004). No primeiro estão, por exemplo, os dicionários eletrônicos e os thesaurus eletrônicos. No segundo, estão “etiquetadores” (analisam partículas do texto para análise morfológica, sintática ou semântica). Todos eles são utilizados de maneira (inter)dependente, um apoiando e aprimorando os demais processos, de maneira que um dicionário fornece informações para a análise morfológica de um determinado verbo em português do Brasil.

Entre os recursos computacionais disponíveis pelo NILC da USP e testados, a ferramenta UNITEX[9], desenvolvida originalmente por laboratórios franceses e uma universidade[10], mostrou bons resultados, pois possui suporte para o português do Brasil, além de se tratar de um *software* livre e de fácil instalação, com manual de operação completo disponível[11]. O recurso vem sendo desenvolvido com a colaboração de pesquisadores no mundo todo, inclusive brasileiros e portugueses que trabalham com partes específicas de nossa língua. A figura 11 mostra uma das telas do aplicativo, na qual se pode ver uma relação de palavras encontradas em determinado documento analisado (lado direito) e, no lado esquerdo acima, a relação de itens simples (*simple-word*) morfologicamente analisados e, no lado esquerdo embaixo, a relação de itens compostos (*compound lexical entries*) examinados também no aspecto morfológico.

[9]<http://www-igm.univ-mlv.fr/~unitex/index.php?page=0>.
[10]<http://infolingu.univ-mlv.fr/>.
[11]<http://www-igm.univ-mlv.fr/~unitex/index.php?page=3>.

Figura 11 – Tela aplicativo UNITEX



Fonte: Aplicativo UNITEX.

Na seção seguinte, são apresentados os resultados obtidos através do aplicativo UNITEX nos documentos de 1 a 53, que compõem a amostra.

12.7 Dados obtidos via UNITEX

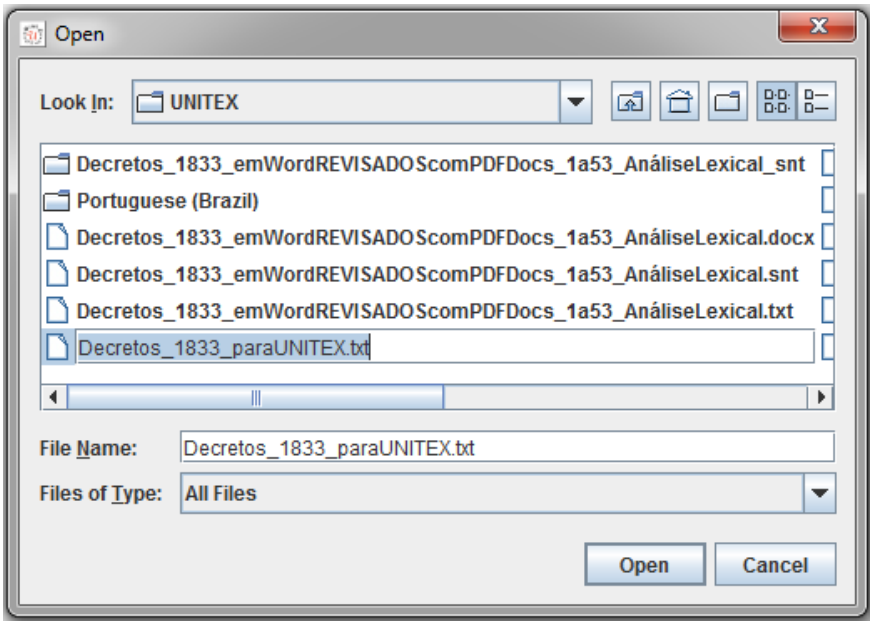
Os primeiros dados a serem obtidos são aqueles do aspecto lexical do registro de informações nos documentos da amostra. O ponto de partida é um arquivo[12] em texto não formatado (.txt) contendo, num único arquivo, todas as informações em todos os documentos da amostra. Esse arquivo será o ponto de partida da análise no aspecto lexical.

12.7.1 Dados brutos inicias

O primeiro passo foi a importação do arquivo do tipo .txt pelo aplicativo UNITEX (que permite importar outros formatos além do TXT, como o HTML ou XML). A *figura 12* ilustra essa ação.

[12] Arquivo Decretos_1833_paraUNITEX.txt, gravado no disco anexo à tese.

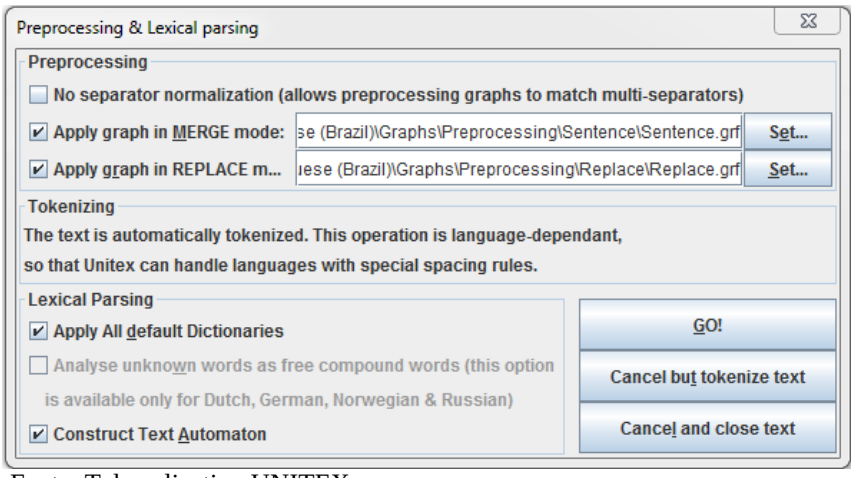
Figura 12 – UNITEX abrindo arquivo



Fonte: Tela aplicativo UNITEX

Ao abrir o arquivo, o aplicativo UNITEX faz um pré-processamento no conteúdo textual, conforme *figura 13*. Várias operações de computação linguística são realizadas sobre o texto, os detalhes podem ser consultados no manual completo do aplicativo (PAUMIER, 2003).

Figura 13 – Pré-processamento UNITEX



Fonte: Tela aplicativo UNITEX

De todas as operações de linguística computacional, interessa em especial o conceito de *token*, como tratado pelo aplicativo, ou seja, qualquer caractere numérico ou não, algarismos (*digits*), espaços em branco, marcas de parágrafo, pontuação ou qualquer outro símbolo como números ordinais, cifrão, inclusive sentenças são consideradas como *tokens*. O aplicativo estabelece a relação respectiva de *tokens* que recebem uma identificação única.

Veja-se um exemplo com a sentença "Decreto nº 2, de 18 de junho de 1833". Para esse exemplo, considera-se uma identificação hipotética de *tokens* a partir do número 0.

Decreto	[espaço]	n	º	2	,	[espaço]	de	Junho	[espaço]	de	1833
0	1	2	3	3		1	4	5	1	4	6

Na *tabela 5*, que organiza os *tokens* em ordem de identificação, a sentença analisada no exemplo hipotético ficará:

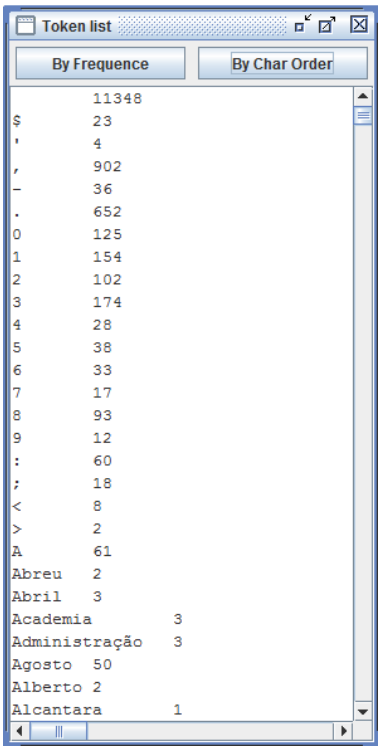
Tabela 5 – Tokens em ordem de identificação

Token ID	Token	Ocorrências
0	Decreto	1
1	[espaço]	3
2	n	1
3	º	1
4	de	2
5	Junho	1
6	1833	1

Fonte: Elaboração própria.

No exemplo da tabela 5, há portanto, sete *tokens*, cinco deles ocorrem uma única vez (*tokens* 0, 2, 3, 5 e 6), um ocorre duas vezes (*token* 4) e outro ocorre três vezes (*token* 1) na sentença considerada. A *figura 14* mostra a janela do aplicativo com os primeiros *tokens* da lista. O primeiro item são os espaços em branco, seguidos de símbolos, pontuação, números arábicos até o início do texto (em ordem alfabética), os números à direita correspondem à ocorrência de cada *token* identificado.

Figura 14 – Tokens no UNITEX



Fonte: Aplicativo UNITEX

Para outros tratamentos, o aplicativo UNITEX gera vários arquivos com os resultados dos produtos de análise do texto original. Nesse momento, o arquivo *tok_by_alph.txt* será útil, pois contém a relação de todos os *tokens*, em ordem alfabética gerada sobre o arquivo original com os documentos 1 a 53. Esse arquivo foi importado para o aplicativo Excel (manipulação de planilhas eletrônicas) a fim de fazer as manipulações necessárias. O arquivo em Excel correspondente é *tokensUNITEX.xlsx*. Com manipulações e análises neste arquivo, foi derivada a *tabela 6*.

Tabela 6 – Tokens e ocorrências

tipo de token	itens	ocorrências
maiúsculas	502	4067
minúsculas	867	6886
espaços em branco	1	11348
símbolos 1	5	1617
dígitos	10	776
símbolos 2	4	88
sentenças	1	491
símbolos 3	5	732
outros	5	17
total	1400	**26022**

Fonte: Elaboração própria.

A *tabela 6* mostra que no total há 26.022 ocorrências de 1.400 tipos de *tokens*.

São considerados *tokens* espaços em branco, sentenças (estudadas nas próximas seções), dígitos, símbolos e palavras em geral (separadas em maiúsculas e minúsculas) a fim de facilitar a identificação da classe de substantivos. Há também um grupo de palavras com acento grave, no caso específico dos textos analisados. Os detalhes de todos os *tokens* obtidos estão no arquivo *tokensUNITEX.xlsx (na aba tokensFromUnitex).* Estes dados não podem ser aqui inseridos pois possuem mais de 1500 linhas, o que significa umas setenta páginas impressas. Como ilustração a figura 16 contem um extrato das primeiras trinta e uma linhas e a figura 15 exibe as cores para identificação visual dos diferentes tipos de *tokens* na figura 16.

Figura 15 – Legenda *tokens* arquivo tokensUNITEX.xlsx

tipo de token	cores
maiúsculas	A
minúsculas	a
error words	á
DLF entries	a,.ABREV:ms
espaços em branco	
símbolos 1	$
dígitos	0
símbolos 2	:
sentenças	{S}
símbolos 3	ª
outros	ácerca

Fonte: Elaboração própria.

Nesse ponto, é fundamental compreender algumas noções sobre como funcionam os dicionários incorporados ao UNITEX e utilizados nos processamentos de texto. Um dicionário eletrônico no contexto de recursos computacionais linguísticos (PLN) como o UNITEX possui uma definição diferente em relação a dicionários tradicionais. Basicamente, cada entrada no dicionário do UNITEX aqui utilizado corresponde a uma ocorrência de um termo e sua morfologia básica (flexões verbais, gênero, contrações e associações entre preposição e artigo). O dicionário utilizado nas análises aqui efetuadas foi criado por uma equipe brasileira (MUNIZ; NUNES; LAPORTE, 2008). O trecho seguinte exemplifica algumas entradas de termos presentes nos documentos da amostra analisada com correspondência no dicionário português brasileiro utilizado no UNITEX (quadro 7).

Quadro 7 – Correspondências no dicionário utilizado

baixo,.A:ms	Adjetivo (masculino singular)
baixo,.ADV	advérbio
baixo,.N:ms	Substantivo (masculino singular)
baixo,baixar.V:P1s	Verbo (primeira pessoa singular)
banco,.N:ms	Substantivo (masculino singular)
banco,bancar.V:P1s	Verbo (primeira pessoa singular)
barca,.N:fs	Substantivo (feminino singular)
barcas,barca.N:fp	Substantivo (feminino plural)
barco,.N:ms	Substantivo (masculino singular)
barcos,barco.N:mp	Substantivo (masculino plural)

Fonte: Elaboração própria.

Ainda nas análises da planilha em Excel (figura 16), é possível identificar que, de todos os *tokens* o conteúdo analisado possui 1.549 entradas com correspondência no dicionário[13] e 338 não estão nesse dicionário (não possuem correspondência). Em geral, nesses 338 *tokens* fora do dicionário, há palavras com grafia antiga, sem acentuação correta, alguns nomes. Esses números podem ser conferidos na primeira aba do arquivo *tokensUNITEX.xlsx* nas colunas da esquerda (DLF entries = no dicionário e Error words = fora do dicionário), vide também figura 16 os primeiros valores da planilha.

[13]note-se que está na contagem cada possível ocorrência, então "banco" (vide quadro 7) possui duas entradas identificadas e espaços e sentenças não possuem, naturalmente, correspondência.

Figura 16 – Resumo geral tokens obtidos

Tipos de Tokens e Ocorrências (#)						Correspondência Dicionário	
maiúsculas	#	minúsculas	#	símbolos	#	Error Words (338)	DLF entries (1549)
A	61	a	231	[espaço branco]	11348	á	a,.ABREV:ms
Abreu	2	abaixo	5	$	23	accionistas	a,.N:ms
Abril	3	abertura	2	'	4	accomodar	a,.PREP
Academia	3	abonar	2	,	902	accôrdo	a,ele.PRO+Pes:A3fs
Administração	3	abrangendo	2	-	36	ácerca	a,o.DET+Art+Def:fs
Agosto	50	absolutamente	1	.	652	acto	a,o.PRO+Dem:fs
Alberto	2	abusos	1	0	125	actualmente	abaixo,.ADV
Alcantara	1	accionistas	2	1	154	additamento	abaixo,abaixar.V:P1s
Aldêa	1	accomodar	1	2	102	admittir	abertura,.N:fs
Alfandega	4	accôrdo	1	3	174	agrarios	abertura,aberturar.V:P3s:Y2s
Alfandegas	2	acertado	1	4	28	ahi	abonar,.V:W1s:W3s:U1s:U3s
Alferes	2	acha	2	5	38	Alcantara	abrangendo,abranger.V:G
Almas	4	achar	1	6	33	Aldêa	abreu,.N+Pr:ms:fs
Alto	1	acharem	2	7	17	aldêa	abril,.N:ms
Alves	2	ache	1	8	93	aldêamento	absolutamente,.ADV
Além	1	acima	3	9	12	algebra	abusos,abuso.N:mp
Amalia	2	acto	4	:	60	alli	academia,.N:fs
Andrade	3	actualmente	3	;	18	allodiaes	academia,academiar.V:P3s:Y2s
Anjos	1	adaptados	1	<	8	alteral	acertado,acertar.V:K
Anna	7	additamento	2	>	2	alumnos	acha,.N:fs
Antas	2	adiar	1	{S}	491	Amalia	acha,achar.V:P3s:Y2s
Antero	4	administradas	1		563	analogos	achar,.N:ms
Antonio	17	administração	1	ª	2	animaes	achar,.V:W1s:W3s:U1s:U3s
Antéro	4	admittir	4	º	8	anno	acharem,achar.V:W3p:U3p
Aos	1	agora	1	º	159	annos	ache,achar.V:S1s:S3s:Y3s
Apiacá	1	agrarios	1	á	19	annuaes	acima,.ADV
Approva	12	ahi	1	ácerca	1	annual	acima,acimar.V:P3s:Y2s
Aquiraz	1	ainda	1	áquella	1	annualde	adaptados,adaptado.A:mp
Aracaty	1	aldeamento	1	ás	7	annualmente	adaptados,adaptar.V:K
Araras	1	aldeamentos	1	é	7	Antéro	adiar,.V:W1s:W3s:U1s:U3s
Araujo	13	aldêa	1	único	1	Apiacá	administração,.N:fs
continua...		continua...				continua...	continua...

(Primeiras 31 Linhas)

Fonte: Elaboração própria em planilha excel.

12.7.2 Dados aspecto lexical

O total de 1.887 *tokens* (1.549 + 338) serão o objeto principal de interesse nas análises linguísticas, especialmente em relação à mudança linguística, e serão chamados de *tokens* relevantes. Criou-se uma aba separada para esses *tokens* (no mesmo arquivo *tokensUNITEX.xlsx*) com o nome *tokensRelevantes* e inclui tanto aqueles termos iniciados em maiúsculas como em minúsculas, foram omitidos símbolos, dígitos e espaços em branco. Como esta aba no arquivo original possui mais de 1500 linhas, extraímos uma tabela com as linhas iniciais apenas para ilustração (tabela 7).

Tabela 7 – Tokens relevantes (maiúsculas e minúsculas).

maiúsculas/ocorrências		DIC	minúsculas/ocorrências		DIC
A	61	0	a	231	0
Abreu	2	0	abaixo	5	0
Abril	3	0	abertura	2	0
Academia	3	0	abonar	2	0
Administração	3	0	abrangendo	2	0
Agosto	50	0	absolutamente	1	0
Alberto	2	0	abusos	1	0
Alcantara	1	1	accionistas	2	1
Aldêa	1	1	accomodar	1	1
Alfandega	4	0	accôrdo	1	1
Alfandegas	2	0	acertado	1	0
Alferes	2	0	acha	2	0
Almas	4	0	achar	1	0
Alto	1	0	acharem	2	0
Alves	2	0	ache	1	0
Além	1	0	acima	3	0

Fonte: Elaboração própria.
Nota: 1 (em vermelho) significa item sem correspondência no dicionário UNITEX.

Observe-se aqui que não pertencer ao dicionário não implica necessariamente que a palavra não exista (erro de digitação) ou não tenha correspondência em dicionários tradicionais, no caso do tipo dicionário histórico, já que estes termos eram utilizados no período do império brasileiro. Nem todos os nomes próprios estão fora do dicionário como no exemplo da tabela 7 para o termo Alcantara.

Uma primeira análise é calcular a razão entre os termos fora do dicionário e aqueles com correspondência no dicionário UNITEX (para o total e para as ocorrências de

termos). As análises numéricas sobre esses dois grupos de *tokens* podem ser verificadas na *tabela 8*, que resume os dados.

Tabela 8 – Tokens relevantes

	Maiúsculas	Ocorrências	Minúsculas	Ocorrências
total	502	4.067	867	6.886
fora DIC	146	992	189	506
% total	29,08%	24,39%	21,80%	7,35%

Fonte: Elaboração própria.

Com base nos dados da tabela 8, a maior parte das ocorrências que estão fora do dicionário eletrônico está no grupo das maiúsculas, 29,08% em relação ao total de maiúsculas consideradas.

Notar que as ocorrências totais de maiúsculas e minúsculas incluir várias classes gramaticais, como substantivos, adjetivos, advérbios e verbos. A fim de possibilitar tratamentos mais precisos criamos a aba *tokensSubstantivos* no arquivo *tokensUNITEX.xlsx* que inclui apenas as ocorrências da classe substantivos, padronizamos todos os elementos iniciando em minúsculas.

Com isso, encerra-se a exploração dos dados no aspecto lexical, nas próximas seções esses dados serão retomados.

12.7.3 Dados aspecto morfológico

Para a análise dos dados obtidos a partir do UNITEX, no aspecto morfológico, foi criada a planilha (aba) **tokensMorfológico** no mesmo arquivo Excel originalmente criado.

Há no total **949** *tokens* retirados da lista tanto de maiúsculas como de minúsculas, exceto aqueles *tokens* que não possuem correspondência no dicionário eletrônico, já que isso é essencial para que ocorra a análise morfológica em si. Notar que é o dicionário que fornece identificações como classe gramatical e número (singular e plural). A nova lista não distingue maiúsculas ou minúsculas. Esses *tokens* podem se desdobrar em várias linhas de análise, já que pode pertencer a várias classes gramaticais. Por exemplo, o *token* "a" desdobra-se em seis análises morfológicas possíveis (quadro 8). No total, ocorreram **1.549** linhas.

Quadro 8 – Token "a"

a,.ABREV:ms	Abreviação
a,.N:ms	Substantivo masculino singular
a,.PREP	preposição
a,ele.PRO+Pes:A3fs	Pronome pessoal, feminino singular
a,o.DET+Art+Def:fs	Artigo definido, feminino singular
a,o.PRO+Dem:fs	Pronome demonstrativo feminino singular

Fonte: Elaboração própria.

As possibilidades de desdobramento são calculadas por algoritmos de processamento de linguagem natural (PLN).

Como apresentado na seção anterior (lexical), há **338** *tokens* do total de elementos linguísticos analisados que não estão no dicionário, portanto não podem ser morfologicamente estudados. A aba *tokensMorfológico* no arquivo *tokensUNITEX.xlsx* contém todas as análises morfológicas extraídas do aplicativo UNITEX.

