XVI ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO

Declaração de Direito Autoral

FONTE:

http://www.ufpb.br/evento/lti/ocs/index.php/enancib2015/enancib2015/paper/viewFile/2798/1004. Acesso em: 22 nov. 2015.

REFERÊNCIA:

BODÊ, Ernesto Carlos; SOUSA, Renato Tarciso Barbosa de. Modelo diacrônico para representação e recuperação de informação. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 16., 2015, João Pessoa. **Anais**... João Pessoa: ANCIB, 2015. Disponível em:< http://www.ufpb.br/evento/lti/ocs/index.php/enancib2015/enancib2015/paper/viewFile/2798/1004>. Acesso em: 22 nov. 2015.

XVI ENANCIB 2015
Informação, Memória e Patrimônio: do documento às redes
26 a 30 de outubro - João Pessoa - PB

XVI Encontro Nacional de Pesquisa em Ciência da Informação **(XVI ENANCIB)**
ISSN 2177-3688

GT 2 – Organização e Representação do Conhecimento
Pôster

MODELO DIACRÔNICO PARA REPRESENTAÇÃO E RECUPERAÇÃO DE INFORMAÇÃO[1]

DIACHRONIC MODEL FOR REPRESENTATION AND INFORMATION RETRIEVAL

Ernesto Carlos Bodê, UnB
bod.ernesto@gmail.com

Renato Tarciso Barbosa de Sousa, UnB
renasou@unb.br

Resumo: O artigo aborda a relação entre a Representação da Informação e os efeitos da Mudança Linguística na Recuperação da Informação que ocorrerá no futuro, considerando um lapso temporal entre o processo de Representação e o de Recuperação longa o suficiente para alterar a língua utilizada num acervo documental. Defende-se que esses efeitos podem afetar a qualidade da Recuperação futura, acarretando, assim, problemas para a preservação da Memória. Após a contextualização do tema e do problema, é feita uma análise da situação atual, indicando onde e em que situações o problema ocorre e onde e em que condições poderá ocorrer no futuro. Para compreender exatamente como ocorre o problema citado, é apresentado um Modelo de Representação & Recuperação que considera as condições necessárias para compreender e possivelmente mitigar o problema. Além disso, defende-se o argumento de que o problema em questão, a partir do modelo apresentado, está relacionado à falta de uma postura adequada para o tratamento de documentos que deverão ser preservados por longos prazos ou indefinidamente.

Palavras-chave: Representação da informação. Recuperação da informação. Sistemas de recuperação da informação. Mudança linguística. Memória.

Abstract: The paper deals with the relationship between the Representation of Information and the effects of Linguistic Change in the Information retrieval to happen in the future, considering a time gap between the process of representation long enough to change the language considered in a document collection. It is argued that these effects can affect the quality of future retrieval, thus leading problems for the preservation of memory. After the contextualization of the problem and

[1] O conteúdo textual deste artigo, os nomes e e-mails foram extraídos dos metadados informados e são de total responsabilidade dos autores do trabalho.

analysis of the current situation, indicating where and when the problem occur and where and under what conditions may occur in the future. To understand exactly how is the problem, we present a Representation & Retrieval Model that considers the conditions necessary to understand and possibly mitigate the problem. In addition, we defend the argument that the problem in question, from the model point of view, is related to the lack of proper posture for the treatment of documents to be preserved for long periods or indefinitely.

Keywords: Information representation. Information retrieval. Information retrieval systems. Linguistica change. Memory.

1 INTRODUÇÃO

O tema de minha pesquisa é o problema da *Preservação da Memória* registrada em *Documentos Digitais Arquivísticos* avaliados como possuindo valor histórico cultural que justifique serem mantidos por longos períodos. Em princípio, indefinidamente. De fato, do ponto de vista de um historiador, "o documento foi definido tradicionalmente como um texto escrito à disposição do historiador" (FUNARI, 2003, p. 14).

Meu ponto de partida contrasta com documentos históricos antigos, produzidos a muitas décadas ou séculos atrás em relação a hoje. Nosso foco abarca os documentos produzidos na contemporaneidade e especificamente considerados como possuidores de características que justifique sua preservação indefinidamente.

O problema específico que tratamos abrange os processos de *Representação e Recuperação de Informação* registrada naqueles documentos. O cenário que consideramos é aquele em que há um grande lapso temporal entre o processo original de *Representação*, executado num primeiro momento, e o processo de *Recuperação* que será executado décadas ou séculos no futuro. Consideramos que a língua vernacular das pessoas utilizada no primeiro momento – tanto no conteúdo dos documentos como também nos produtos e instrumentos do processo de representação – está sujeita à *Mudança Linguística*[2]. Se considerarmos também que as pessoas no futuro terão que expressar suas necessidades de informação através de sua língua, pode-se antever vários possíveis entraves à *Recuperação da Informação* nesses documentos e, nesse caso, o que pode ser feito para mitigá-los? Essa é nossa pergunta fundamental.