Dos **1.549** *tokens* com associação ao dicionário, a análise morfológica identifica verbos, substantivos, adjetivos, advérbios, artigos e preposições. Para o caso dos substantivos, por exemplo, foi criada uma planilha (aba) específica de nome **TokensSubstantivos** na planilha Excel original. Nela, podem-se visualizar 636 substantivos classificados em ms (masculino singular), mp (masculino plural), fs (feminino singular), fp (feminino plural), Pr+ms (prenome masculino singular), Pr+fs (prenome feminino singular) e Pr+ms:fs (prenome masculino singular e feminino singular). Detalhes específicos podem ser conferidos na planilha disponível.

12.7.4 Dados aspecto sintático

O aspecto sintático das informações registradas na amostra documental considerada refere-se às sentenças identificadas pelo aplicativo. As regras para identificação das sentenças baseiam-se principalmente na identificação do ponto final "." e do caractere de parágrafo. Assim, sentenças como:

{S}Aureliano de Souza e Oliveira Coutinho. (última linha no documento 1) ou

{S}Palácio do Rio de Janeiro, em dezoito de junho de mil oitocentos trinta e três, decimo segundo da independência e Imperio. (última sentença da finalização no documento 3)

São sentenças identificadas com base nessas regras. O símbolo {S} é inserido no começo de cada sentença identificada. No total, há 491 identificadores, mas, como a primeira sentença identificada não possui esse identificador, então são 492 sentenças identificadas.

A fim de obter uma melhor visualização e análise, importou-se do aplicativo UNITEX um arquivo .DOC e foram destacados todos os 491 identificadores de sentenças em vermelho e negrito no arquivo *SentençasDestacadas_491.doc*, gravado em disco no anexo. Note-se que esses 491 identificadores de sentenças não significam que há 491 sentenças no sentido de uma análise sintática tradicional, já que podem surgir erros no processamento. De fato, analisando o arquivo *SentençasDestacadas_491.doc* com os identificadores de sentença, há várias que não foram explicitamente identificadas. Para incluir todas as sentenças possíveis seria necessário fazer modificações no documento original para sanar os casos em que o aplicativo UNITEX não resolveu corretamente alguns trechos. Como exemplo, observem-se as sentenças seguintes (quadro 9):

Quadro 9 – Sentenças analisadas UNITEX

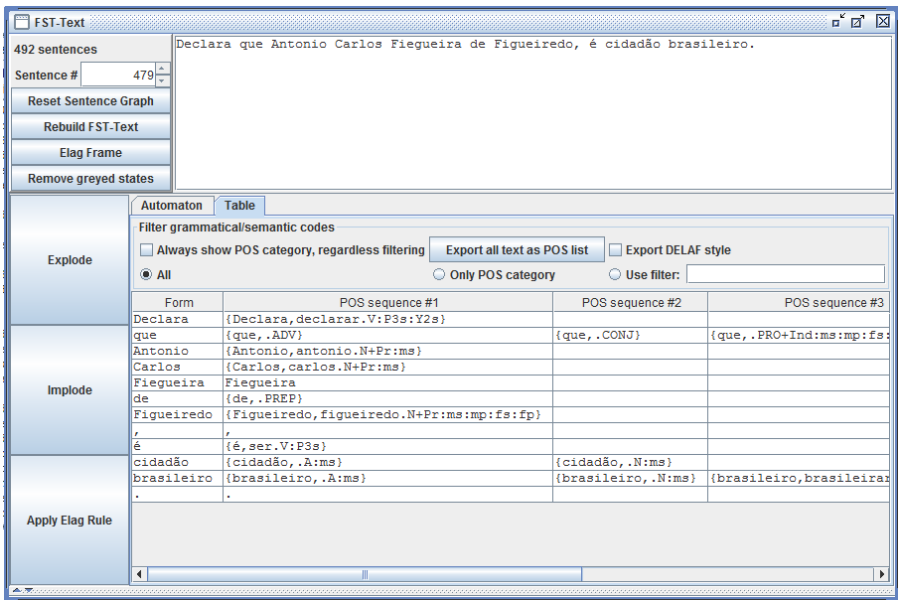
474.{S}Palacio do Rio de Janeiro, em dez de Outubro de mil oitocentos trinta e tres, decimo segundo da Independencia e do Imperio.
475.{S}FRANCISCO DE LIMA E SILVA.
476.{S}João Braulio Moniz.
477.{S}Aureliano de Souza e Oliveira Coutinho.
478.{S}Decreto nº 65, de 10 de Outubro de 1833
479.{S}Declara que Antonio Carlos Fiegueira de Figueiredo, é cidadão brasileiro.
480.{S}A Regencia Permanente, em Nome do Imperador o Senhor D. Pedro II, Ha por bem Sanccionar e Mandar que se execute a seguinte Resolução da Assembléa Geral:
Artigo unico.
481.{S} Antonio Carlos Figueira de Figeiredo é cidadão brasileiro, na conformidade do artigo sexto paragrapho segundo do Titulo segundo da Constituição Politica do Imperio.

Fonte: Elaboração própria.

No exemplo anterior, nota-se que, entre a sentença 480 e 481, deveria haver outra igualmente identificada. O problema ali ocorrido pode ter sido causado pelos dois pontos ":" anteriores a "Artigo unico." ou por vários outros fatores indetermináveis. De qualquer forma, para efeitos da análise efetuada, isso não deve afetar os resultados, pois é possível exemplificar os procedimentos com base nas sentenças corretamente identificadas. A tarefa de correção das sentenças não identificadas pode ser necessária para análises de um ponto de vista estritamente linguístico.

O processo de identificação de sentenças é feito para que seja possível analisa-las em seus elementos constituintes. A *figura 17* ilustra essa análise para a sentença 479.

Figura 17 – Tela UNITEX exemplificada (479)



Fonte: Aplicativo UNITEX.

A *figura 17* mostra todos os componentes da sentença considerada, o elemento "Fiegueira" não está analisado, pois não possui correspondente no dicionário de português brasileiro utilizado. Trata-se do mesmo motivo da falta de análise morfológica em alguns elementos na seção anterior.

12.7.5 Dados aspecto semântico-pragmático

O aplicativo UNITEX pode explorar algumas informações semânticas, mas elas devem estar inseridas nas entradas dos dicionários associados à análise. O quadro 10 ilustra alguns possíveis códigos para informações semânticas com exemplos que ajudam a contextualizar o uso (aspecto pragmático).

Quadro 10 – Códigos para informações semânticas

Code	Description	Example
z1	general language	joke
z2	specialized language	floppy disk
z3	very specialized language	serialization
Abst	abstract	patricide
Anl	animal	horse
AnlColl	collective animal	flock
Conc	concrete	chair
ConcColl	collective concrete	rubble
Hum	human	teacher
HumColl	collective human	parliament
T	transitive verb	kill
I	intransitive verb	agree

Fonte: Manual UNITEX.

Um exemplo, retirado do manual UNITEX, é:

bottle,.N+Conc:s (onde "bottle" (garrafa) é um substantivo concreto no singular (:s)

Do ponto de vista semântico, o dicionário utilizado não apresenta significados com explicações mais detalhadas, como encontramos num dicionário tradicional. Este fato limita bastante as análises neste aspecto da língua. Nas informações analisadas a partir dos textos nos documentos da amostra a única informação semântica foi Pr (Prenome), como em "Pedro":

Pedro{Pedro,pedro.N+Pr:ms}

No exemplo, o *token* Pedro é apresentado em duas formas (com maiúsculas e sem maiúsculas e indica-se que é um prenome.

Nas seções seguintes, são apresentadas análises dos dados até aqui destacados.

12.8 Descrição dos dados

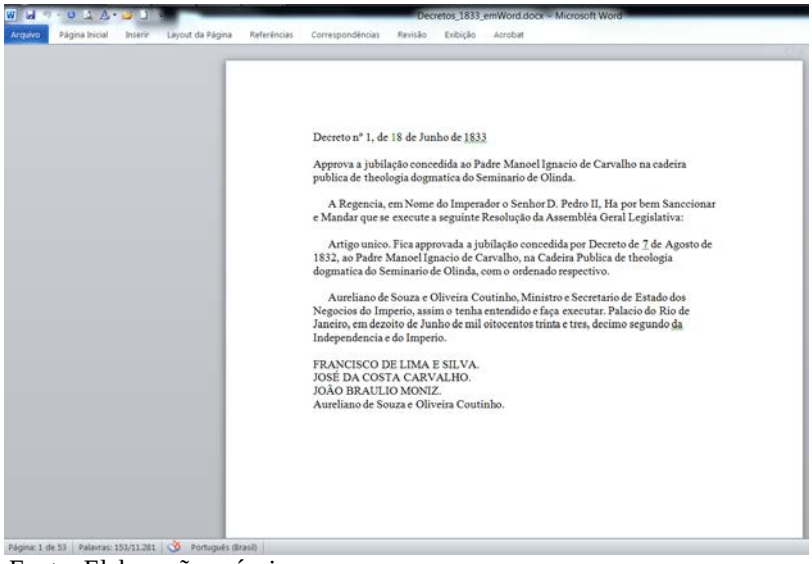
A seguir, fazem-se análises mais detidas tanto com base nos dados já obtidos através do aplicativo UNITEX como também com novos dados obtidos nesta seção. Também serão utilizados os conceitos apresentados no quadro sinóptico apresentado no final da revisão de literatura.

12.8.1 Aspecto lexical

O aspecto lexical é particularmente útil para a caracterização da mudança linguística de maneira quantitativa, pois permite agrupamentos e comparações entre palavras utilizadas de maneira bastante ampla, como se verá a seguir.

A primeira ação de análise consiste em verificar, para cada um dos 53 documentos na amostra, a quantidade de palavras presentes, excetuados os símbolos de pontuação e espaços em branco. Para isso, utilizou-se uma versão em texto de cada um dos documentos no no arquivo *Decretos_1833_emWord.docx* (também presente no disco em anexo). Esse formato eletrônico de arquivo permite a contagem precisa das palavras presentes. A *figura 18* ilustra essa ação através do aplicativo MS-Word, mas, de fato, praticamente qualquer aplicativo editor de texto permite essa ação.

Figura 18 – Exemplo contagem palavras



Fonte: Elaboração própria.

No exemplo da *figura 18*, o documento número 1 (decreto n. 1 de 18 de junho de 1833) contém 153 palavras no total (canto direito inferior da figura). Note-se que em relação ao documento original digitalizado (vide arquivo *Decretos 1833.pdf*), aqui se considera apenas até a última assinatura em cada decreto. A *tabela 9* resume os dados encontrados.

Tabela 9 – Dados encontrados na pesquisa

doc n.	palavras	doc n.	palavras	doc n.	palavras	doc n.	palavras
1	153	16	193	31	211	46	365
2	255	17	186	32	180	47	170
3	151	18	193	33	194	48	179
4	321	19	247	34	165	49	157
5	236	20	155	35	367	50	194
6	163	21	216	36	241	51	158
7	161	22	130	37	217	52	133
8	306	23	353	38	267	53	310
9	208	24	206	39	202		
10	221	25	241	40	185		
11	151	26	405	41	242		
12	241	27	149	42	219		
13	202	28	213	43	162		
14	163	29	188	44	213		
15	189	30	162	45	193		

Fonte: Elaboração própria.

O próximo procedimento consistiu em identificar todas as palavras que estão fora da ortografia corrente na língua. Nesse levantamento, desconsideram-se os erros de acentuação de palavras em função do grande risco de serem causados pela própria técnica tipográfica utilizada. Ainda no exemplo da *figura 18*, há seis palavras com registro ortográfico fora do padrão atual, são elas: “approva”, “theologia” (duas ocorrências), “sanccionar”, “assembléa” e “approvada”. O verificador ortográfico do aplicativo ajudou nesta verificação de dados. A tabela 10 apresenta os elementos fora da ortografia para todos os cinquenta e três documentos.

Tabela 10 – Palavras fora da ortografia atual da língua portuguesa

doc n.	fora ortogr.	doc n.	fora ortogr.	doc n.	fora ortogr.	doc n.	fora ortogr.
1	6	16	13	31	4	46	23
2	13	17	11	32	3	47	5
3	5	18	7	33	7	48	11
4	14	19	11	34	5	49	9
5	9	20	9	35	17	50	5
6	5	21	8	36	16	51	4
7	6	22	2	37	10	52	3
8	8	23	14	38	12	53	7
9	11	24	9	39	11		
10	8	25	14	40	19		

continua...

continuação

11	4	26	25	41	8
12	10	27	6	42	6
13	5	28	12	43	7
14	6	29	7	44	12
15	5	30	5	45	9

Fonte: Elaboração própria.

Com os dados obtidos nas tabelas 9 e 10, pode-se obter a razão dos números de palavras fora da ortografia atual em relação ao total de palavras em cada documento.

Por exemplo, no documento 1, há 153 palavras no total (tabela 9) e seis delas estão fora do padrão ortográfico (tabela 10), essas seis palavras correspondem a 3,92% do total no documento 1. A tabela 11 resume esses dados para todos os documentos.

A razão obtida no procedimento anterior permite uma análise nova, no aspecto lexical, sobre a percentagem de mudança linguística obtida em relação ao ano de produção dos documentos e o ano atual, no caso (2015 – 1833) = 182 anos. Ressalte-se que se trata de apenas um dado, que pode ser levado em consideração, mas deve sê-lo em conjunto com todos os demais dados obtidos.

Tabela 11 – Razão de itens fora da ortografia em relação ao total de itens

doc n.	razão	doc n.	razão	doc n.	razão	doc n.	Razão
1	3,92%	16	6,74%	31	1,90%	46	6,30%
2	5,10%	17	5,91%	32	1,67%	47	2,94%
3	3,31%	18	3,63%	33	3,61%	48	6,15%
4	4,36%	19	4,45%	34	3,03%	49	5,73%
5	3,81%	20	5,81%	35	4,63%	50	2,58%
6	3,07%	21	3,70%	36	6,64%	51	2,53%
7	3,73%	22	1,54%	37	4,61%	52	2,26%
8	2,61%	23	3,97%	38	4,49%	53	2,26%
9	5,29%	24	4,37%	39	5,45%		
10	3,62%	25	5,81%	40	10,27%		
11	2,65%	26	6,17%	41	3,31%		
12	4,15%	27	4,03%	42	2,74%		
13	2,48%	28	5,63%	43	4,32%		
14	3,68%	29	3,72%	44	5,63%		
15	2,65%	30	3,09%	45	4,66%		

Fonte: Elaboração própria.

A média obtida é igual a **4,16** e o desvio padrão, igual a **0,01604** (vide aba *AspectoLexical* no arquivo *tokensUNITEX.xlsx* na coluna C em relação a A, no disco em anexo).

Outro dado complementar, talvez mais significativo como indicativo de mudança

linguística, é a razão entre as palavras com alteração ortográfica em relação ao total de palavras do documento menos nomes significativos (nomes de pessoas, coisas, lugares). Por exemplo, “Rio das Antas”, “Goyaz”, “Villa de Alcântara”, “Aureliano de Souza e Oliveira Coutinho” e também números. Efetuar o cálculo da razão sem essas palavras pode ser mais significativo, pois trata-se de palavras que podem não possuir mais uma relação coisa-nome atualmente, como é certamente o caso dos nomes das pessoas, os quais podem ainda ser utilizados, mas apontando para outras pessoas. Além disso, a mudança em nomes pode seguir regras diferentes da mudança linguística em termos gerais. Assim, o arquivo *Decretos_1833_Lexical.doc* apresenta os mesmos documentos originais, excluindo os nomes e números do texto original. As palavras fora da ortografia atual aparecem destacadas em vermelho. A tabela 12 apresenta os dados obtidos calculando a razão fora da ortografia em relação ao total de itens para este último arquivo. Note-se que o número de palavras total em cada documento agora é menor que no primeiro procedimento (coluna doc n.), essa contagem de palavras foi obtida da mesma forma que na contagem do número total de palavras na tabela 9.

Tabela 12 – Análise fora ortografia com texto sem números e nomes

doc	tot.	%	doc	tot.	%	doc	tot.	%	doc	tot.	%
1	110	**5,45%**	16	167	**7,78%**	31	160	**2,50%**	46	333	**6,91%**
2	220	**5,91%**	17	159	**6,92%**	32	135	**2,22%**	47	133	**3,76%**
3	98	**5,10%**	18	166	**4,22%**	33	159	**4,40%**	48	154	**7,14%**
4	270	**5,19%**	19	201	**5,47%**	34	137	**3,65%**	49	127	**7,09%**
5	174	**5,17%**	20	120	**7,50%**	35	328	**5,18%**	50	137	**3,65%**
6	121	**4,13%**	21	185	**4,32%**	36	192	**8,33%**	51	126	**3,17%**
7	125	**4,80%**	22	96	**2,08%**	37	167	**5,99%**	52	98	**3,06%**
8	235	**3,40%**	23	324	**4,32%**	38	207	**5,80%**	53	224	**3,13%**
9	160	**6,88%**	24	180	**5,00%**	39	156	**7,05%**			
10	162	**4,94%**	25	183	**7,65%**	40	157	**12,10%**			
11	125	**3,20%**	26	328	**7,62%**	41	206	**3,88%**			
12	201	**4,98%**	27	121	**4,96%**	42	178	**3,37%**			
13	164	**3,05%**	28	174	**6,90%**	43	116	**6,03%**			
14	127	**4,72%**	29	153	**4,58%**	44	170	**7,06%**			
15	153	**3,27%**	30	122	**4,10%**	45	166	**5,42%**			

Fonte: Elaboração própria.

Pelos dados na tabela 12, pode-se verificar que há um aumento no percentual, já que a razão foi obtida a partir de um número menor (total de palavras = coluna tot.). A média dessa tabela foi de **5,18%** e o desvio padrão, de **0,01872**, conforme aba *AspectoLexical* no arquivo *tokensUNITEX.xlsx* (vide coluna C em relação a B).

Também foi verificada esta mesma razão considerando, no número de palavras fora da ortografia atual também as palavras sem acentuação, as quais como já explicado, podem estar nesse estado em função de erros tipográficos. De qualquer forma, trata-se de mais um dado para a análise. A tabela 13 apresenta os dados. A razão foi obtida em função do número total de palavras em cada documento sem nomes ou números.

Tabela 13 – Análise da ortografia considerando erros de acentuação

doc	tot.	%	doc	tot.	%	doc	tot.	%	doc	tot.	%
1	110	**12,73%**	16	167	**15,57%**	31	160	**9,38%**	46	333	**12,61%**
2	220	**9,09%**	17	159	**13,21%**	32	135	**10,37%**	47	133	**12,03%**
3	98	**8,16%**	18	166	**10,24%**	33	159	**10,06%**	48	154	**13,64%**
4	270	**11,11%**	19	201	**10,95%**	34	137	**10,22%**	49	127	**14,17%**
5	174	**10,34%**	20	120	**14,17%**	35	328	**9,76%**	50	137	**9,49%**
6	121	**9,09%**	21	185	**9,73%**	36	192	**14,06%**	51	126	**9,52%**
7	125	**10,40%**	22	96	**8,33%**	37	167	**11,38%**	52	98	**10,20%**
8	235	**6,81%**	23	324	**9,26%**	38	207	**11,11%**	53	224	**8,04%**
9	160	**12,50%**	24	180	**10,56%**	39	156	**14,74%**			
10	162	**9,88%**	25	183	**15,85%**	40	157	**17,20%**			
11	125	**11,20%**	26	328	**11,28%**	41	206	**9,71%**			
12	201	**9,45%**	27	121	**14,05%**	42	178	**11,80%**			
13	164	**10,37%**	28	174	**14,94%**	43	116	**13,79%**			
14	127	**11,02%**	29	153	**13,73%**	44	170	**13,53%**			
15	153	**13,07%**	30	122	**11,48%**	45	166	**11,45%**			

Fonte: Elaboração própria.

Os dados obtidos nessa última tabela apresentam os maiores valores de média e desvio padrão, **11,45%** e **0,02237** respectivamente, conforme aba *AspectoLexical* no arquivo *tokensUNITEX.xlsx* (vide coluna C + D em relação a B). Esses percentuais se explicam já que há muitas palavras nos documentos analisados em desconformidade com o padrão atual de acentuação gráfica. A figura 19 apresenta os resultados gerais das análises anteriores.

Figura 19 – Resultados gerais de análise lexical/ortográfica

Doc(s) #	Itens considerados		Itens fora padrão		Percentuais		
	A	B	C	D	C em relação a A	C em relação a B	(C + D) em relação a B
1	153	110	6	8	3,92%	5,45%	12,73%
2	255	220	13	7	5,10%	5,91%	9,09%
3	151	98	5	3	3,31%	5,10%	8,16%
4	321	270	14	16	4,36%	5,19%	11,11%
5	236	174	9	9	3,81%	5,17%	10,34%
6	163	121	5	6	3,07%	4,13%	9,09%
7	161	125	6	7	3,73%	4,80%	10,40%
8	306	235	8	8	2,61%	3,40%	6,81%
9	208	160	11	9	5,29%	6,88%	12,50%
10	221	162	8	8	3,62%	4,94%	9,88%
11	151	125	4	10	2,65%	3,20%	11,20%
12	241	201	10	9	4,15%	4,98%	9,45%
13	202	164	5	12	2,48%	3,05%	10,37%
14	163	127	6	8	3,68%	4,72%	11,02%
15	189	153	5	15	2,65%	3,27%	13,07%
16	193	167	13	13	6,74%	7,78%	15,57%
17	186	159	11	10	5,91%	6,92%	13,21%
18	193	166	7	10	3,63%	4,22%	10,24%
19	247	201	11	11	4,45%	5,47%	10,95%
20	155	120	9	8	5,81%	7,50%	14,17%
21	216	185	8	10	3,70%	4,32%	9,73%
22	130	96	2	6	1,54%	2,08%	8,33%
23	353	324	14	16	3,97%	4,32%	9,26%
24	206	180	9	10	4,37%	5,00%	10,56%
25	241	183	14	15	5,81%	7,65%	15,85%
26	405	328	25	12	6,17%	7,62%	11,28%
27	149	121	6	11	4,03%	4,96%	14,05%
28	213	174	12	14	5,63%	6,90%	14,94%
29	188	153	7	14	3,72%	4,58%	13,73%
30	162	122	5	9	3,09%	4,10%	11,48%
31	211	160	4	11	1,90%	2,50%	9,38%
32	180	135	3	11	1,67%	2,22%	10,37%
33	194	159	7	9	3,61%	4,40%	10,06%
34	165	137	5	9	3,03%	3,65%	10,22%
35	367	328	17	15	4,63%	5,18%	9,76%
36	241	192	16	11	6,64%	8,33%	14,06%
37	217	167	10	9	4,61%	5,99%	11,38%
38	267	207	12	11	4,49%	5,80%	11,11%
39	202	156	11	12	5,45%	7,05%	14,74%
40	185	157	19	8	10,27%	12,10%	17,20%
41	242	206	8	12	3,31%	3,88%	9,71%
42	219	178	6	15	2,74%	3,37%	11,80%
43	162	116	7	9	4,32%	6,03%	13,79%
44	213	170	12	11	5,63%	7,06%	13,53%
45	193	166	9	10	4,66%	5,42%	11,45%
46	365	333	23	19	6,30%	6,91%	12,61%
47	170	133	5	11	2,94%	3,76%	12,03%
48	179	154	11	10	6,15%	7,14%	13,64%
49	157	127	9	9	5,73%	7,09%	14,17%
50	194	137	5	8	2,58%	3,65%	9,49%
51	158	126	4	8	2,53%	3,17%	9,52%
52	133	98	3	7	2,26%	3,06%	10,20%
53	310	224	7	11	2,26%	3,13%	8,04%
				Média	4,16%	5,18%	11,45%
				Desvio	0,01604	0,01872	0,02237

Fonte: Elaboração própria.

Os dados obtidos a partir do UNITEX comparando *tokens* com correspondência no dicionário e aqueles sem correspondência podem ser comparados aos dados acima. Nos dados do UNITEX, o número de ocorrência de *tokens* com inicial maiúscula (4.067) somado ao número de ocorrência de *tokens* sem inicial maiúscula (6.886) resulta em (4.067+6.886=10.953, vide tabela 8), as ocorrências sem correspondência no dicionário foram 992 e 506 (maiúsculas e minúsculas) que resulta em (992+506=1.498). A razão desse número com o anterior é 1.498/6886=0,21 ou 21,75%.

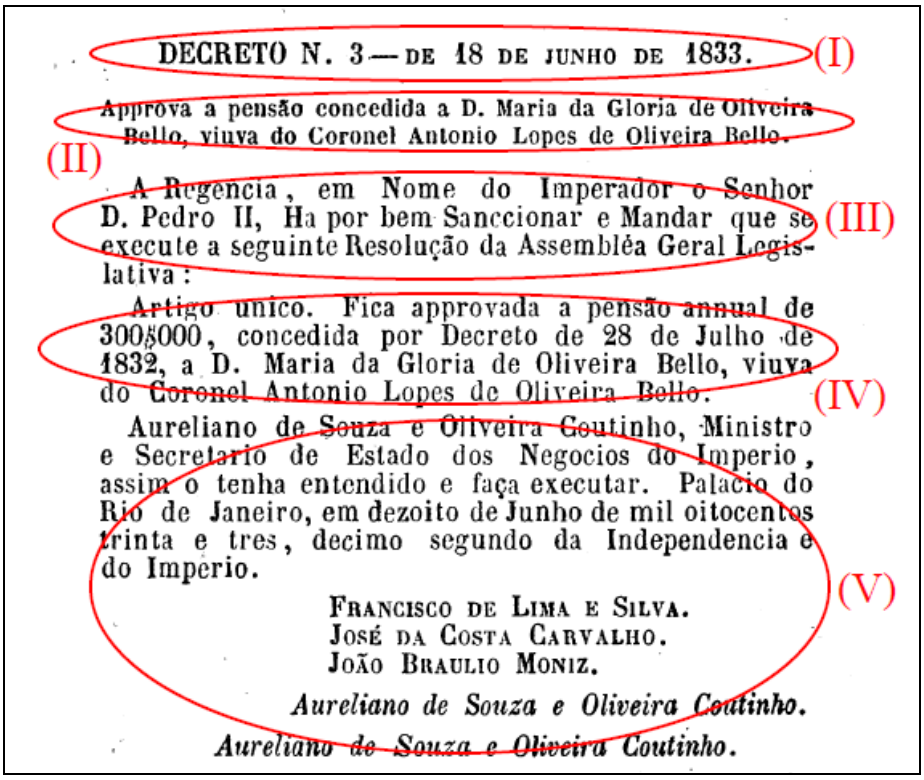
Os percentuais obtidos através da análise manual diretamente nos arquivos digitalizados revelam um percentual entre 4,16% e 11,45% e o percentual com dados diretamente do UNITEX revelam o percentual de 21,75%. Os primeiros percentuais são mais confiáveis, pois foram obtidos através de leitura e análise manual, o percentual de 21,75%, apesar de obtido com mais dados, foi obtido a partir de dados fora do dicionário, o que não significa necessariamente efeitos de mudança linguística.

Um número que pode exprimir a média de todos esses valores e indicar um percentual de mudança linguística no aspecto lexical entre a produção das informações registradas na amostra e o presente é 4,16+11,45+21,75/3=12,45 ou **12,45%.**

12.8.2 Aspecto morfológico-sintático

Com relação ao aspecto sintático, os documentos analisados apresentam uma estrutura padronizada e aplicada a todos os itens documentais na amostra. Isso ocorre já que todos são de uma mesma categoria de documentos legal (decretos) e todos emitidos por um mesmo poder (imperial) num mesmo ano. Analisados os documentos, todos eles possuem nessa ordem: (I) título da lei com data, (II) ementa, (III) cabeçalho, (IV) artigos (que pode ser único) e (V) encerramento com assinaturas dos responsáveis. Vide o texto do decreto 3 na *figura 20*.

Figura 20 – pdf das leis do império, dec. 1



DECRETO N. 3 — DE 18 DE JUNHO DE 1833. (I)

Approva a pensão concedida a D. Maria da Gloria de Oliveira Bello, viuva do Coronel Antonio Lopes de Oliveira Bello. (II)

A Regencia, em Nome do Imperador o Senhor D. Pedro II, Ha por bem Sanccionar e Mandar que se execute a seguinte Resolução da Assembléa Geral Legislativa: (III)

Artigo unico. Fica approvada a pensão annual de 300$000, concedida por Decreto de 28 de Julho de 1832, a D. Maria da Gloria de Oliveira Bello, viuva do Coronel Antonio Lopes de Oliveira Bello. (IV)

Aureliano de Souza e Oliveira Coutinho, Ministro e Secretario de Estado dos Negocios do Imperio, assim o tenha entendido e faça executar. Palacio do Rio de Janeiro, em dezoito de Junho de mil oitocentos trinta e tres, decimo segundo da Independencia e do Imperio. (V)

FRANCISCO DE LIMA E SILVA.
JOSÉ DA COSTA CARVALHO.
JOÃO BRAULIO MONIZ.

Aureliano de Souza e Oliveira Coutinho.

Aureliano de Souza e Oliveira Coutinho.

Fonte: Elaboração própria na publicação Leis do Império Brasileiro.

Nessa estrutura, há sempre a frase (com poucas variações) no cabeçalho: “Ha por bem Sanccionar e Mandar que se execute a seguinte Resolução da Assembléa Geral Legislativa”. O último artigo em cada decreto quase sempre contém (com poucas variações) “Ficam revogadas todas as.....”. E no encerramento encontra-se a construção “[...] assim o tenha entendido e faça executar”. A ementa sempre começa com uma frase que está em ordem inversa no primeiro artigo, no exemplo da figura: “Approva a pensão concedida a [...]” e no artigo único “Fica approvada a pensão [...]”.

As construções lexicais mais significativas, pois dão o sentido legal a cada decreto individualmente, são as frases na ementa (e primeiro artigo). A tabela 14 apresenta todas as frases mais significativas nas ementas e as sentenças comuns a todos os documentos.

Tabela 14 – Análise sintática

comuns:	"Ha por bem Sanccionar e Mandar que se execute a seguinte Resolução da Assembléa Geral Legislativa"; "[...], assim o tenha entendido e faça executar"; "Ficam revogadas todas as disposições em contrario"; "Ficam revogadas todas as Leis, Ordens, e mais disposições em contrario"; "Ficam derogadas quaesquer Leis, ou disposições em contrario"
1	"Approva a jubilação concedida ao Padre Manoel Ignacio de Carvalho"
2	"Approva as disposições dos estatutos da Academia das Bellas Artes"
3	"Approva a pensão concedida a D. Maria da Gloria de Oliveira Bello"
4	"Isenta de pagar dizimos e mais tributos os individuos [...] e manda supprir com [...]"
5	"Erige em Villa o Arraial de Bomfim"
6	"Approva os ordenados marcados pelo Presidente do Maranhão"
7	"Crêa no Arraial do Rio Claro, na Provincia de Goyaz, uma escola"
8	"Erige em Villa o arraial de Jaguará"
9	"Erige em freguezia a capella curada"
10	"Approva as pensões concedidas a [...]"
11	"Crêa uma cadeira de primeiras letras"
12	"Crêa na villa da Laguna"
13	"Faz extensiva a Provincia de Santa Catharina a Resolução do Conselho geral [...]"
14	"Approva a jubilação concedida ao Padre Francisco de Paula e Oliveira"
15	"Eleva os ordenados dos professores"
16	"Crêa na Capital da Provincia do Piauhy uma cadeira"
17	"Manda que se colloquem boias entre o pharól da Ilha de Santa Anna, e a barra do Maranhão"
18	"Providencia sobre o provimento das cadeiras de primeiras letras"
19	"Approva os ordenados de diversas cadeiras"
20	"Approva a pensão concedida a Francisco Rodrigues da Silva Mello"
21	"Determina sobre a fórma dos exames para o gráo de Doutor e provimento das cadeiras"
22	"Declara que Jacintho Vieira do Couto Soares é cidadão brasileiro"
23	"Determina sobre o julgamento dos processos anteriores á publicação do Codigo do Processo"
24	"Faz extensiva a todos os Tribunaes de Justiça do Imperio a disposição da Resolução [...]"
25	"Erige em Freguezia a Capella"
26	"Desmembra da Freguezia do Senhor Bom Jesus do Cuyaba, e erige em Freguezias [...]"
27	"Crêa escolas"
28	"Autoriza os Directores dos Cursos Juridicos de [...] admitir a"
29	"Autorisa o Governo a conceder a [...] privilegio exclusivo por dez annos"
30	"Approva a Tença concedida a D. Constança Clara de Souza Gonzaga"
31	"Autoriza o Governo a mandar abonar a "
32	"Autoriza o Governo a mandar pagar ao "
33	"Autoriza o Governo a mandar passar carta de serventia vitalicia "
34	"Approva a aposentadoria concedida a "
35	"Autoriza a construcção"
36	"Erige em Matriz a Capella de "
37	"Erige em Freguezia o Districto de "
38	"Erige em Freguezia a Capella do"
39	"Eleva á Igreja Parochial a Capella de "
40	"Crêa na Villa de Campos cadeiras de"
41	"Autoriza o Governo para fazer executar em todas as [...] o Regulamento de [...], e o additamento de [...], e para alteral-os nas suas disposições legislativas.
42	"Determina ácerca dos moveis e alfaias da extincta Congregação"
43	"Approva a pensão annual concedida a

continua...

continuação

44	"Autoriza o Director de qualquer dos Cursos Juridicos a admittir "
45	"Determina que o Distribuidor dos extinctos Juizos sirva conjunctamente com o Distribuidor e Contador do Civel e Crime para a distribuição dos feitos"
46	"Crêa Guardas policiaes em cada um dos districtos "
47	"Desliga do morgado pertencente ao Conde de Linhares e converte em bens allodiaes"
48	"Autoriza a Governo a contractar com quaesquer companhias, nacionaes ou estrangeiras o exclusivo da navegação"
49	"Manda dividir pelos accionistas os metaes preciosos "
50	"Approva a pensão annual concedida aos quatro filhos do finado Desmbargador
51	"Dispensa ao Bacharel [...], do intersticio exigido pela Lei para poder obter carta de naturalisação.
52	"Declara que [...], é cidadão brasileiro"
53	"Erige em Freguezia o Curato de "

Fonte: Elaboração própria.

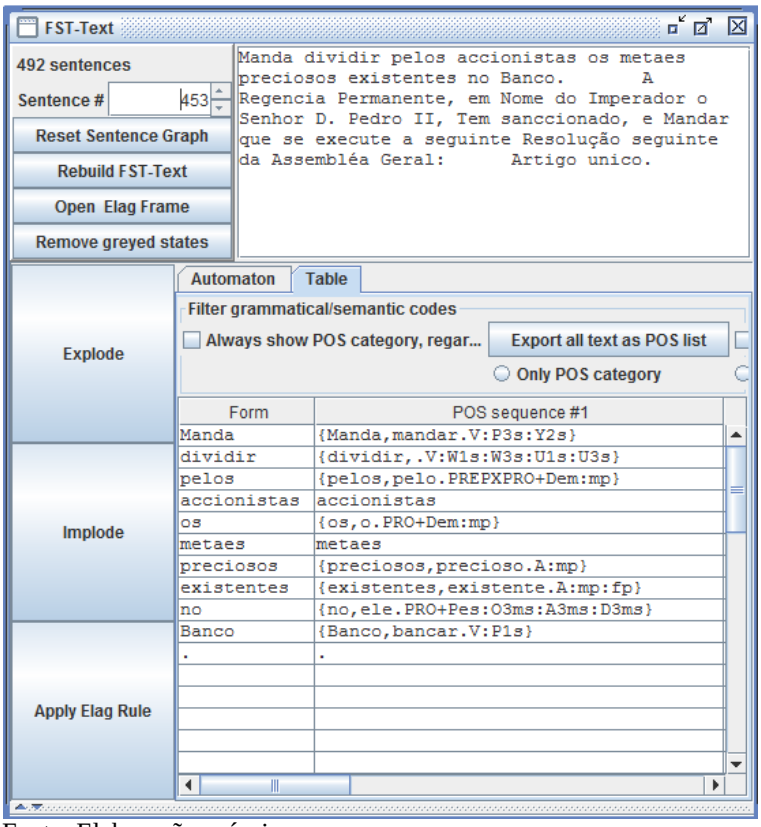
Em resumo, para cada documento, há em média três construções comuns a todos, no cabeçalho e no último artigo e no encerramento e uma frase que expressa o conteúdo do decreto (a ação principal) na ementa de cada documento.

Cada uma dessas frases pode ser analisada com relação aos elementos com ou sem correspondência no dicionário eletrônico utilizado pelo UNITEX. Pode-se procurar a frase no arquivo *SentençasDestacadas_491.doc* que relacionou as 492 sentenças e, através do aplicativo UNITEX, visualizar os dados.

Tome-se como exemplo a frase no documento 49 "Manda dividir pelos accionistas os metaes preciosos". Essa frase no arquivo .DOC corresponde à sentença 453, a qual, visualizada no UNITEX, resulta nas informações da *figura 21*.

Pela figura 21, é possível observar que as formas "accionistas" e "metaes" não puderam ser analisadas em função da não correspondência no dicionário. Praticamente todas as sentenças analisadas através do UNITEX com o procedimento anterior apresentam pelo menos um elemento sem correspondência no dicionário, o que sugere que o ponto de vista sintático, ou seja, das relações entre os elementos morfológicos entre si, é bastante afetado pelos efeitos da mudança linguística em seu aspecto lexical.

Figura 21 – Sentença 453 analisada



Fonte: Elaboração própria.

Esses dados também mostram a (co)relação entre elementos linguísticos com aspecto lexical alterado, dificuldade para análise morfológica e consequentes problemas na análise sintática entre esses elementos. Notar que cada uma das sentenças precisa ser analisada com o aplicativo UNITEX em funcionamento após abrir o arquivo com todos os decretos no formato .TXT.

12.8.3 Aspecto semântico-pragmático

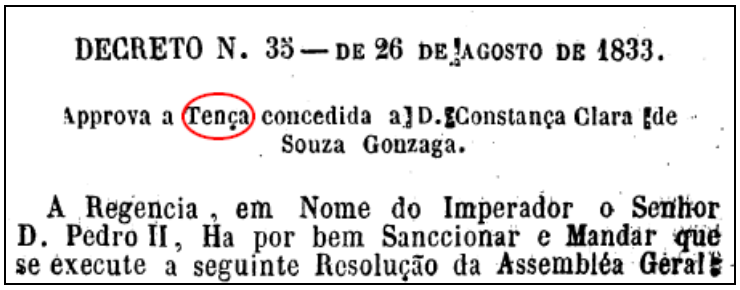
No caso do aspecto semântico-pragmático, é fundamental compreender que certo grau de subjetivismo na análise parece inevitável. Aplicativos com o UNITEX são limitados em relação a esse tipo de análise. Na análise desses aspectos o que se pretende é verificar quantos elementos linguísticos (termos ou palavras) apresentam defasagem em relação ao seu sentido/significado (semântico) e seu uso naquele contexto histórico (pragmático) rem relação ao sentido/uso atual dos termos.

A objetividade para essa tarefa pode ser obtida através do uso de um dicionário, não como o utilizado no UNITEX, mas um atual da língua portuguesa e/ou bases de dados.

No entanto, o processo de destacar as palavras que podem ter sido alteradas em função da mudança linguística é um processo que resulta em escolhas diferentes para diferentes pessoas que façam a escolha. Nesse sentido, é importante esclarecer que a escolha das palavras para análise reflete o ponto de vista deste autor como usuário nativo da língua portuguesa (na versão brasileira) e seu grau de cultura, daí o subjetivismo.

A *figura 22* exemplifica esse problema no documento 30.

Figura 22 – Token "tença"


DECRETO N. 35 — DE 26 DE AGOSTO DE 1833.

Approva a Tença concedida a D. Constança Clara de Souza Gonzaga.

A Regencia, em Nome do Imperador o Senhor D. Pedro II, Ha por bem Sanccionar e Mandar que se execute a seguinte Resolução da Assembléa Geral:

Fonte: Elaboração própria a partir da publicação Leis do Império.

Na *figura 22*, a palavra “tença” foi destacada, pois parece ser um termo que aponta para um significado não mais existente, ou que atualmente é descrito por outro termo. Outra pessoa fazendo a mesma análise poderia acreditar que se trata de um termo tão comum como “ordenado”, “mulher” ou “mil réis”. Nesse caso, coincidentemente, não há correspondência com o dicionário eletrônico do UNITEX para o termo "tença".

De qualquer forma, a análise qualitativa permitirá verificações mais detalhadas e assim indicará, com mais precisão, os possíveis efeitos da mudança linguística nos aspectos semântico e pragmático. A tabela 15 apresenta os dados obtidos mediante a leitura individual de cada documento. O arquivo *Decretos_1833_Semântico_Pragmático.docx* contém os termos em destaque que sugerem defasagem no sentido e/ou uso semântico/pragmático. A aba *Semântico_Pragmático* no arquivo *tokensUNITEX.xlsx* contém outras observações registradas.

Tabela 15 – Aspecto semântico-pragmático

Doc #	Termo	Total
1	jubilação	3
	cadeira publica	
	theologia dogmatica	
2	lente	4
	professor de miologia	
	professor de physiologia das paixões	

continua...

continuação

	professor osteologia	
3	pensão	1
4	aldeamento	3
	cavallar	
	gado vaccum de criar	
6	ordenados	2
	professor de ensino primário	
7	escola de primeiras letras	2
	methodo individual	
9	freguezia	2
	capella curada	
10	guardas municipaes permanentes	2
	pensões	
11	cadeira de primeiras letras	3
	povoação	
	ordenado	
12	escola de primeiras letras para meninas	4
	ordenado	
	mestra das meninas	
	professor de primeiras letras	
13	legislação peculiar	1
14	Jubilação	4
	Cadeira	
	ordenado por inteiro	
	philosohia racional e moral	
15	Ordenado	3
	professor de primeiras letras	
	Província	
16	Cadeira	3
	pôr a concurso	
	francez e geografia	
17	cofres nacionais	3
	commissão de marítimos	
	systema de boias	
18	cadeira de primeiras letras	2
	methodo Lencastriano	
19	Ordenado	2
	ensino mutuo	
20	pensão anual	2
	sciencias juridicas, e sociais	
21	Lentes	2
	gráo de doutor	
23	processos crimes	3
	de libelo	
	jury de sentença	
24	Causas cíveis e crimes	1
25	freguezia, capella e districto	6
	Côngrua	

continua...

continuação

	Guizamento	
	Conhecenças	
26	Parochia	3
	Côngrua	
	capellas coladas	
27	escolas de primeiras letras para meninas	1
28	fazer acto das matérias	2
	carta de bacharel formado	
30	Tença	1
31	Vencimentos	1
32	Soldos	1
33	carta de serventia vitalícia	2
	escrivão da mesa grande	
34	escrivão da mesa da estiva	1
35	animaes vacum	3
	Cavalar	
	Officiaes	
36	povoação	4
	Côngrua e guizamentos	
	ordinario da diocese	
37	arroio	1
38	congrua	1
40	rhetorica, philosophia, francez, arithmetica, geometria e algebra	6
41	additamento	1
42	alfaia	1
	prelado diocesano	1
43	pensão annual	1
44	fazer acto das materias	1
45	Conselho da fazenda	4
	Contador geral do cível e crime	
	Correções do cível e crime da corte	
	Juízo da Corôa	
46	Guardas policiaes	2
	vencimento diários	
47	bens allodiaes	2
	morgado	
48	exclusivo da navegação	1
50	pensão annual	1
53	freguezia, curato	4
	natureza collativa	

Fonte: Elaboração própria.

A partir dessa tabela, apenas oito documentos (15% do total) não contêm nenhum termo que sugere dúvidas em relação ao seu significado ou uso naquele contexto, são os documentos 5, 8, 22, 29, 39, 49, 51 e 52.

12.9 Conclusões da análise documental

A análise documental realizada explorou estes aspectos linguísticos: lexical, morfológico, sintático, semântico e pragmático. Inicialmente, foi efetuada uma análise por meio de recursos computacionais. Esse tipo de análise permite precisão e rapidez, notadamente quando se trata de grandes quantidades de informações. Como a amostra documental não é grande, como compensação também foi realizada uma análise manual também se apoiando nos dados obtidos pelo aplicativo UNITEX. Esta última análise foi importante para complementar principalmente os aspectos semântico-pragmáticos que requerem uma análise individualizada.

É um limite importante a relatividade dessa análise com o uso do estado atual da língua portuguesa por quem analisa. Não apenas é possível obter resultados diferentes a partir do ponto de vista de outras pessoas, como também é fundamental esclarecer que a análise no tempo presente interfere nos resultados. As mesmas análises agora realizadas, quando repetidas em cinquenta ou cem anos, devem produzir resultados diferentes, pois haverá uma distância maior entre os documentos registrados (estado da língua no período do império) e o uso da língua portuguesa no futuro. Espera-se um tendência para que os mesmos dados, analisados em décadas no futuro, indiquem maior defasagem por exemplo no aspecto lexical e semântico.

As análises feitas demonstram que é possível quantificar aspectos dos efeitos da mudança linguística num acervo documental considerado. Se não é possível derivar leis ou tendências gerais para qualquer acervo documental, fica demonstrado que é possível fazê-lo para um acervo específico, esses dados podem ser associados a outros fatores então:

> Conforme o desenvolvimento de uma investigação, a pesquisa documental poderá ser uma fonte de dados e informações auxiliar, subsidiando o melhor entendimento de achados e também corroborando evidências coletadas por outros instrumentos e outras fontes, possibilitando a confiabilidade de achados através de triangulações de dados e de resultados. (MARTINS, THEÓPHILO, 2007, p. 86).

Das análises realizadas, o aspecto lexical (ortografia e acentuação) e os aspectos semântico-pragmáticos parecem ser aqueles que mais evidenciam efeitos com relação à mudança linguística. Mas fica claro também que todos os aspectos analisados se (inter)relacionam e são (inter)dependentes, por exemplo, na análise do aspecto sintático no aplicativo UNITEX, em que a falta de elementos associados ao dicionário (por serem de outra época) não permitiu uma análise morfológica, à qual por sua vez impacta na análise sintática.

13 CONCLUSÕES GERAIS

Retomando o objetivo geral desta pesquisa, qual seja:

Descrever como a mudança linguística pode afetar a recuperação de documentos de arquivo textuais de caráter permanente, através de sistemas de recuperação da informação que serão utilizados nos novos estados da língua portuguesa.

E considerando também o nível de pesquisa exploratório que este trabalho se propôs a desenvolver, com base na revisão de literatura e nos dados obtidos e analisados, é possível derivar nossas conclusões finais.

Uma primeira conclusão envolve questões conceituais sobre os efeitos da mudança linguística, entre estados de uma língua, em sistemas de recuperação da informação. O quadro sinóptico que apresentamos ao final da revisão de literatura apresenta um contexto de conceitos e relações que julgamos essencial para a compreensão dos efeitos da mudança linguística em sistemas de recuperação da informação. Conceitos como língua, língua escrita e linguagem são fundamentais neste contexto. Seria proveitoso, para a melhor compreensão do problema dos efeitos da mudança linguística em sistemas de recuperação no longo prazo, se distinções entre esses termos e o termo "língua" ocorressem de maneira clara. Até porque, no caso do termo "língua escrita", este é de grande importância, pois está presente tanto no conteúdo dos documentos, sua informações, como também não representações sobre esses documentos, inclusive nas representações sobre documentos não textuais, embora estes estivessem fora do escopo da pesquisa.

Ainda com relação às conclusões num nível conceitual, alguns conceitos mais complexos se mostraram importantes para a melhor compreensão dos efeitos da mudança linguística, a longo prazo, em sistemas de recuperação da informação, como a distinção entre "variação" e "mudança linguística", além de suas mútuas relações.

É importante esclarecer que o problema do vocabulário (do ponto de vista das pessoas envolvidas na criação, representação, busca e recuperação) e as ambiguidades nas *queries* de pesquisa (do ponto de vista dos sistemas) são dois problemas que ocorrem sincronicamente e continuamente ao longo dos diferentes estados de uma língua considerada. Os efeitos da mudança linguística são um elemento independente dos primeiros, mas que potencializam tanto o problema do vocabulário como as ambiguidades nas pesquisas. Essa distinção parece fundamental. E ainda com relação a essas relações, também é relevante destacar que sistemas de recuperação da informação sempre podem ser programados para tentar contornar os efeitos da mudança linguística, o problema mais importante é como obter

os dados sobre o que mudou linguisticamente a fim de alimentar a programação de sistemas.

Ainda que o escopo de pesquisa compreenda os efeitos futuros, no longo prazo, em sistemas de recuperação, a análise documental efetivada permitiu derivar importantes conclusões – com base em documentos produzidos no passado – acerca do que ocorrerá no futuro com acervos de documentos de arquivo com valor histórico, produzidos na atualidade.

A primeira conclusão da análise documental é a caracterização de como a mudança linguística afeta as informações nos documentos considerados. Como um sistema complexo, a língua muda em vários aspectos, afetando mutuamente cada um desses aspectos. Nos dados analisados na pesquisa documental, foi possível identificar, por exemplo, alterações no aspecto semântico-pragmático diretamente ligadas a alterações no aspecto lexical e vice-versa. Os percentuais de alterações, pelo menos no aspecto lexical e semântico, podem até sugerir um possível índice do quanto a mudança linguística afetou os elementos linguísticos num acervo considerado.

Como já mencionado na revisão de literatura, sistemas de recuperação da informação podem ser programados para contornar os efeitos detectados da mudança linguística. No entanto, mudanças no aspecto semântico-pragmático podem ser difíceis de reconstruir em relação ao uso original, incluindo um risco de não compreensão dos significados e usos originais, o que sugere um risco, potencialmente maior conforme o lapso temporal, de não recuperação de informações no futuro sobre o presente.

Na introdução sobre a parte que trata de metodologia, destaca-se uma citação sobre estudos exploratórios, a qual aqui é repetida: “as pesquisas exploratórias têm como principal finalidade desenvolver, esclarecer e modificar conceitos e ideias, tendo em vista, a formulação de problemas mais precisos ou hipóteses pesquisáveis para estudos posteriores” (GIL, 2006, p. 43). Considera-se que a conclusão mais importante, além da compreensão mais precisa do problema de pesquisa tratado, é a possibilidade de aventar uma hipótese de pesquisa que permita avançar na investigação da questão da mudança linguística e a recuperação de informações no longo prazo, principalmente, se for possível propor procedimentos mitigatórios aos problemas apontados.

A caracterização de como a mudança linguística pode afetar a língua escrita num acervo documental histórico considerado e os riscos de no futuro não ser possível contornar as alterações em pelo menos alguns aspectos como o semântico-pragmático sugerem que, para os acervos de documentos de arquivo com valor histórico que são produzidos hoje, todos os aspectos da mudança linguística devem ser monitorados (acompanhados e registrados), inclusive no aspecto fonético-fonológico, embora não tenha feito parte deste estudo devido às

características dos documentos considerados.

O monitoramento da mudança linguística pode permitir – já que essa hipótese deverá ser confirmada no futuro – a redução substantiva de problemas decorrentes de alterações linguísticas no longo prazo. Isso permitirá que os sistemas de recuperação no futuro, seja qual for a tecnologia utilizada, possam contornar as alterações em relação aos estados futuros da língua considerada.

Certamente, ainda há muito que investigar sobre a hipótese do monitoramento da mudança linguística num acervo de documentos de arquivo ao longo do tempo. Concretamente, há desafios teóricos e práticos sobre como fazer tal monitoramento. Avançou-se um pouco em relação a esta investigação que segue na seção seguinte como estudos futuros. Nesses estudos, aborda-se a possibilidade de integrar este monitoramento com um novo modelo de recuperação da informação, já que a investigação sobre modelos de RI, na revisão da literatura, indica que a mudança linguística não é devidamente abordada proporcionalmente aos riscos para a recuperação futura. Também se avançou um pouco em relação a procedimentos práticos na seção sobre “como monitorar as mudanças”.

14 ESTUDOS FUTUROS

Esta seção inclui outros resultados de pesquisas nesta tese sobre os efeitos da mudança linguística, a longo prazo, em sistemas de recuperação da informação, mas, por estarem ainda em estado não totalmente completo, optou-se por apresentar como estudos futuros.

Com relação à discussão teórica, tratou-se de um possível modelo no qual os procedimentos de monitoramento da mudança linguística (última conclusão exposta na seção anterior) pudessem ser integrados. Também se avançou com considerações sobre possíveis procedimentos práticos para efetivar tal monitoramento da mudança linguística. O ponto de vista aqui ainda é o mesmo do restante da tese, ou seja, aplica-se a acervos de documentos de arquivo, de guarda permanente, que são produzidos contemporaneamente e deverão ser recuperados pelas futuras gerações utilizando outros estados da língua portuguesa.

14.1 Necessidade de um modelo com características específicas

As análises dos principais modelos contemporâneos para busca e recuperação da informação efetuadas na revisão de literatura sugerem a importância de um novo modelo que incorpore as características necessárias. A estratégia adotada aqui foi apresentar um esboço de tal modelo, propondo suas características iniciais.

Nas subseções seguintes, são discutidas as etapas fundamentais nesse modelo, como o tempo e a ML podem operar nesse modelo, o problema do vocabulário e o consequente papel das pessoas, a visão da língua como um sistema complexo e algumas considerações de procedimentos sobre como monitorar as mudanças num acervo determinado. Espera-se que todos esses elementos formem um esboço do modelo pretendido, que deverá ser mais bem desenvolvido por futuras pesquisas.

14.2 Etapas necessárias, tempo e mudança

Um modelo adequado para mitigar os efeitos da ML sobre SRIs deve incluir os processos tanto na fase de representação, bem como na fase de recuperação das informações registradas em documentos de arquivo (documentos de guarda permanente) produzidos na atualidade e principalmente que o elemento tempo e ML estejam claramente identificados.

Os conceitos de processo e subprocesso ou etapas de um mesmo processo podem

ser relativizados, dependendo do ponto de vista de quem faz a análise. Assim, uma sequência de operações pode ser vista como um processo. Isso foi feito com o P. de representação da informação (como é considerado em Ciência da Informação). Mas, pode-se analisar de um ponto de vista mais amplo e entender como sendo um subprocesso (ou outra etapa), se se considerar também o outro processo importante P. de recuperação de informações. Em princípio, considerando a análise dos dois processos, tratar-se-á de ambos como duas etapas de um processo maior.

Considerando os dois parágrafos anteriores, um conjunto de requisitos básicos que podem ser tomados como ponto de partida para os estudos sobre esse modelo podem ser assim elencados, *quadro 11*:

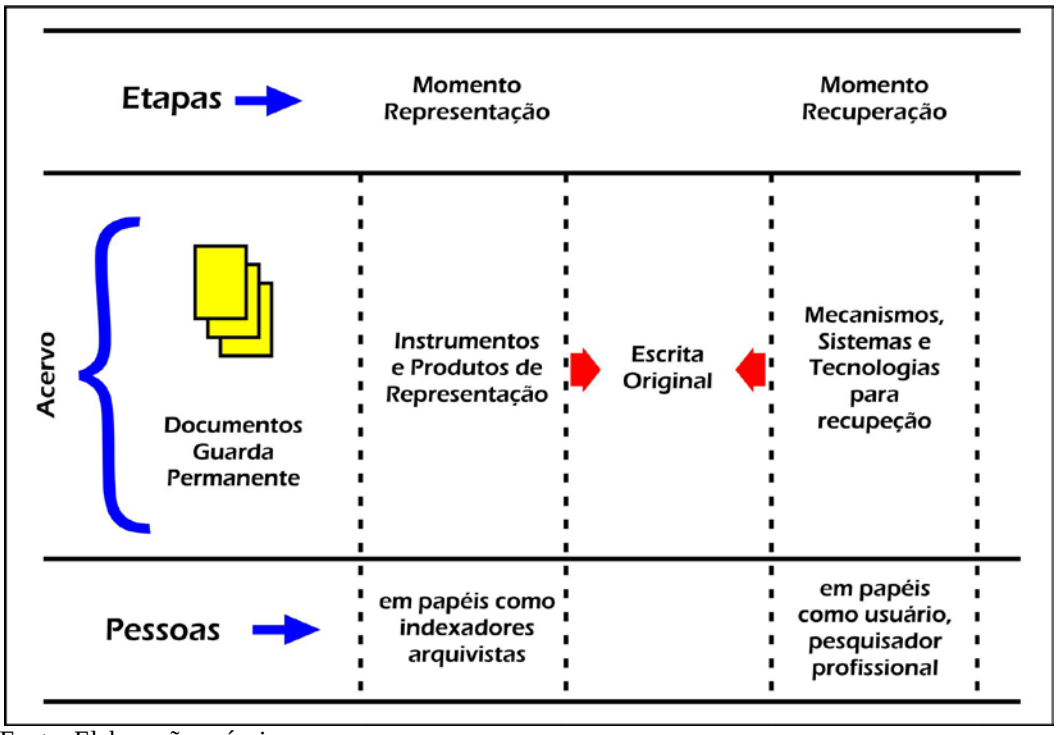
Quadro 11 – Requisitos básicos para o modelo proposto

Req.	Descrição
1	Considerar a representação e a recuperação como duas etapas relacionadas e (inter)dependentes.
2	Abstração de procedimentos práticos para a representação e recuperação.
3	Abstração de sistemas digitais utilizados tanto para a representação como para a recuperação.
4	Abstrair o conteúdo documental e destacar o papel da língua escrita e respectivos elementos linguísticos nos instrumentos/produtos das etapas de representação e recuperação.
5	Destacar o papel das pessoas envolvidas nas etapas que consideramos importantes nas duas etapas: representação & recuperação.
6	O modelo aplica-se para documentos de arquivo de guarda permanente, produzidos na atualidade.

Fonte: Elaboração própria.

A *figura 23* ilustra graficamente um esboço para esse modelo. A disposição gráfica na figura corresponde aos momentos temporais em que as etapas são executadas, a partir da esquerda. O ponto de partida é a existência de um acervo de documentos de arquivo de guarda permanente.

Figura 23 – Elementos do modelo proposto



Fonte: Elaboração própria.

A descrição das etapas sugeridas na figura 23 é a seguinte:

14.2.1 Primeira etapa: representação.

A primeira etapa da figura corresponde à representação de informações registradas nos documentos daquele acervo considerado que é composto por documentos de arquivo de guarda permanente. Na parte debaixo dessa etapa, são representadas as pessoas que executam os correspondentes processos específicos. Estes podem ser a elaboração de instrumentos típicos para a descrição de documentos de arquivo como guias, inventários ou catálogos.

Independentemente dos instrumentos/produtos específicos utilizados na etapa representação, o fato é que sempre haverá elementos linguísticos associados e produzidos a partir desses processos, como demonstrado anteriormente. Considere-se, por exemplo, o caso de um guia do arquivo (fundo) tratando do acervo como um todo ou resumos criados a partir de alguns documentos específicos escolhidos.

É importante ressaltar que essa especificação se aplica a acervos com documentos considerados de guarda permanente, ou seja, a serem mantidos e potencialmente recuperados indefinidamente no futuro, mas contemporâneos às pessoas que participam dessa etapa de

representação: Pessoa = $P_{Representação}$.

À totalidade dos elementos linguísticos presentes nos instrumentos/produtos nessa primeira etapa, contemporaneamente aos documentos produzidos, esta será chamada de escrita representada originalmente ($E_{RepresentaOriginal}$). Essa distinção é importante, pois é possível que, ao longo do tempo, várias outras tentativas de representação das informações naquele acervo sejam implementadas, substituindo ou não a representação original. Ter-se-á então:

Elementos linguísticos futuros = $E_{RepresentaV1....Vn}$ (onde V1 a Vn são as futuras representações sobre aquele mesmo acervo).

A $E_{RepresentaOriginal}$, por ser contemporânea simultaneamente aos documentos no acervo e às pessoas ($P_{Representação}$) presentes nessa etapa, é a que possui as melhores chances de representar corretamente o conteúdo linguístico. Quando os processos de representação são executados sobre acervos antigos por pessoas na atualidade, ou seja, quando há um lapso temporal grande entre a produção dos documentos e o momento em que as pessoas executam a representação (preparam os instrumentos e/ou produtos), haverá um risco maior de problemas em relação à descrição do conteúdo linguístico. Isso ocorre porque a ML pode afetar de várias maneiras o conteúdo original nos documentos do acervo antigo em relação às pessoas na atualidade. Considere os exemplos abaixo.

Exemplo 1: Em documentos em um acervo antigo, utilizava-se a palavra "Robe de Chambre"[1]. Ocorre que essa palavra não é mais utilizada na atualidade e, ainda que seja possível fazer pesquisas sobre as características da língua utilizada naquela época (fontes históricas), há um risco de a informação nessas fontes não ser precisa ou mesmo não ser nelas encontrada.

Exemplo 2: Em documentos de um acervo antigo, utilizava-se o termo "Chapelaria"[2], que possuía um sentido específico naquele contexto institucional naquele momento. Em algum momento na atualidade ou no futuro distante, as pessoas ($P_{Representação}$) envolvidas na etapa de representação terão grandes dificuldades para compreender aquele termo no sentido utilizado originalmente, se é que será possível compreendê-lo.

O que há de comum nos dois termos acima "Robe de Chambre" e "Chapelaria" é

[1]Designação antiga para o que é chamado hoje "roupão" (VILLAS, 2012).

[2]A Câmara dos Deputados em Brasília até hoje utiliza internamente o termo Chapelaria para designar um local do complexo de prédios. Mas não há chapéus ou chapelaria lá, o nome foi mantido quando há mais de três décadas chegou a existir uma chapelaria de verdade, onde as pessoas, principalmente deputados podiam deixar seus chapéus.

que eles eram utilizados em outro estado da língua (portuguesa brasileira) e, provavelmente possuíam sentidos precisos em determinados setores da sociedade ou mesmo sentido restrito a uma instituição (Chapelaria).

Os problemas em potencial elencados acima serão maiores e mais importantes de forma diretamente proporcional ao tamanho do lapso temporal entre a produção dos documentos no acervo considerado e o momento em que as pessoas ($P_{Representação}$) atuam na representação daqueles documentos antigos, já que a ML é um processo constante para uma língua qualquer considerada.

Isso tudo não significa que necessariamente a $E_{RepresentaOriginal}$ estará isenta de problemas linguísticos entre o conteúdo dos documentos no acervo e as pessoas ($P_{Representação}$). Por exemplo, alguém que tente preparar um resumo textual sobre um documento pode interpretar incorretamente o uso de um termo qualquer. Mas, em casos como esses, fala-se de fatores imponderáveis. O fato é que a $E_{RepresentaOriginal}$ produzida por pessoas ($P_{Representação}$) contemporâneas é a que possui melhores chances, excetuados erros pessoais, de melhor representar o documento original.

A análise documental implementada antes também permite a visualização com exemplos do modo como as mudanças passam a ser perceptíveis se comparadas entre dois estados da língua portuguesa.

14.2.2 Segunda etapa: recuperação.

Na etapa da recuperação das informações registradas nos documentos do acervo de documentos considerado, serão usados outros instrumentos e mecanismos específicos que também poderão utilizar a $E_{RepresentaOriginal}$ como base para os processos de recuperação. Como procurou-se demonstrar, a $E_{RepresentaOriginal}$ é a melhor alternativa, se estiver disponível, do ponto de vista da representação da informação, mas será também a melhor alternativa do ponto de vista da RI?

A resposta a essa pergunta dependerá, na verdade, de outro fator que tem de ser considerado no esboço de modelo aqui considerado. Trata-se do lapso temporal entre a etapa de representação e a etapa de recuperação.

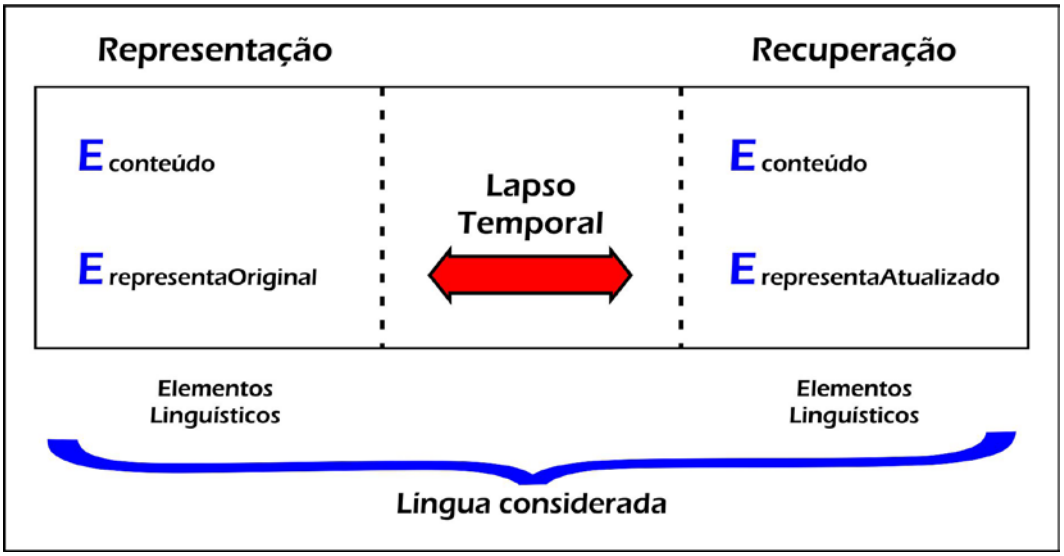
Considere-se, por exemplo, o caso em que há um lapso temporal de 120 anos entre as duas etapas. É razoável esperar que, nessas condições, as pessoas que atuam na RI ($P_{Recuperação}$) sejam diferentes das pessoas na etapa de representação ($P_{Representação}$). E mais importante ainda, em função da ML, o uso da língua deve ser diferente entre essas pessoas no

passado e 120 anos à frente, isso em vários aspectos.

É possível resumir a problemática nos seguintes termos. Se houver uma representação sobre as informações registradas em um acervo documental num determinado momento, contemporâneo à produção documental, essa representação tende a ser mais precisa ou com maior qualidade. Por exemplo, pode-se lançar uma nota num resumo explicando o histórico do termo institucional "Chapelaria" ou descrever esse termo em relação ao que realmente é, ou seja, a denominação de uma chapelaria de verdade ou apenas um nome tradicional para um local dentro da instituição. Mas, quanto maior for o lapso temporal ocorrido, as pessoas no futuro tendem a não utilizar termos no sentido original. Imagine-se que em 1970 havia uma chapelaria na Câmara dos Deputados, mas em 2070, ao se procurar por documentos com esse termo, não se estará buscando uma chapelaria de verdade, mas apenas o endereço ou localização interna. Claro que o problema será maior conforme a complexidade e a importância do termo procurado no futuro.

É importante que um modelo que contemple a recuperação da informação e os efeitos da ML especifique também o fator tempo ou lapso temporal entre as etapas representação e recuperação. E também destaque que esse lapso tenha que ser longo o suficiente para que fiquem evidentes os efeitos dessa mudança em vários aspectos de uma língua considerada, A *figura 24* ilustra graficamente esse aspecto em relação ao lapso temporal.

Figura 24 – Lapso temporal como fator a ser considerado



Fonte: Elaboração própria.

14.3 Como monitorar as mudanças

De um ponto de vista estritamente científico, não é possível prever o futuro. Por isso, infelizmente, ações concretas em relação ao que orienta os objetivos desta pesquisa são necessariamente limitadas. Porém, é possível preparar-se para o futuro com base no que se sabe sobre o passado e presente. Se não é possível saber como será a tecnologia utilizada em SRIs no futuro de longo prazo – de fato, é difícil prever mesmo uma ou duas décadas a frente –, por outro lado, existem outras informações úteis.

Primeiro, é sabido que a ML existe de fato com base em muitos dados empíricos, na verdade esse é um fato incontestável. Segundo, os sistemas de recuperação da informação registrada em documentos dependem da língua (na forma escrita) para recuperar informações. Consequentemente, pode-se deduzir que falantes de língua portuguesa que tentarem recuperar informações, talvez daqui a um século, registradas em documentos arquivísticos produzidos e representados hoje tentarão fazê-lo com uma futura versão da língua portuguesa atual. E esse descompasso entre as duas versões pode significar falhas no processo de recuperação.

Mas tome-se uma porção de língua utilizada num acervo documental específico. Se é perfeitamente plausível que, com base em informações atualizadas sobre a evolução da língua portuguesa nesse acervo, os futuros sistemas poderão contornar (com procedimentos técnicos) as eventuais alterações ocorridas (lexicalmente, morfologicamente, sintaticamente, semanticamente ou até mesmo fonética e pragmaticamente), então o que se pode fazer hoje, em termos procedimentais? Podem-se registrar as evoluções parciais (monitoramento) ocorridas nessa porção de língua portuguesa. Esses procedimentos podem, inclusive, fazer parte da especificação do modelo cujo esboço é aqui apresentado com suas principais características.

Durante o século XIX, o que hoje se denomina linguística histórica desenvolveu e aprimorou métodos científicos para analisar a evolução fonética de línguas, trata-se do método comparativo. Um dos objetivos era estabelecer famílias de línguas e até mesmo (re)construir as línguas antigas já mortas que deram origem às atuais, essa é uma das justificativas para a existência de uma linguística histórica, "disciplina que estuda as alterações (supressões, acréscimos, adaptações, substituições, rearranjos) que se operam nas línguas através do tempo" (BORBA, 1986, p. 279). Em grande parte, essas análises comparativas eram baseadas em registros escritos, mas algumas leis foram derivadas e permitiram a dedução hipotética de formas antigas sem registro escrito (o que, em linguística histórica, é representado com um asterisco junto à palavra obtida, por exemplo, **fader*).

O que é útil desse trabalho dos neogramáticos, filólogos e linguistas históricos é justamente um exemplo sobre comparação de textos de diferentes épocas, ainda que com objetivos muito distintos dos deste estudo e não considerando todos os aspectos modernos de estudo das línguas. "A comparação de dois textos de diferentes épocas evidencia diferenças em todos os componentes da estrutura linguística: o fonológico, o gramatical (morfologia e sintaxe) e o léxico." (BORBA, 1986, p. 279). As diferenças detectadas ocorriam através da comparação de elementos com séculos de diferença.

Veja-se um exemplo de análise comparativa em línguas derivadas do latim extraído de Lyons (1987) para o termo "cavalo". Em latim *caballus*, em francês *cheval*, em italiano *cavallo* e em espanhol *caballo*. Destaca-se nesse exemplo o interesse na comparação dos sons nas diferentes línguas, pois na verdade em latim o termo mais adequado seria "*equus*". O termo "*caballus*" era utilizado com sentido próximo, mas diferente do atual no latim antigo. Evidencia-se daí que o sentido das palavras muda com o tempo. Outro exemplo importante é o termo "*canis*", o qual a partir do latim deu origem a "*chien*" (F) e "*cane*" (I), mas não há correspondente sonoro em espanhol, pois esse termo desapareceu nessa língua (LYONS, 1987). Os cães em espanhol são chamados de "*perro*", palavra que não possui nenhuma relação fonética com "*canis*", ainda que tenha uma relação semântica.

O ponto de vista na Ciência da Informação é diferente daquele na linguística histórica. Especificamente, àquela interessam os possíveis efeitos deletérios de mudanças – como as exemplificadas no parágrafo anterior – para os futuros SRIs. Mas há elementos em comum, qual seja, observar a língua portuguesa em sua forma escrita, no caso em documentos de arquivo de guarda permanente e suas representações, para monitorar e documentar as mudanças.

Mas esse monitoramento não pode ser feito por linguistas históricos? E não estão esses profissionais, de seu ponto de vista, mais qualificados para executar esses procedimentos? Os linguistas não só podem como estão fazendo esse procedimento e, atualmente, em relação a todos os aspectos da língua portuguesa (como também em outras línguas no mundo). Porém, os objetivos deles não são os mesmos do ponto de vista da Ciência da Informação, nem é o mesmo o nível de profundidade do estudo sobre essa mudança. Por exemplo, em linguística há preocupação em entender por que uma língua muda. Além disso, em função da profundidade científica do que é feito por linguistas, o trabalho é lento em relação a uma língua totalmente considerada.

Esta proposta é sobre o monitoramento da ML em relação aos elementos linguísticos presentes num acervo determinado, um universo muito menor do que o total de

elementos numa língua qualquer, por maior que seja esse acervo de documentos. Isso significa que a emergência das informações sobre a ML nesse acervo pode ser mais facilmente disponibilizada, evitando, assim, os efeitos indesejados nos SRIs. No fundo, o que se defende é uma postura não reativa quanto ao problema da ML num acervo documental, ou seja, não esperar que essa evolução produza efeitos indesejados, mas monitorar e possibilitar a permanente adaptação dos sistemas envolvidos. Isso é coerente, pois a preservação da memória é um problema do ponto de vista da Ciência da Informação. Aliás, como ficou evidente da sondagem na produção de CI na última década e meia (teses), o tema da memória tem sido recorrentemente abordado.

Outra diferença fundamental em relação ao que a linguística histórica executa é que, quando se defende o "monitoramento da ML", não se está sugerindo estudar o que ocorre em relação a causas ou efeitos, mas apenas DOCUMENTAR as alterações linguísticas detectadas ao longo do tempo. E, nesse ponto, os produtos dos estudos linguísticos em relação à ML serão valiosas fontes de informações. Ao lado de outras fontes como a história da instituição que deu origem aos documentos, estudos e produtos arquivísticos, seus termos e idiossincrasias específicas, ou seja, há várias fontes para alimentar esse monitoramento, que, insista-se, limita-se a documentar as mudanças linguísticas apenas quanto aos elementos linguísticos no acervo considerado.

Uma obra que pode ilustrar adequadamente o estudo da língua portuguesa e sua evolução do ponto de vista linguístico é o livro *Fundamentos Histórico-Linguísticos do Português do Brasil* (ELIA, 2003). Nesse trabalho, o autor faz um levantamento das primeiras obras que trataram da língua portuguesa de um ponto de vista científico no Brasil, vide *quadro 12*.

Quadro 12 – Primeiras obras brasileiras sobre nossa língua

Obra	Autor	Ano
O dialeto caipira	Amadeu Amaral	1920
O português do Brasil	Renato Mendonça	1936
O linguajar carioca	Antenor Nascentes	1922
O falar mineiro	José A. de Oliveira	1934
Linguagem de Goiás	José A. de Oliveira	1944
A influência africana no português do Brasil	Renato Mendonça	1933
A língua portuguesa no Brasil	Virgílio Lemos	1916
A gramática e a evolução da língua portuguesa	Herbert Parentes Fortes	1933
A língua do Brasil	Luís Viana Filho	1936
O problema da língua brasileira	Sílvio Elia	1940

continua...

continuação

Língua brasileira	Edgard Sanches	1940
A língua do Brasil	Gladstone Chaves de Melo	1946
Introdução ao estudo da língua portuguesa	Serafim da Silva Neto	1950

Fonte: extraído de Elia (2003).

Com base nas obras do quadro 12 e várias outras, o livro apresenta análises sobre o quadro histórico, literatura e língua (incluindo análises sintáticas, de léxico e morfológica) para o período colonial (dividido entre século XVI, XVII e XVIII) e período do Brasil independente (ELIA, 2003).

O monitoramento que se propõe como procedimento de gestão do acervo de documentos arquivísticos de guarda permanente possui quatro características essenciais:

(1) delimitação de acervo;
(2) utilização de todas as fontes documentais disponíveis;
(3) intervalos regulares de registro;
(4) independência tecnológica.

Primeiro, é preciso delimitar o acervo que será considerado no procedimento. Em termos práticos, ao não delimitar o acervo considerado, podem surgir limites de recursos que impeçam o trabalho. Trata-se de procedimento não trivial e envolve várias etapas indefinidamente ou enquanto o acervo continuar existindo. Documentos que não tenham sido definidos como sendo de guarda permanente deveriam ser excluídos do conjunto total. Caso se trate de um fundo aberto (ainda recebendo novos documentos de guarda permanente), será inevitável receber novos documentos, mas em geral essa quantidade é menor, desde que os procedimentos de gestão documental (avaliação e seleção) tenham sido corretamente executados.

Segundo, a pesquisa permanente sobre o monitoramento da mudança deve se basear em fontes confiáveis como dicionários de primeira linha, normas (como acordos ortográficos) e estudos linguísticos como o citado anteriormente nesta seção. Estudos arquivísticos sobre as origens e funcionamento da instituição ou pessoa que deu origem aos documentos também podem ser considerados. Da mesma forma, produtos e procedimentos de análise documentária como tesauros, vocabulários controlados e socioterminologias podem ser também considerados como fontes importantes para o registro das mudanças.

Terceiro, a periodicidade das atualizações deve ser constante. O intervalo para atualizar os registros dependerá de vários fatores, como o tamanho do acervo e

disponibilidade de recursos (pelo menos pessoas). Além disso, trata-se de uma variável que deve ser experimentada para definir, com mais segurança, o melhor intervalo de anos para que as atualizações ocorram. Intervalos de uma ou duas décadas parecem um valor adequado. Com intervalos muito longos, corre-se o risco de assumir uma postura (re)ativa. O que significa essa postura será tema da seção seguinte.

Quarto, com independência tecnológica, quer-se dizer que os procedimentos de registro da ML devem evitar ao máximo a dependência tanto de sistemas como de instrumentos que podem vir a ser desativados, comprometendo, assim, o prosseguimento do monitoramento, o qual deve ser permanente. Por exemplo, veja-se uma instituição com uma (socio)terminologia que englobe todos os elementos linguísticos no acervo e, além disso, nela existam procedimentos para atualização da terminologia em questão. Ocorre, porém, que, em algum momento, pode-se preterir esse instrumento para utilizar alguma outra tecnologia, ou mesmo não fazer nada em substituição. Se o monitoramento linguístico dos elementos do acervo depender exclusivamente desse instrumento, então isso aumenta o risco de os efeitos da ML afetarem a recuperação no futuro, já que haverá uma interrupção no monitoramento.

Finalmente, é preciso tecer comentários sobre a aplicabilidade das informações obtidas no processo de monitoramento para uso nos futuros SRIs. Apesar de ser sabido que qualquer tecnologia para recuperação da informação não pode prescindir de elementos linguísticos tanto na representação quanto na recuperação e que essa situação não deve mudar ao longo do tempo, ao mesmo tempo não é possível saber qual tecnologia específica será utilizada no futuro. Como será a indexação e a recuperação automática em texto integral? De que maneira ocorrerá a integração com ontologias ou linguagens documentárias? Em relação ao longo prazo, é impossível responder a essas perguntas hoje. Esse é um limite neste trabalho. Simplesmente não é possível antecipar como será a integração com os futuros sistemas, pois não se pode prever como serão esses sistemas. O princípio adotado aqui é o de subsidiar, com informações linguísticas sobre a ML, esses futuros sistemas.

Uma última questão sobre o monitoramento vem de uma proposta teórica recente da linguística sobre a língua e sua evolução. Trata-se da hipótese de que uma língua só pode ser compreendida em seus variados "contextos de uso". "A associação entre pressupostos variacionistas e pressupostos funcionalistas se apoia essencialmente num ponto de partida comum: o de que a língua só pode ser entendida nos seus variados contexto de uso" (WEINREICH; LABOV; HERZOG, 2006, p. 147). Estando essa hipótese correta, pelo menos em parte, isso implica monitorar as mudanças numa língua, a partir de seu uso original (inclusive contextualizando a instituição que produziu determinados elementos linguísticos),

muito mais do que tentar "resgatar" os aspectos linguísticos do passado.

Todos os elementos linguísticos, em todos os aspectos linguísticos presentes no acervo considerado, devem ser relacionados e submetidos ao monitoramento. A análise documental executada na parte sobre metodologia e métodos ilustra esse procedimento para o caso de documentos produzidos no passado.

14.4 Possíveis soluções já disponíveis

Uma pergunta pertinente se estamos propondo uma possível solução ao problema dos efeitos da ML em SRIs no futuro é: existe(m) solução(ões) já disponíveis que possam ser utilizadas para o mesmo propósito que o monitoramento das mudanças? (como adiante apresentamos) e esta(s) solução(ões) são mais eficientes que a solução do monitoramento por nós aventada?

Uma linha de pesquisa para responder estas perguntas pode ser traçada a partir de outras perguntas mais específicas. Algumas delas são apresentadas a seguir.

O monitoramento da ML por nós proposto tem como objetivo principal subsidiar com informações linguísticas, no escopo da ML, SRIs. De forma a possibilitar a solução de ambiguidades especificamente derivadas da ML. Sendo assim, há outras soluções que tenham o mesmo objetivo principal?

O procedimento de monitoramento da ML é específico para registro da ML num acervo considerado. Sendo assim, há outras soluções específicas para acervos documentais?

O procedimento de monitoramento da ML é específico para documentos históricos que precisem ser mantidos indefinidamente. Na verdade, acompanhar os efeitos da ML é algo que só faz sentido durante um longo prazo: muitas décadas e séculos. Sendo assim, há outras soluções que sejam utilizadas para o mesmo escopo e tipo de documento?

Finalmente, o procedimento de monitoramento da ML é específico para subsidiar SRIs no longo prazo. Sendo assim, há outras soluções que também sejam planejadas para esta perspectiva?

Estas são perguntas iniciais que podem ser utilizadas numa pesquisa de sondagem e comparação com outros recursos informacionais já disponíveis. Na área de linguística, as pesquisas sobre filologia e linguística histórica podem suprir e responder estas perguntas? Em ciência da informação, instrumentos como socioterminologias, ontologias ou tesauros podem suprir estas necessidades? Áreas novas como as humanidades digitais (que tratamos na revisão de literatura) poderiam suprir e responder estas perguntas?

Além de responder estas perguntas, é preciso também analisar até que ponto outras possíveis soluções são mais eficientes que a solução do monitoramento da ML aqui proposta, ainda que em fase inicial. Esta pergunta só poderá ser respondida na medida em que a solução do monitoramento da ML estiver mais maduro e testado, o que ainda não ocorreu.

REFERÊNCIAS

ABBAGNANO, Nicola. **Dicionário de filosofia**. 5. ed. São Paulo: Martins Fontes, 2007.

AFONSO, Alexandre Ribeiro. **B2**: um sistema para indexação e agrupamento de artigos científicos em português brasileiro utilizando computação evolucionária. 2013. 158 f. Tese (Doutorado em Ciência da Informação)–Programa de Pós-Graduação em Ciência da Informação, Faculdade de Ciência da Informação, Universidade Federal de Brasília, Brasília, 2013.

AGUIAR, Francisco Lopes de. **O controle de vocabulário como dispositivo metodológico para a organização, tratamento e recuperação da informação arquivística**. 2008. 267 f. Dissertação (Mestrado)–Programa de Pós-Graduação em Ciência da Informação, Pontífícia Universidade Católica de Campinas, São Paulo, 2008. Disponível em: <http://www.bibliotecadigital.puc-campinas.edu.br/tde_busca/arquivo.php?codArquivo=437>. Acesso em: 15 ago. 2015.

ALMEIDA, Carlos Candido de. **Elementos de linguística e semiologia na organização da informação.** São Paulo: Cultura Acadêmica, 2011.

ALMEIDA, Daniela Pereira dos Reis et al. Paradigmas contemporâneos da ciência da informação: a recuperação da informação como ponto focal. **Revista Eletrônica Informação e Cognição**, Marilia, v. 6, n. 1, p. 16-27, 2007. Disponível em: <http://www2.marilia.unesp.br/revistas/index.php/reic/article/view/745>. Acesso em: 07 out. 2015.

ALVARENGA, Lídia. Representação do conhecimento na perspectiva da Ciência da Informação em tempo e espaços digitais. **Revista Eletrônica Biblioteconomia e Ciência da Informação**, Florianópolis, n. 15, p. 18-40, 2003. Disponível em: <https://periodicos.ufsc.br/index.php/eb>. Acesso em: 19 mar. 2015.

ALVES, I. M. Neologia e níveis de análise linguística. In: ISQUERDO, A. N.; ALVES, I. M. **As ciências do léxico**. Campo Grande: Editora UFMS, 2007. v. 2, p. 77-92.

ALVES, Rachel Cristina Vesu et al. Ciência da Informação, Ciência da Computação e Recuperação da Informação: algumas considerações sobre os métodos e tecnologias da informação utilizados ao longo do tempo. **Revista Eletrônica Informação e Cognição**, Marilia, v. 6, n. 1, p. 28-40, 2007. Disponível em: <http://www.brapci.ufpr.br/download.php?dd0=8410>. Acesso em: 15 ago. 2015.

ARAMPATZIS, Avi et al. Linguistically motivated information retrieval. **Encyclopedia of Library and Information Science**, v. 69, december, 2000. Online. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.31.7928>. Acesso em: 07 out. 2015.

ARBOIT, Aline Elis. **O processo de institucionalização sociocognitiva do domínio de organização do conhecimento a partir dos trabalhos científicos dos congressos da ISKO.** 2014. 285 f. Tese (Doutorado em Ciência da Informação)–Programa de Pós-Graduação em Ciência da Informação, Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2014.

ARELLANO, Miguel Ángel Márdero. Preservação de Documentos Digitais. **Ciência da Informação**, Brasília, v. 33, n. 2, p. 15-27, maio/ago. 2004. Disponível em: <http://revista.ibict.br/index.php/ciinf/article/view/305/270>. Acesso em: 31 jan. 2016.

ARQUIVO NACIONAL. **Dicionário brasileiro de terminologia arquivística**. Rio de Janeiro, 2005.

CAMARGO, Ana Maria de; BELLOTTO, Heloísa Liberalli. **Dicionário de Terminologia Arquivística**. Associação dos Arquivistas Brasileiros: São Paulo, 1996.

BAGNO, Marcos. **Preconceito linguístico**: o que é, como se faz. São Paulo: Loyola, 1999.

BAIR, Sheila A.; CARLSON, Sharon. Where keywords fail: using metadata to facilitate digital humanities scholarship. **University Libraries Faculty & Staff Publications**, [S.l.], paper 12, 2008. Disponível em: <http://scholarworks.wmich.edu/cgi/viewcontent.cgi?article=1012&context=library_pubs>. Acesso em: 29 jul. 2015.

BARANOW, Ulf Gregor. Perspectivas na contribuição da linguística e de áreas afinas à ciência da informação. **Ciência da Informação**, Brasília, v. 12, n. 1, p. 23-35, 1983.

BEL-ENGUIX, Gemma; JIMÉNEZ-LÓPEZ, M. Dolores (Ed.). **Language as a complex system**: interdisciplinary approaches. NewCastle: Cambridge Scholars Publishing, 2010.

BILETZKI, Anat; MATAR, Anat. Ludwig Wittgenstein. **Stanford Encyclopedia of Philosophy**, 3 Mar. 2014. Disponível em: <http://plato.stanford.edu/archives/spr2014/entries/wittgenstein/>. Acesso em: 10 set. 2015.

BOCCATO, Vera Regina Casari. **Avaliação do uso de linguagem documentária em catálogos coletivos de bibliotecas universitárias**: um estudo sociocognitivo com protocolo verbal. 2009. 301 f. Tese (Doutorado em Ciência da Informação)–Programa de Pós-Graduação em Ciência da Informação, Universidade Estadual Paulista, Marília, 2009.

BOOTH, Wayne C.; COLOMB, Gregory G.; WILLIAMS, Joseph M. **The craft of research.** 3. ed. Chicago: The university of Chicago Press, 2008.

BORBA, Francisco da Silva. **Introdução aos estudos linguísticos**. 9. ed. São Paulo: Editora Nacional, 1986.

BOTTÉRO, Jean. A escrita e a formação da inteligência na mesopotâmia antiga. In: _____. **Cultura, Pensamento e Escrita**. São Paulo: Ática, 1995. p. 9-45.

BRASIL. Senado Federal. **Constituição da República Federativa do Brasil**. Brasília, 1988.

BRÉAL, Michel. **Semantics**: studies in the science of meaning. New York: Dover Publications, 1964.

_____. **Semantics**: studies in the science of meaning. Translated by Mrs. Henry Cust. London: William Heinemann, 1900. Disponível em: <https://archive.org/details/semanticsstudie00postgoog>. Acesso em: 15 jan. 2016.

BRUNELLE, Marc. Diglossia, bilingualism, and the revitalization of written eastern cham. **Language, documentation & conservation**, [S.l.], v. 2, n. 1, p. 28-46, Jun. 2008. Disponível em: <http://hdl.handle.net/10125/1848>. Acesso em: 12 ago. 2015.

BURKE, Peter. **Uma história social do conhecimento**: de gutenberg a diderot. Rio de Janeiro: Zahar, 2003.

CALVET, Louis-Jean. **Saussure:** pró e contra para um linguística social. São Paulo: Editora Cultrix, 1975.

CAMPOS, Maria Luiza de Almeida. Linguagem documentária: teorias que fundamentam sua elaboração. Niterói: Eduff, 2001.

CAMPOS, Maria Luiza de Almeida. Indexação e descrição em arquivos: a questão da representação e recuperação de informações. **Arquivo & Administração**, Rio de Janeiro, v. 5, n. 1, p. 17-32, jan./jun. 2006.

CANDIDO JUNIOR, Arnaldo. **Criação de um ambiente para o processamento de córpus de Português Histórico.** 2008. 142 f. Dissertação (Mestrado em Ciências da Computação)– Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2008.

CAPUANO, Ethel Airton. **Mineração e modelagem de conceitos como praxis de gestão do conhecimento para inteligência competitiva.** 2010. 236 f. Tese (Doutorado em Ciência da Informação)–Programa de Pós-Graduação em Ciência da Informação, Faculdade de Ciência da Informação, Universidade de Brasília, Brasília, 2010.

CERVANTES, Brígida Maria Nogueira. **A construção de tesauros com a integração de procedimentos terminográficos.** 2009. 209 f. Tese (Doutorado em Ciência da Informação)– Programa de Pós-Graduação em Ciência da Informação, Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2009.

CERVO, Amado Luiz; BERVIAN, Pedro Alcino. **Metodologia científica.** 5. ed. São Paulo: Pearson Prentice Hall, 2002.

CHILD, William. **Wittgenstein**. Porto Alegre: Penso, 2013.

CHOMSKY, Noam. **Linguagem e mente**: pensamentos atuais sobre antigos problemas. Brasília: Universidade de Brasília, 1998.

CINTRA, Anna Maria Marques. Elementos de linguística para estudos de indexação. **Ciência da Informação**, Brasília, n. 12, p. 5-22, 1983.

COELHO; Izete Lehmkuhl et al. **Para conhecer sociolinguística**. São Paulo: Contexto, 2015.

CONSELHO INTERNACIONAL DE ARQUIVOS. **ISAD(G):** Norma geral internacional de descrição arquivística: adotada pelo Comitê de Normas de Descrição, Estocolmo, Suécia, 19-22 de setembro de 1999, versão final aprovada pelo CIA. 2. ed. Rio de Janeiro: Arquivo Nacional, 2000.

COULMAS, Florian. **Escrita e sociedade**. São Paulo: Parábola Editorial, 2014.

CRYSTAL, David. **Que é linguística**? Rio de Janeiro: Ao livro técnico, 1981.

DALBELLO, Marija; VAMANU, Iulian. Conceptualizations of cultural heritage in information science. **ASSIST**, Pittsburgh, 22-27 Oct. 2010. Disponível em: <https://www.asis.org/asist2010/proceedings/proceedings/ASIST_AM10/submissions/375_Final_Submission.pdf>. Acesso em: 12 ago. 2015.

DAL'EVEDOVE, Paula Regina. **O tratamento temático da informação em abordagem sociocultural**: diretrizes para definição de política de indexação em bibliotecas universitárias. 2014. 266 f. Tese (Doutorado em Ciência da Informação)–Programa de Pós-Graduação em Ciência da Informação, Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2014.

DALGAAARD, Rune. Hypertext and the scholarly archive: intertexts, paratexts and metatexts at work. **ACM**, New York, p. 175-184, 2001. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.23.4783&rep=rep1&type=pdf>. Acesso em: 15 ago. 2015.

DERRIDA, Jacques. **Gramatologia**. São Paulo: Perspectiva, 2011.

DERVIN, Brenda. Sense-making theory and practive: an overview of user interests in knowledge seeking and use. **Journal of Knowledge Management**, [S.l.], v. 2, n. 2, p. 36-46, Dec. 1998. Disponível em: <http://www.emeraldinsight.com/doi/abs/10.1108/13673279810249369>. Acesso em: 05 out. 2015.

DIAS, Fernando Skackauskas. **Migração conceitual entre sistemas de recuperação e ciências cognitivas**. 2011. 175 f. Tese (Doutorado em Ciência da Informação)–Programa de Pós-Graduação em Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2011.

DODEBEI, Vera. Cultura digital: novo sentido e significado de documento para a memória social? **DataGramaZero**, Rio de Janeiro, v. 12, n. 2, abr. 2011. Disponível em: <http://www.dgz.org.br/abr11/Art_01.htm>. Acesso em: 19 mar. 2015.

DRAAISMA, Douwe. **Metáforas da memória**: uma história das ideias sobre a mente. Bauru: Edusc, 2005.

DUBOIS, Jean et al. **Dicionário de linguística**. 10. ed. São Paulo: Cultrix, 1998.

DUCROT, Oswald; TODOROV, Tzvetan. **Dicionário enciclopédico das ciências da linguagem**. São Paulo: Perspectiva, 2010.

ECO, Umberto. **Como se faz uma tese**. São Paulo: Perspectiva, 2006.

_____. **Tratado Geral de Semiótica**. São Paulo: Perspectiva, 2012.

EDMONDSON, Ray. **Memória do mundo:** diretrizes para a salvaguarda do patrimônio documental. [S.l.]: Divisão da Sociedade da Informação; Organização das Nações Unidas para Educação, Ciência e Cultura, 2002. Disponível em: <http://www.portalan.arquivonacional.gov.br/Media/Diretrizes%20para%20a%20salvaguarda%20do%20patrim%C3%B4nio%20documental.pdf>. Acesso em: 20 jan. 2016.

ELIA, Sílvio. **Fundamentos histórico-linguísticos do português do Brasil**. Rio de Janeiro: Lucerna, 2003.

_____. **Sociolinguística:** uma introdução. Rio de Janeiro: Padrão, 1987.

ESTEBAN NAVARRO, Miguel Ángel. La representación y la organización del conocimiento en los archivos. In: MARCO, Fco. Javier Garcia (Ed.). **Organización del conocimiento en sistemas de información y documentación**: Atas del I Encuentro de ISKO-Espanha, Madrid, 4-5 novembro, 1993. Zaragoza: [s.n], 1995.

FARACO, Carlos Alberto. **Linguística histórica**: uma introdução ao estudo da história das línguas. São Paulo: Parábola Editorial, 2005.

FERNEDA, Edberto. **Recuperação de informação**: análise sobre a contribuição da ciência de computação para a Ciência da Informação. 2003. 147 f. Tese (Doutorado em Ciências da Comunicação)–Escola de Comunicação e Artes, Universidade de São Paulo, São Paulo, 2003.

FINNEMANN, Niels Ole. **Hypertext and the representational capacities of the binary alphabet**. [S.l.; s.n.], 1999. Disponível em: <http://www.hum.au.dk/ckulturf/pages/publications/nof/hrc.pdf>. Acesso em: 15 ago. 2015.

FIORIN, José Luiz. A criação dos cursos de letras no Brasil e as primeiras orientações da pesquisa linguística universitária. **Línguas & Letras**, Cascavel, v. 7, n. 12, p. 11-25, 2006.

_____ (Org.). **Linguística? O que é isso?** São Paulo: Contexto, 2013.

FISCHER, S. R. **A history of writing**. London: Reaktion Books, 2003.

FUNARI, Pedro Paulo Abreu. **Antiguidade clássica**: a história e a cultura a partir dos documentos. 2. ed. Campinas: Editora UNICAMP, 2003.

FURNAS, G. W. et al. The vocabulary problem in human-system communication. **Communications of the ACM**, New York, v. 30, n. 11, p. 964-971, Nov. 1987. Disponível em: <http://dl.acm.org/citation.cfm?id=32212>. Acesso em: 29 jul. 2015.

GARCIA, Rodrigo Moreira. **Modelos de comportamento de busca de informação**: contribuições para a organização da informação. 2007. 139 f. Dissertação (Mestrado)–Programa de Pós-Graduação em Ciência da Informação, Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2007.

GARDIN, Jean-Claude. Document analysis and linguistic theory. **Journal of Documentation**, [S.l.], v. 29, n. 2, p. 137-168, 1973. Disponível em: <http://dx.doi.org/10.1108/eb026553>. Acesso em: 05 set. 2015.

GARRETT, Jeffrey. KWIC and Dirty? human cognition and the claims of full-text searching. **The Journal of Electronic Publishing**, [S.l.], v. 9, n. 1, winter 2006. Disponível em: <http://quod.lib.umich.edu/j/jep/3336451.0009.106?view=text;rgn=main>. Acesso em: 29 jul. 2015.

GELB, I. J. **A study of writing**: a discussion of the general principles governing the use and evolution of writing. Chicago: Chicago University Press, 1952.

GIL, Antônio Carlos. **Métodos e técnicas de pesquisa social**. 5. ed. São Paulo: Atlas, 2006.

GODOIS, Janette Mariano; DALPIAN, Laurindo. Semântica: um estudo diacrônico. In: SEMINÁRIO INTERNACIONAL EM LETRAS: LÍNGUA E LITERATURA NA PÓS-MODERNIDADE, 12., 2012, Santa Maria. **Anais...** Santa Maria: UNIFRA, 2012. Disponível em: <http://www.unifra.br/eventos/inletras2012/Trabalhos/4693.pdf>. Acesso em: 15 ago. 2015.

GOMES, Daniel et al. Introducing the portuguese web archive initiative. In: INTERNATIONAL WEB ARCHIVING WORKSHOP, 8., 2008, Aarhus. **Annals…** Aarhus: [s.n.], 2008. Disponível em: <http://comum.rcaap.pt/handle/123456789/470>. Acesso em: 15 ago. 2015.

GONÇALVES, Maria Filomena; BANZA, Ana Paula (Coords.). **Património textual e humanidades digitais**: da antiga à nova Filologia. Évora: CIDEHUS, 2013.

GREIMAS, A. J.; COURTÉS, J. **Dicionário de semiótica**. 2. ed. São Paulo: Contexto, 2012.

GRETE, Seland. **User revealment revisited:** knowledge formation in the prefocus stage of information-based work tasks. 2014. 464 f. Tese (PhD)–Institut for Kommunikation, Aalborg Universitet, Aalborg, 2014. Disponível em: <http://vbn.aau.dk/files/201270036/PhdThesis_SelandGrete_20140622.pdf>. Acesso em: 29 jul. 2015.

GUNDER, Ana. Forming the Text, Performing the Work - Aspects of Media, Navigation, and Linking. **Humanit,** Borás, v. 5, n. 2-3, 2001. Disponível em: <http://etjanst.hb.se/bhs/ith/23-01/ag.htm>. Acesso em: 15 jan. 2016.

GUY, Gregory Riordan; ZILLES, Ana. **Sociolinguística quantitativa:** instrumental de análise. São Paulo: Parábola Editorial, 2007.

HIEMSTRA, Djoerd. A linguistically motivated probabilistic model of information retrieval. **ECDL**, [S.l.], p. 569-584, 1998. Proceedings. Disponível em: <http://link.springer.com/chapter/10.1007%2F3-540-49653-X_34>. Acesso em: 07 out. 2015.

HJORLAND, Birger. Documents, memory institutions and information science. **Journal of Documentation**, [S.l.], v. 56, n. 1, p. 27-41, 2000. Disponível em: <http://dx.doi.org/10.1108/EUM0000000007107>. Acesso em: 12 ago. 2015.

HUISMAN, Denis. **Dicionário de obras filosóficas**. São Paulo: Martins Fontes, 2000.

HUNNEX, Milton D. **Filósofos e correntes filosóficas em gráficos e diagramas**: conheça melhor os filósofos e as correntes filosóficas por meio de gráficos e diagramas temáticos e cronológicos. São Paulo: Editora Vida, 2003.

INGWERSEN, Peter. Cognitive perspectives of information retrieval interaction: elements of a cognitive IR theory. **Journal of Documentation**, [S.l.], v. 52, n. 1, p. 3-50, 1996. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.116.2558&rep=rep1&type=pdf>. Acesso em: 07 out. 2015.

_____; JARVELIN, Kalervo. **The turn**: integration of information seeking and retrieval in contexto. Netherlands: [s.n.], 2005.

JANSON, Tore. **A história das línguas**: uma introdução. São Paulo: Parábola Editorial, 2015.

JATOWT, Adam; DUH, Kevin. A framework for analyzing semantic change of words across time. **JCDL,** London, p. 229-238, Sep. 2014. Disponível em: <http://dl.acm.org/citation.cfm?id=2740809&dl=ACM&coll=DL&CFID=531915645&CFTOKEN=32321217>. Acesso em: 29 jul. 2015.

JONES, Karen Sparck; KAY, Martin. **Linguistics and information science**. New York: Academic Press, 1973.

KARLGREN, Jossi. **Information retrieval**: statistics and linguistics. Sweden: Swedish Institute of Computer Science, 2000. Disponível em: <http://ccl.pku.edu.cn/doubtfire/NLP/Information_Retrieval/Introduction/Information%20Retrieval%20Statistics%20and%20Linguistics%20by%20Jussi%20Karlgren.pdf>. Acesso em: 05 set. 2015.

KOBASHI, Nair Yumiko. Fundamentos semânticos e pragmáticos da construção de instrumentos de representação de informação. **DataGramaZero**, Rio de Janeiro, v. 8, n. 6, dez. 2007. Disponível em: <http://www.dgz.org.br/>. Acesso em 19 mar. 2015.

KOOLEN, Marijn et al. **Unified Access to Heterogeneous Data in Cultural Heritage**. [200-?]. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=8BB1830B1BC4A6B4016649BCCFB2B64E?doi=10.1.1.154.9781&rep=rep1&type=pdf>. Acesso em: 15 jan. 2016.

KOOLEN, Marijn; KAMPS, Jaap; KEIJZER, Vincent. Information Retrieval in cultural heritage. **Interdisciplinary science reviews**, [S.l.], v. 34, n. 2-3, p. 267-284, 2009. Disponível em: <http://humanities.uva.nl/~kamps/publications/2009/kool:info09.pdf>. Acesso em: 20 jan. 2016.

KRAAIJ, Wessel. **Variations on language modeling for information retrieval**. 2004. 287 f. Tese (Doutorado)–Centre for telematics and information technology, Netherlands, Enschede,

2004. Disponível em: <http://doc.utwente.nl/41478/1/thesis_Kraaij.pdf>. Acesso em: 15 ago. 2015.

LANCASTER, F. W. **Indexação e resumos**: teoria e prática. 2. ed. rev. atual. Brasília: Briquet de Lemos, 2004.

LAVILLE, Christian; DIONNE, Jean. **A construção do saber**: manual de metodologia da pesquisa em ciências humanas. Porto Alegre: Artmed, 1999.

LEIKIN, Mark; IBRAHIM, Raphiq; EGHBARIA, Hazar. The influence of diglossia in arabic on narrative ability: evidence from analysis of the linguistic and narrative structure of discourse among pre-school children. **Reading and Writing**, [S.l.], v. 27, n. 4, p. 733-747, Apr. 2014. Disponível em: <http://link.springer.com/article/10.1007%2Fs11145-013-9462-3>. Acesso em: 15 ago. 2015.

LEITE, Maria Angélica de Andrade. **Recuperação da informação, representação do conhecimento, ontologia, sistemas difusos**. 2009. Tese (Doutorado)–Programa de Pós-Graduação em Engenharia Elétrica, Faculdade de Engenharia Elétrica e de Computação, Universidade Estadual de Campinas, Campinas, 2009.

LEROY, Maurice. **As grandes correntes da linguística moderna**. São Paulo: Cultrix, 1967.

LEWIS, David; JONES, Karen Sparck. Natural language processing for information retrieval. **Communications of the ACM**, New York, v. 99, n. 1, p. 92-101, 1996. Disponível em: <https://www.cl.cam.ac.uk/archive/ksj21/ksjdigipapers/cacm96.pdf>. Acesso em: 07 out. 2015.

LIEBERMAN, Philip. The evolution of human speech: its anatomical and neural bases. **Current Anthropology**, [S.l.], v. 48, n. 1, p. 39-66, Feb. 2007.

LIMA, Vânia Mara Alves. **Da classificação do conhecimento científico aos sistemas de recuperação de informação**: enunciação de codificação e enunciação de decodificação da informação documentária. 2004. 156 f. Tese (Doutorado em Ciências da Comunicação)–Escola de Comunicação e Artes, Universidade de São Paulo, São Paulo, 2004.

_____; BOCCATO, Vera Regina Casari. O desempenho terminológico dos descritores em ciência da informação do vocabulário controlado do SIBi/USP nos processos de indexação manual, automática e semi-automática. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 14, n. 1, p. 131-151, jan./abr. 2009. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-99362009000100010>. Acesso em: 29 jul. 2015.

_____; VITORIANO, Marcia Cristina de Carvalho Pazin; BARBANTI, Cristina Hilsdorf. Organização do conhecimento e o patrimônio industrial em São Paulo: o projeto eletromemória. In: GUIMARÃES, José Augusto Chaves; DODEBEI, Vera (Orgs.). **Organização do conhecimento e diversidade cultural**. Marília: ISKO-Brasil; FUNDEPE, 2015. p. 565-573. (Estudos Avançados em Organização do Conhecimento, v. 3).

LYONS, John. **Linguagem e linguística**: uma introdução. Rio de Janeiro: LTC, 1987.

MAIMONE, Giovana Deliberali; TÁLAMO, Maria de Fátima Gonçalves Moreira. Linguística e terminologia: contribuições para a elaboração de tesauros em ciência da informação. **DataGramaZero**, Rio de Janeiro, v. 12, n. 2, abr. 2011. Disponível em: <http://www.dgz.org.br/abr11/Art_05.htm>. Acesso em: 05 set. 2015.

MANNING, Christopher; RAGHAVAN, Prabhakar; SCHUTZE, Hinrich. **An introduction do information retrieval**. Cambridge: Cambridge University Presse, 2009. Disponível em: <nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf>. Acesso em: 15 ago. 2015.

MARCONDES, Danilo. **Textos básicos de linguagem**: de Platão a Foucault. Rio de Janeiro: Zahar, 2010.

MARTINS, Agnaldo Lopes. **O uso do sintagma nominal na recuperação de documentos**: proposta de um mecanismo automático para classificação temática de textos digitais. 2014. 192 f. Tese (Doutorado em Ciência da Informação)–Programa de Pós-Graduação em Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2014a.

MARTINS, Gilberto de Andrade; THEÓPHILO, Carlos Renato. **Metodologia da investigação científica para ciências sociais aplicadas.** 2. ed. São Paulo: Atlas, 2007.

MARTINS, Gracy Kelli. **Institucionalização cognitiva e social da organização e representação do conhecimento na ciência da informação no Brasil**. 2014. 184 f. Tese (Doutorado em Ciência da Informação)–Programa de Pós-Graduação em Ciência da Informação, Universidade Estadual Paulista, Marília, 2014b.

MARTINS, Marcio Souza; LIMA, Vânia Mara Alves. A abordagem social na recuperação da informação: frente e tendências de pesquisa. **Biblios:** Revista de Bibliotecología y Ciencias de la Información, Pittsburgh, n. 52, p. 1-15, 2013.

MAUTNER, Thomas. **Dicionário de filosofia**. Lisboa: Edições 70, 2011.

MELO, Fabio J. Dantas de; BRÄSCHER, Marisa. **Fundamentos da linguística para a formação do profissional da informação**. Brasília: Centro Editorial, 2011.

MENDES, Ronald Beline. Língua e variação. In: FIORIN, José Luiz (Org.). **Linguística? O que é isso?** São Paulo: Contexto, 2013.

MENDONÇA, Ercilia Severina. A linguística e a ciência da informação: estudos de uma interseção. **Ciência da Informação**, Brasília, v. 29, n. 3, p. 50-70, set./dez. 2000. Disponível em: <http://www.scielo.br/pdf/ci/v29n3/a06v29n3.pdf>. Acesso em: 05 set. 2015.

MILLER, Alexander. **Filosofia da linguagem**. São Paulo: Paulus, 2010.

MONTEIRO, Silvana; CARELLI, Ana; PICLKER, Maria Elisa. Representação e memória no ciberespaço. **Ciência da Informação**, Brasília, v. 35, n. 3, p. 115-123, set./dez. 2006.

MONTGOMERY, Christine A. Linguistics and information science. **Journal of the American Society for Information Science**, [S.l.], v. 2, n. 3, p. 195-219, May/Jun. 1972. Disponível em: <http://eric.ed.gov/?id=EJ060830>. Acesso em: 05 set. 2015.

MOREIRA, Manoel Palhares. **Ambiente para geração e manutenção semi-automática de tesauros.** 2005. 197 f. Tese (Doutorado em Ciência da Informação)–Programa de Pós-Graduação em Ciência da Informação, Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2005.

MORSY, Sara; KARYPIS, George. Accounting for language changes over time in document similarity search. **Technical Report**, Minneapolis, p. 1-12, Jul. 2015. Disponível em: <https://www.cs.umn.edu/sites/cs.umn.edu/files/tech_reports/15-011.pdf>. Acesso em: 29 jul. 2015.

MUNIZ, Marcelo Caetano Martins. **A construção de recursos lingüístico-computacionais para o português do Brasil**: o projeto Unitex-PB. 2004. 92 f. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional)–Instituto de Ciências Matemáticas e de Computação, São Carlos, 2004.

_____; NUNES, Maria das Graças V.; LAPORTE, Eric. UNITEX-PB, a set of flexible language resources for Brazilian Portuguese. In: CONGRESSO SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 25., 2008, São Leopoldo. **Anais...** São Leopoldo: UNISINOS, 2008, p. 2059-2058. Disponível em: <http://www.nilc.icmc.usp.br/til/til2005/arq0102.pdf>. Acesso em: 31 jan. 2016.

NOVELLINO, Maria Salet Ferreira. Instrumentos e metodologias de representação da informação. **Informação & Informação**, Londrina, v. 1, n. 2, p. 37-45, jul./dez. 1996.

OGDEN, C. K; RICHARDS, I. A. **O significado de significado.** Rio de Janeiro: Zahar Editores, 1972.

_____; _____. **The meaning of meaning**: a study of the influence of language upon thought and of the science of symbolism. San Diego: Harcourt Brace Jovanovich, 1989.

OLIVEIRA, Eliane Braga; RODRIGUES, Georgete Medleg. O conceito de memória na ciência da informação: análise das teses e dissertações dos programas de pós-graduação no Brasil. **Liinc em Revista**, Rio de Janeiro, v. 7, n. 1, p. 311-328, mar. 2011. Disponível em: <http://dx.doi.org/10.18225/liinc.v7i1.416>. Acesso em: 15 ago. 2015.

ORTEGA, Cristina Dotta. **Informática documentária**: estado da arte. 2002. 259 f. Dissertação (Mestrado em Ciências da Comunicação)–Escola de Comunicações e Artes, Universidade de São Paulo, São Paulo, 2002. Disponível em: <http://www.teses.usp.br/teses/disponiveis/27/27143/tde...155935/.../Ortega.pdf>. Acesso em: 15 ago. 2015.

PAUMIER, Sébastien. **Unitex 3.0:** user manual. Champs-sur-Marne: université paris-est marne-la-vallée, 2003. Disponível em: <http://www-igm.univ-mlv.fr/~unitex/UnitexManual3.0.pdf>. Acesso em: 15 jan. 2016.

PAVÃO, Caterina Marta Groposo. **Comportamento de busca e recuperação da informação em serviços de descoberta em rede no contexto acadêmico**. 2014. 225 f. Tese (Doutorado em Comunicação e Informação)–Pós-Graduação em Comunicação e Informação, Faculdade de Biblioteconomia e Comunicação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2014.

PEIRCE, Charles Sanders. **Semiótica**. São Paulo: Perspectiva, 2012.

POTTIER, Bernard. **Nuevos enfoques sobre diacronía de sistemas**. Alicante: Biblioteca Virtual Miguel de Cervantes, 1992a. Disponível em: <http://www.cervantesvirtual.com/nd/ark:/59851/bmcq81q5>. Acesso em: 15 ago. 2015.

_____. **Sémantique générale**. Paris: Presses Universitaires de France, 1992b.

RAJAGOPALAN, Kanavillil. Os caminhos da pragmática no Brasil. In: **DELTA:** Documentação de Estudos em Linguísticas Teórica e Aplicada, São Paulo, v. 15, n. especial, p. 323-338, 1999.

RIBEIRO, Célia Pereira; PIRES, Erik André de Nazaré. A preservação da informação em relação ao patrimônio cultural na atualidade. **DataGramaZero**, Rio de Janeiro, v. 16, n. 1, fev. 2015.

RICHARDSON, Roberto Jarry. **Pesquisa social:** métodos e técnicas. 3. ed. São Paulo: Atlas, 2010.

ROLIM, Elizabeth Almeida; CENDÓN, Beatriz Valadares. Modelos teóricos de estudos de usuários na ciência da informação. **DataGramaZero**, Rio de Janeiro, v. 14, n. 2, abr. 2013. Disponível em: <http://www.dgz.org.br/abr13/Art_06.htm>. Acesso em: 07 out. 2015.

ROSS, Anthony John Charles. **Correspondents theory 1800/2000**: philosophical reflections upon epistolary technics and praxis in the analogue and digital. 2012. 229 f. Tese (Doctor of Philosophy)–Humanities Advanced Technology and Information Institute, College of Arts, University of Glasgow, Glasgow, 2012. Disponível em: <http://theses.gla.ac.uk>. Acesso em 15 ago. 2015.

ROTH, Wolfgang. A semântica histórica: um campo abandonado da linguística. **Filologia e Lingüística Portuguesa**, [S.l.], n. 2, p. 61-79, 1998.

SAMPSON, Geoffrey. **Sistemas de escrita**: tipologia, história e psicologia. São Paulo: Ática, 1996.

SANDERSON, Mark; CROFT, Bruce. The history of information retrieval research. **Proceedings of the IEEE**, [S.l.], v. 100, p. 1444-1451, May 2012. Disponível em: <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6182576>. Acesso em: 15 ago. 2015.

SARACEVIC, Tefko. Modeling interaction in information retrieval: a review and proposal. **Proceedings of the American Society for Information Science**, [S.l.], v. 33, p. 3-9, 1996. Disponível em: <http://eric.ed.gov/?id=EJ557152>. Acesso em: 15 ago. 2015.

SAUSSURE, Ferdinand de. **Curso de linguística geral**. 28. ed. São Paulo: Cultrix, 2012.

SEVERINO, Antonio Joaquim. **Metodologia do trabalho científico**. 18. ed. São Paulo: Cortez, 1992.

SHANNON, C. E. A Mathematical Theory of Communication. **ACM SIGMOBILE Mobile Computing and Communications Review,** New York, v. 5, n. 1, p. 3-55, Jan. 2001. Disponível em: <http://dl.acm.org/citation.cfm?doid=584091.584093>. Acesso em: 10 out. 2015.

SILVA, Daniela Lucas da. **Ontologias para representação de documentos multimídia**: análise e modelagem. 2014. 442 f. Tese (Doutorado em Ciência da Informação)–Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2014.

SILVA, Eliezer Pires da Silva; ORRICO, Evelyn Goyannes Dill. O trabalho de descrição de acervo arquivístico no Brasil. In: DOBEDEI, Vera; GUIMARÃES, José Augusto Chaves (Orgs.). **Complexidade e organização do conhecimento**: desafios do nosso século. Rio de Janeiro: ISKO-Brasil; Marília: FUNDEPE, 2013. p. 211-216. (Estudos Avançados em Organização do Conhecimento, v. 2). Disponível em: <http://isko-brasil.org.br/wp-content/uploads/2013/02/Estudos-avan%C3%A7ados-2.pdf>. Acesso em: 05 fev. 2016.

SILVA, Rosa Virgínia Mattos e. **Caminhos da linguística histórica**: ouvir o inaudível. São Paulo: Parábola editorial, 2008.

SILVA NETO, Serafim da. **História da língua portuguesa**. Rio de Janeiro: Livros de Portugal, 1952.

SILVEIRA, Fabrício José Nascimento da. Biblioteca pública, memória e discursos identitários: uma leitura sócio-histórica dos depoimentos colhidos pelo Projeto Memória Oral da Biblioteca Mário de Andrade (BMA). **Tendências da Pesquisa Brasileira em Ciência da Informação**, Belo Horizonte, v. 5, n. 1, p. 1-23, 2012.

SMIT, Johanna Wilhelmina; KOBASHI, Nair Yumiko. **Vocabulário controlado para aplicações em arquivos**. São Paulo: Arquivo do Estado, 2003.

_____; TÁLAMO, Maria de Fátima Gonçalves Moreira. Documentation: la mémoire e t les systèmes de recherche d`information. **Sciences de la Société**, Toulouse, n. 68, p. 177-190, 2006.

SOBEL, Karen; BEALL, Jeffrey. Humanities research, book digitization, and the problem of linguistic change. **Journal of Library Innovation,** [S.l.], v. 2, n. 2, p. 3-15, 2011. Disponível em: <http://www.libraryinnovation.org/article/view/99>. Acesso em: 29 jul. 2015.

SOUSA, Renato Tarciso Barbosa. A classificação como função matricial do que-fazer arquivístico. In: SANTOS, Vanderlei Batista dos; INNARELLI, Humberto Celeste; SOUSA, Renato Tarciso Barbosa de (Orgs.). **Arquivística:** temas contemporâneos: classificação, preservação digital, gestão do conhecimento. Brasília: SENAC, 2007. p. 79-163.

SOUZA, Maria Clara Paixão. Conceito material de "texto digital": um ensaio. **Texto Digital**, Florianópolis, v. 5, n. 2, p. 159-187, 2009. Disponível em: <http://dx.doi.org/10.5007/1807-9288.2009v5n2p159>. Acesso em: 31 jan. 2016.

SOUZA, William Eduardo Righini de; CRIPPA, Giulia. O campo da ciência da informação e o patrimônio cultural: reflexões iniciais para novas discussões sobre os limites da área.

Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação, Florianópolis, v. 15, n. 29, p. 1-23, 2010.

STYLTSVIG, Henrik Bulskov. **Ontology-based information retrieval**. 2006. 196 f. Dissertation (Doctor of Philosophy)–Faculties of Roskild University, Denmark, 2006. Disponível em: <http://akira.ruc.dk/~bulskov/thesis_bib.pdf>. Acesso em: 19 mar. 2015.

TAMBA-MECZ, Irène. **A semântica**. São Paulo: Parábola Editorial, 2006.

TARALLO, Fernando. **A pesquisa socio-linguística.** 6. ed. São Paulo: Editora Ática, 1999.

TRASK, R. L. **Dicionário de linguagem e linguística**. 2. ed. São Paulo: Contexto, 2006.

ULLMANN, Stephen. **Semântica**: uma introdução à ciência do significado. 3. ed. Lisboa: Fundação Calouste Gulbenkian, 1964.

_____. **The principles of semantics**: a linguistic approach to meaning. 2. ed. Glasgow: Jackson, Son & Co, 1957.

VANDERMEERSCH, Léon. Escrita e língua gráfica na china. In: BOTTÉRO, J. et al. (Orgs.). **Cultura, pensamento e escrita**. São Paulo: Ática, 1995. p. 47-63. (Coleção Múltiplas Escritas).

VAREJÃO, Filomena de Oliveira Azevedo. O português do Brasil: revisitando a história. **Cadernos de Letras da UFF**, Niterói, n. 39, p. 119-137, 2009. Disponível em: <http://www.uff.br/cadernosdeletrasuff/39/artigo6.pdf>. Acesso em: 10 ago. 2015.

VECHIATO, Fernando Luiz. **Encontrabilidade da informação**: contributo para uma conceituação no campo da ciência da informação. 2013. 206 f. Tese (Doutorado em Ciência da Informação)–Programa de Pós-graduação em Ciência da Informação, Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2013.

VERRI, Gilda Maria Whitaker. Das fontes do passado à memória em construção. **Tendências da Pesquisa Brasileira em Ciência da Informação**, Belo Horizonte, v. 5, n. 1, jan./dez. 2012. Disponível em: <http://inseer.ibict.br/ancib/index.php/tpbci/article/viewArticle/78>. Acesso em: 31 jan. 2015.

VILLAS, Alberto. **Pequeno dicionário brasileiro da língua morta.** São Paulo: Globo, 2012.

VIOTTI, Evani. Mudança linguística. In: FIORIN, José Luiz (Org.). **Linguística? Que é isso?** São Paulo: Contexto, 2013. p. 137-179.

WEEDWOOD, Barbara. **História concisa da linguística**. São Paulo: Parábola Editorial, 2002.

WEINREICH, Uriel; LABOV, William; HERZOG, Marvin I. **Fundamentos empíricos para uma teoria da mudança linguística**. São Paulo: Parábola Editorial, 2006.

APÊNDICE A – TABELA DE OCORRÊNCIAS DE TERMOS

	História			Linguística								Computação		Ciência da Informação						Grupo (1)	Grupo (2)	Grupo (3)	Grupo (4)
Teses	**Memória**	**Preservação**	**Doc Arquivo**	**Linguagem**	**Linguística**	**Linguística Histórica**	**Mudança Linguística**	**Variação Linguística**	**Sociolinguística**	**Diacronia**	**Língua Escrita**	**Sist. RI**	**Proc. Linguagem Natural**	**Arquivologia**	**Repr. Info**	**Rec. Info**	**Tesauro**	**Ontologia**	**Voc. Controlado**	Memória + Linguística	(memória) AND (linguística) AND (sist. De Rec. Da Info) AND (Repres. Da Info) AND (Recup. Da Info)	Memória + Linguística + Representação da Informação + Recuperação da Informação + Sistemas de RI + Mudança Linguística	Memória + Linguística + Representação da Informação + Recuperação da Informação + Sistemas de RI + Variação Linguística
001_UFMG	11	5	0	176	6	0	0	0	0	0	0	3	0	1	4	8	22	1546	9	1	1	0	0
002_UFMG	20	0	0	122	19	0	0	0	0	0	0	5	29	0	0	16	0	1	3	1	0	0	0
003_UFMG	6	4	0	27	0	0	0	0	0	0	0	0	1	0	0	7	21	256	0	0	0	0	0
004_UFMG	1	0	0	105	8	0	0	0	0	0	0	0	0	0	3	2	0	444	1	1	0	0	0
005_UFMG	3	8	0	18	0	0	0	0	0	0	0	2	0	2	0	9	5	2	2	0	0	0	0
006_UFMG	3	4	1	129	8	0	0	0	0	0	0	3	0	16	1	6	279	3	6	1	1	0	0
007_UFMG	29	4	0	135	0	0	0	0	0	0	0	0	2	1	0	22	127	21	4	0	0	0	0
008_UFMG	0	5	0	415	138	0	0	0	0	0	0	0	137	0	0	3	7	28	1	0	0	0	0
009_UFMG	12	0	0	37	0	0	0	0	0	0	0	5	0	0	0	45	9	2	5	0	0	0	0
010_UFMG	28	0	0	35	15	0	0	0	0	0	0	21	0	0	1	36	3	1	0	1	1	0	0
011_UFMG	2	0	0	64	2	0	0	0	0	0	1	0	23	0	1	15	0	81	1	1	0	0	0
012_UFMG	2	4	0	37	11	0	0	0	0	0	0	0	0	1	0	5	7	8	6	1	0	0	0
013_UFMG	11	1	0	93	38	0	0	0	1	0	1	8	26	0	3	24	1	6	0	1	1	0	0
014_UFMG	311	1	0	193	0	0	0	0	0	0	0	2	0	0	1	4	5	581	1	0	0	0	0
015_UFMG	0	1	0	111	40	0	0	0	0	6	0	0	18	1	22	4	5	4	0	0	0	0	0
016_UFMG	5	0	0	10	6	0	0	0	0	0	0	0	0	0	0	3	0	0	0	1	0	0	0
017_UFMG	18	10	0	112	0	0	0	0	0	0	0	17	11	0	0	48	15	2	2	0	0	0	0
018_UFMG	0	6	0	52	0	0	0	0	0	0	0	5	11	3	2	25	2	4	1	0	0	0	0
019_UFMG	20	20	11	40	0	0	0	0	0	0	0	0	0	121	1	9	27	150	1	0	0	0	0
020_UFMG	0	502	9	23	0	0	0	0	0	0	0	1	0	73	1	4	0	3	3	0	0	0	0
021_UFMG	76	0	15	0	0	0	0	0	0	0	0	0	0	108	0	0	0	0	0	0	0	0	0
022_UFMG	5	1	0	11	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
023_UFMG	49	17	0	10	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
001_UFRJ	41	30	0	788	67	0	0	0	2	3	1	0	0	6	0	7	5	20	0	1	0	0	0
002_UFRJ	2	0	0	688	0	0	0	0	0	0	0	1	0	0	3	18	3	2	0	0	0	0	0
001_UNB	31	1	0	22	0	0	0	0	0	0	0	0	0	0	1	21	17	1	1	0	0	0	0
002_UNB	7	1	0	18	12	0	0	0	0	0	0	0	0	0	16	4	0	0	1	1	0	0	0
003_UNB	27	5	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0
004_UNB	64	8	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
005_UNB	8	3	0	0	0	0	0	0	0	0	0	0	0	2	1	5	3	180	0	0	0	0	0
006_UNB	2	0	0	184	42	0	0	0	0	0	0	0	176	1	1	1	21	258	2	1	0	0	0
007_UNB	33	28	0	2	20	0	0	0	0	0	0	0	0	2	1	2	0	0	0	1	0	0	0
008_UNB	0	5	0	130	0	0	0	0	0	0	0	3	0	0	9	38	3	2	7	0	0	0	0
009_UNB	1	2	0	48	0	0	0	0	0	0	0	0	0	0	1	9	5	231	12	0	0	0	0
010_UNB	6	0	0	195	19	0	0	0	0	0	0	4	42	1	5	39	3	70	0	1	1	0	0
011_UNB	7	1	0	81	0	0	0	0	0	0	0	0	1	0	16	7	16	27	0	0	0	0	0
012_UNB	14	16	0	54	7	0	0	0	0	0	0	0	0	1	0	1	2	0	0	1	0	0	0
013_UNB	5	0	0	74	74	0	0	0	2	0	0	1	3	29	2	26	16	41	0	1	1	0	0
001_UNESP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
002_UNESP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
003_UNESP	2	3	0	36	1	0	0	0	0	0	0	0	5	0	0	1	21	377	0	1	0	0	0
004_UNESP	0	3	0	15	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0

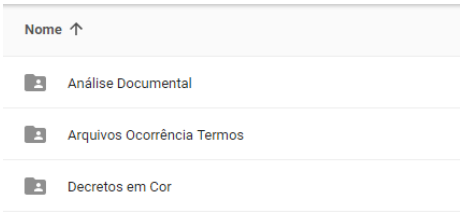
006_UNESP	118	15	0	9	1	0	0	0	0	1	0	1	0	3	0	6	0	0	0	1	0	0	0
007_UNESP	22	4	0	578	7	0	0	1	1	0	0	38	0	3	20	204	57	45	41	1	1	0	1
008_UNESP	0	3	0	29	0	0	0	0	0	0	0	1	0	0	4	13	4	0	2	0	0	0	0
009_UNESP	69	20	2	127	15	0	0	0	0	0	0	0	0	14	0	5	1	6	0	1	0	0	0
010_UNESP	18	6	0	85	25	0	0	0	0	0	0	0	0	5	19	11	18	27	3	1	0	0	0
011_UNESP	3	2	0	90	12	0	0	0	0	0	0	1	0	0	1	0	12	0	3	1	0	0	0
012_UNESP	16	24	1	15	0	0	0	0	0	0	0	0	0	696	4	3	0	5	1	0	0	0	0
014_UNESP	27	2	8	100	110	0	0	0	1	1	0	0	0	626	2	1	0	1	0	1	0	0	0
015_UNESP	7	2	0	75	4	0	0	0	0	0	0	12	0	4	24	44	17	3	3	1	1	0	0
016_UNESP	1	7	0	24	4	0	0	0	0	0	0	0	0	3	6	3	8	2	1	1	0	0	0
017_UNESP	2	1	0	138	0	0	0	0	0	0	0	1	14	0	0	11	0	2	15	0	0	0	0
019_UNESP	9	2	0	187	2	0	0	0	0	2	0	2	0	2	0	12	21	3	16	1	0	0	0
020_UNESP	125	6	0	318	65	0	0	0	3	15	0	2	1	4	5	34	104	39	1	1	1	0	0
021_UNESP	1	1	0	44	7	0	0	0	0	0	0	0	0	0	20	21	109	1	9	1	0	0	0
022_UNESP	7	9	0	79	5	0	0	0	0	0	0	1	0	10	15	13	3	20	0	1	1	0	0
023_UNESP	44	6	2	175	0	0	0	0	0	0	0	1	0	1	0	2	2	0	2	0	0	0	0
026_UNESP	0	8	2	30	1	0	0	0	0	0	0	0	0	12	0	1	0	0	0	0	0	0	0
027_UNESP	106	7	0	36	3	0	0	0	0	0	0	0	0	7	0	0	1	0	1	1	0	0	0
028_UNESP	4	7	0	63	0	0	0	0	0	0	0	0	0	0	2	6	5	1	3	0	0	0	0
031_UNESP	1	1	0	22	0	0	0	0	0	0	0	0	0	0	7	0	0	10	0	0	0	0	0
032_UNESP	3	0	0	291	47	0	0	0	0	2	0	11	0	4	10	31	805	14	58	1	1	0	0
035_UNESP	5	13	1	20	1	0	0	0	0	0	0	0	0	131	5	0	2	1	0	1	0	0	0
001_USP	7	0	1	94	0	0	0	0	0	0	0	0	19	0	1	6	0	19	1	0	0	0	0
002_USP	23	0	0	96	1	0	0	0	0	0	0	4	0	0	0	11	34	4	1	1	0	0	0
003_USP	1	0	0	141	15	0	0	0	0	0	0	0	0	0	1	3	6	353	3	1	0	0	0
004_USP	4	0	0	36	0	0	0	0	0	0	0	0	0	4	6	4	0	0	0	0	0	0	0
005_USP	11	0	0	36	6	0	0	0	0	0	0	6	0	62	0	22	3	2	1	1	0	0	0
006_USP	8	0	0	97	0	0	0	0	0	0	0	0	0	10	0	33	60	165	0	0	0	0	0
007_USP	59	0	0	165	10	0	0	0	0	0	0	0	0	3	1	8	8	8	1	1	0	0	0
008_USP	3	0	0	180	15	0	0	2	1	0	0	0	0	3	2	2	75	0	2	1	0	0	0
009_USP	1	0	0	72	15	0	0	0	0	0	0	0	4	2	2	19	257	88	3	1	0	0	0
010_USP	0	0	0	117	62	0	0	1	0	0	1	3	0	0	1	19	66	295	0	0	0	0	0
011_USP	0	0	0	73	0	0	0	0	0	0	0	0	0	0	3	4	0	1	3	0	0	0	0
012_USP	0	0	0	49	12	0	0	0	0	0	0	0	0	1	3	1	30	4	0	0	0	0	0
013_USP	35	0	0	53	0	0	0	0	0	0	0	3	0	29	1	27	15	1	11	0	0	0	0
014_USP	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	165	0	0	0	0	0
015_USP	0	0	0	6	5	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
016_USP	14	0	0	158	1	0	0	0	0	0	0	0	0	0	0	0	0	8	0	1	0	0	0
017_USP	27	0	0	51	4	0	0	0	0	0	0	0	1	0	0	0	0	3	0	1	0	0	0
018_USP	4	0	0	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Soma	**1648**	**845**	**53**	**8280**	**988**	**0**	**0**	**4**	**11**	**30**	**4**	**168**	**524**	**2011**	**261**	**1057**	**2373**	**5650**	**254**				
Média	**20**	**10**	**1**	**101**	**12**	**0**	**0**	**0**	**0**	**0**	**0**	**2**	**6**	**25**	**3**	**13**	**29**	**69**	**3**				
Mínimo	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**				
Máximo	**311**	**502**	**15**	**788**	**138**	**0**	**0**	**2**	**3**	**15**	**1**	**38**	**176**	**696**	**24**	**204**	**805**	**1546**	**58**				

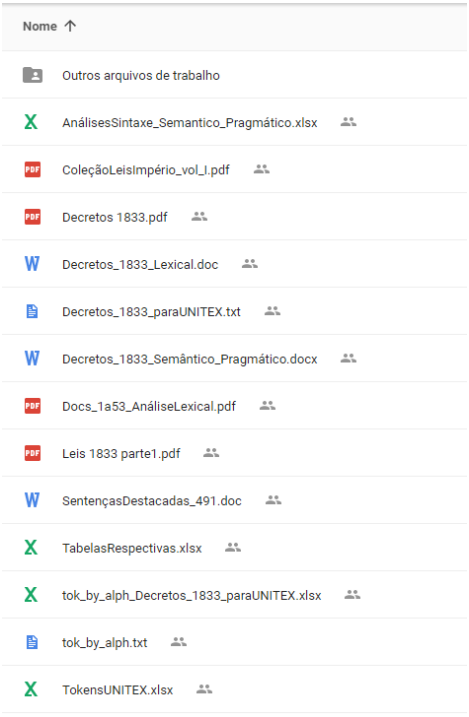
APÊNDICE B – ARQUIVOS NO DISCO ANEXO

O disco em anexo contém arquivos utilizados durante a execução dos métodos de pesquisa, tanto na parte que tratou da análise de ocorrência de termos na seção que tratou da metodologia como também na pesquisa documental. A seguir, seguem orientações para uso dos arquivos.

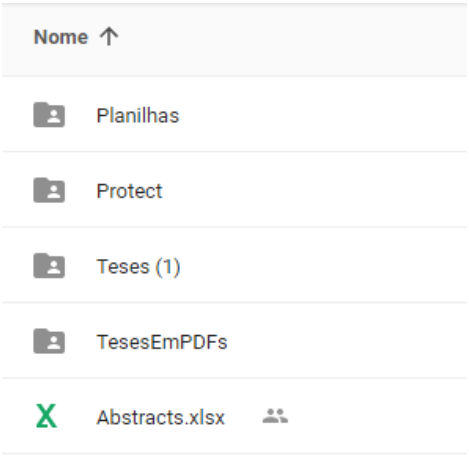
A estrutura de arquivos abaixo mostra a raiz principal do disco. Na pasta “Análise Documental” estão todos os arquivos utilizados na Pesquisa Documental com os documentos do período império brasileiro. Na pasta “Arquivos Ocorrência de Termos” estão os arquivos utilizados na pesquisa de ocorrência de termos nas teses de doutorado de departamentos de ciência da informação no Brasil. Finalmente, a pasta “DecretosEmTIFF” contém os decretos da análise documental numa versão especialmente digitalizada para esta pesquisa, está presente apenas como uma referência extra dos documentos na pesquisa documental.



Já na estrutura de arquivos abaixo estão os arquivos utilizados na pesquisa documental. O arquivo ColeçãoLeisImpério_vol_I.pdf contém os documentos originais. Os demais são tratamentos executados. Notar principalmente o arquivo *tokensUNITEX.xlsx* que corresponde a uma planilha com os resultados totais obtidos.



A estrutura de arquivos abaixo corresponde ao que está presente na pasta da pesquisa de ocorrência de termos. Em TesesEmDocs e TesesEmPDFs estão as teses analisadas. Na pasta planilhas estão todas aquelas que foram utilizados no aplicativo Excell e no aplicativo Ransack. Esse foi utilizado na coleta das ocorrências e co-ocorrências de termos nas teses.



ANEXO A – RELAÇÃO DE TESES CONSIDERADAS

001_UFMG	Ontologias para representação de documentos multimídia: análise e modelagem	2014
002_UFMG	Recuperação da informação através de busca comparada em domínio específico, baseado em expressões miltipalavras	2013
003_UFMG	Modelagem para organização e representação do conhecimento em ontologias de domínio: uma experiência na área da cultura do sorgo	2010
004_UFMG	A linguagem médica utilizada em prontuários e suas representações em sistemas de informação: as ontologias e os modelos de informação	2013
005_UFMG	Análise do padrão brasileiro de metadados de teses e dissertações segundo o modelo entidade-relacionamento	2005
006_UFMG	Ambiente para geração e manutenção semi-automática de tesauros	2005
007_UFMG	Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais	2005
008_UFMG	Processamento de linguagem natural: caracterização da produção científica dos pesquisadores brasileiros	2010
009_UFMG	Sistema de recuperação de informação visual em desenhos técnicos de engenharia e arquitetura: modelo conceitual, esquema de classificação e protótipo	2007
010_UFMG	Migração conceitual entre sistemas de recuperação da informação e ciências cognitivas: uma análise discursiva	2011
011_UFMG	Sirilico: uma proposta para um sistema de recuperação de informação baseado em teorias da linguística computacional e ontologia	2005
012_UFMG	Projeto de sistemas de recuperação de informação corporativa: uma abordagem de análise de domínio baseada na análise facetada	2014
013_UFMG	O uso do sintagma nominal na recuperação de documentos: proposta de um mecanismo automático para classificação temática de textos digitais	2014
014_UFMG	Um modelo baseado em ontologias para representação da memória organizacional	2006
015_UFMG	Encenações linguageiras, jogos argumentativos e redes terminológicas nas eleições presidenciais brasileiras de 2010: a representação da informação em domínios dinâmicos	2014
016_UFMG	Categorização de documentos a partir de suas citações: um método baseado em redes neurais artificiais	2012
017_UFMG	Mineração de textos e gestão do conhecimento: aplicação na experiência operacional em geração de energia nuclear nas usinas de angra I e II	2007
018_UFMG	Uso de sintagmas nominais na classificação automática de documentos eletrônicos	2008
019_UFMG	Análise do domínio organizacional na perspectiva arquivística: potencialidade no uso da metodologia DIRKS	2010
020_UFMG	Gestão arquivística na era do cinema digital: formação de aceervos de documentos digitais provindos da prática cinematográfica	2007
021_UFMG	A preservação de documentos eletrônicos de caráter arquivístico: novos desafios, velhos problemas	2004
022_UFMG	Bases de saber: arqueologia da informação sobre transgênicos	2008
023_UFMG	Fontes de informação de antiquários: proposta de um modelo de análise e de categorização	2006
001_UFRJ	Uma filosofia da ciência da informação: organização dos saberes, linguagem e transgramáticas	2012
002_UFRJ	Filosofia da linguagem e ciência da informação: jogos de linguagem e ação comunicativa no contexto das ações de informação em tecnologias virltuais	2008
001_UNB	Usabilidade da imagem na recuperação da informação no catálogo público de acesso em linha	2004
002_UNB	A representação das necessidades de informação gerencial nos núcleos de informação para avaliação e gestão de empreendimentos	2012
003_UNB	Proposta de modelo de representação do capital intelectual de organizações que desenvolvem software: um estudo no distrito federal	2008
004_UNB	Gestão da imagem organizacional da biblioteca pública na sociedade da informação: as bibliotecas polos do estado do ceará	2013
005_UNB	Uma proposta de interdicisplinaridade entre arquitetura da informação e ciência da computação: linguagem SOWL para as ontologias da web utilizando o formalismo dos grafos conceituais	2013

006_UNB	Processamento de linguagem natural para indexação automática semântico-ontológica	2013
007_UNB	Segurança contra roubo e furto de livros raros: uma perspectiva sob a ótica da economia do crime e da teoria da dissuasão	2014
008_UNB	Em busca dos objetivos bibliográficos: um estudo sobre catálogos	2011
009_UNB	Web semântica e repositórios digitais educacionais na área de saúde: uma modelagem com foco no objetivo de aprendizagem para refinar resultados de busca	2013
010_UNB	Mineração e modelagem de conceitos como praxis de gestão do conhecimento para inteligência competitiva	2010
011_UNB	Autoria de documentos para a Web semântica: um ambiente de produção de conhecimento baseado em ontologias	2006
012_UNB	Identidade/diversidade cultural no ciberespaço: práticas informacionais e de inclusão digital nas comunidades indígenas, o caso dos kariki-xocó e pankararu no braasil	2010
013_UNB	B2: um sistema para indexação e agrupamento de artigos científicos em português brasileiro utilizando computação evolucionária	2013
001_UNESP	Representação iterativa: um modelo para repositórios digitais	2010
002_UNESP	Elementos de interoperabilidade na catalogação descritiva: configurações contemporâneas para a modelagem de ambientes informacionais digitais	2012
003_UNESP	Desenvolvimento e utilização de ontologias em bibliotecas digitais: uma proposta de aplicação	2010
004_UNESP	Representação e persistência para acesso a recursos informacionais digitais gerados dinamicamente em sítios oficiais do governo federal	2013
006_UNESP	A fé documentada: perspectivas metodológicas de organização da informação fotográfica sobre romarias de juazeiro do norte	2014
007_UNESP	Avaliação do uso de linguagem documentária em catálogos coletivos de bibliotecas universitárias: um estudo sociocognitivo com protocolo verbal	2009
008_UNESP	Metadados como elementos do processo de catalogação	2010
009_UNESP	A gênese do arquivo fotográfico de sebastião leme: uma leitura da acumulação	2009
010_UNESP	Institucionalização cognitiva e social da organização e representação do conhecimento na ciência da informação no brasil	2014
011_UNESP	BIAS na representação de assunto: uma discussão de oposições binárias nos functional requirements for subject authority data	2014
012_UNESP	A organização e a representação do conhecimento no domínio da arquivística	2012
014_UNESP	A representação arquivística: uma análise do discurso teórico e institucional a partir dos contextos espanhol, canadense e brasileiro	2014
015_UNESP	O tratamento temático da informação em abordagem sociocultural: diretrizes para definição de política de indexação em bibliotecas universitárias	2014
016_UNESP	Arquitetura da informação pervasiva: contribuições conceituais	2013
017_UNESP	SRDIGITAL: proposta de um modelo baseado na linguagem natural e controlada como instrumentos de apoio ao agente computacional do processo de referência	2012
019_UNESP	Aspectos éticos em representação do conhecimento em temáticas relativas à homossexualidade masculina: uma análise da precisão em linguagens de indexação brasileiras	2010
020_UNESP	O processo de institucionalização sociocognitiva do domínio de organização do conhecimento a partir dos trabalhos científicos dos congressos da ISKO	2014
021_UNESP	A representação documentária do domínio da economia: análise de estruturas de representação em linguagens documentárias e documentos específicos de economia	2014
022_UNESP	Encontrabilidade da informação: contributo para uma conceituação no campo da ciência da informação	2013
023_UNESP	Uma contribuição da teoria literária para a análise de conteúdo de imagens publicitárias do fim do século XIX e primeira metade do século XX, contemplando aspectos da natureza brasileira	2008
026_UNESP	O assunto do e-mail como indício de fraude: contribuições da organização da informação para a prevenção criminal	2008
027_UNESP	O colecionador público documentalista: museu histórico e de ordem geral Plínio Travassos dos Santos	2009
028_UNESP	Acessibilidade em ambientes informacionais digitais	2010
031_UNESP	Modelos conceituais de dados como parte do processo da catalogação: perspectivas de uso dos FRBR no desenvolvimento de catálogos bibliográficos digitais	2010
032_UNESP	A construção de tesauros com a integração de procedimentos terminográficos	2009

035_UNESP	Modelagem conceitual dilam: princípios descritivos de arquivos, bibliotecas e museus para o recurso imagético digital	2015
001_USP	Recuperação de informação: análise sobre a contribuição da ciência da computação para a ciência da informação	2003
002_USP	Da classificação do conhecimento científico aos sistemas de recuperação de informação: enunciação de codificação e enunciação de decodificação da informação documentária	2004
003_USP	O universo e as relações de significação da web: semiose nas ontologias	2012
004_USP	A trajetória da autoria na representação documental	2013
005_USP	Os registros de informação dos sistemas documentários: uma discussão no âmbito da representação descritiva	2009
006_USP	Organização da informação em sistemas eletrônicos abertos de informação científica & tecnológica: análise da plataforma lattes	2007
007_USP	Leitura e significados nos fluxos de informação	2009
008_USP	Informação legislativa ao alcance do cidadão: contribuição dos sistemas de organização do conhecimento	2015
009_USP	Interoperabilidade e mapeamentos entre sistemas de organização do conhecimento na busca e recuperação de informações em saúde: estudo de caso em ortopedia e traumatologia	2015
010_USP	A construção de informações documentárias: aportes da linguística documentária, da terminologia e das ontologias	2010
011_USP	Indexação automática e visualização de informações: um estudo baseado em lógica paraconsistente	2011
012_USP	Capital social e capital científico na produção científica sobre linguagens documentárias e sistemas de organização do conhecimento no campo da knowledge organization nos idiomas espanho, francês e português	2014
013_USP	Análise documentária de fotografias: um referencial de leitura de imagens fotográficas para fins documentários	2002
014_USP	A produção científica brasileira em odontologia e sua visibilidade nacional e internacional	2006
015_USP	Dinâmica das relações entre ciência e tecnologia: estudo bibliométrico e cientométrico de múltiplos indicadores de artigos e patentes em biodiesel	2010
016_USP	Formato: condição para a escrita do jornalismo digital de bases de dados	2011
017_USP	A mulher na sociedade da comunicação ciberdigital	2010
018_USP	Os sentidos pluralistas do cotidiano da cultura nas reportagens da revista realidade nos anos de 1966 a 1968	2006