Um resultado parcial de nossa pesquisa aqui apresentado é a proposta do que chamamos *Modelo Diacrônico*. Este modelo possibilita (I) uma compreensão mais clara da

[2] A Mudança Linguística é um fenômeno que vem sendo observado e estudado principalmente por linguístas de maneira efetivamente científica desde o final do século XIX (FARACO, 2009). Sobre a observação do passado, desde que isso é possível (origem da escrita), podemos concluir que esse é um processo que está em curso e não há motivos para se acreditar que não afetará nossa língua atual.

problemática acima descrita e (II) representa uma mudança de postura em relação a como esse problema vem sendo tratado e percebido (quando é percebido).

2 PERCEPÇÃO ATUAL DO PROBLEMA

Ainda que nossa meta seja tratar o problema de tal maneira a mitigar os efeitos da *Mudança Linguística* no processo futuro de *Recuperação da Informação (RI)*, a melhor maneira de ilustrar concretamente o problema e o modo como ele é tratado hoje é abordar os casos de documentos produzidos no passado.

Abordando especificamente 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 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 acima são emblemáticos em relação aos problemas como ortografias antigas ou léxico utilizados apenas nos séculos anteriores. Mas a língua muda em todos os seus aspectos: fonético, morfológico, sintagmático, semântico e pragmático. 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, do ponto de vista da Ciência da Informação, 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 (GARRETT, 2006).

Os sistemas de RI atuais já lidam há muito tempo com os efeitos do que vem sendo chamado **Problema do Vocabulário**, que é potencializado pela *Mudança Linguística*:

> 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; LANDAUER; GOMEZ; DUMAIS, 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, p. 70-71). 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 ao longo do tempo.

Apenas nos últimos anos surgiram 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).

Não está claro se essas novas propostas para automatizar a identificação e o tratamento dos efeitos da *Mudança Linguística* terão sucesso. Cabe lembrar que a *Indexação Automática* até hoje não é uma realidade plena, pois depende de textos com características bem definidas e padronizadas (LIMA; BOCCATO, 2009, p. 143).

Do ponto de vista da Ciência da Informação, não parece haver modelos para os processos de *Representação e Recuperação da Informação* que contemplem, de maneira explícita, a língua e seu registro escrito em sua relação com a mudança linguística e o papel das pessoas e, principalmente, o **tempo** como fator de potencialização dos efeitos dessa mudança. Pode-se ilustrar isso nos "problemas fundamentais da recuperação da informação" (LANCASTER, 2003, p. 286). Apresentamos uma proposta de modelo que preenche essa lacuna a seguir.

3 O MODELO DIACRÔNICO

O que pretendemos em nossa pesquisa é a elaboração de produtos (modelos e ferramentas) que permitam o monitoramento da *Mudança Linguística*, mitigando, desta forma, seus efeitos em Sistemas de RI que serão utilizados no futuro, visto que, assim como nos casos de acervos antigos digitalizados, a recuperação de documentos contemporâneos também será afetada negativamente no futuro, se nada for feito agora. O caso específico dos documentos arquivísticos, normalmente material textual produzido no âmbito corporativo, é ainda mais grave com relação ao aspecto da língua utilizada, pois é frequente encontrar jargões próprios de cada instituição. Pelo menos uma parte dos significados exatos desses jargões muitas vezes só pode ser determinada no funcionamento real (cultura institucional) de cada instituição.

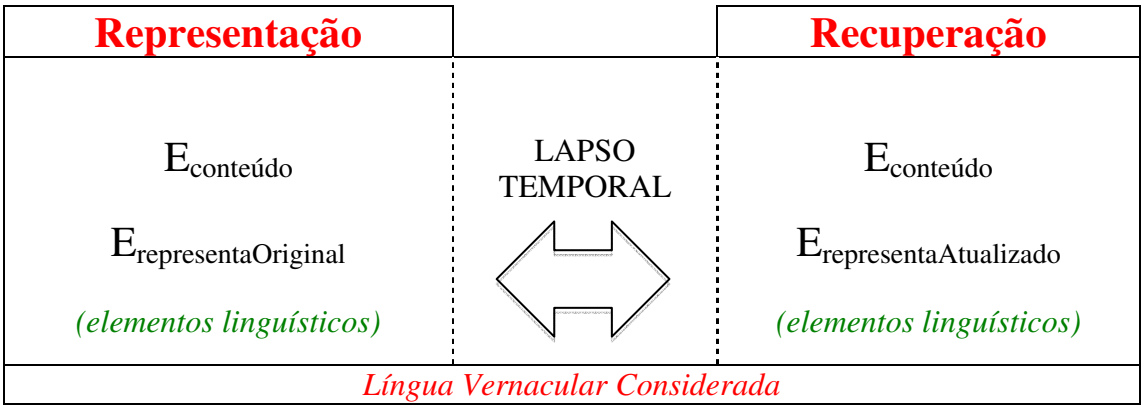
O sucesso do modelo que apresentamos depende de que ele cumpra dois objetivos: (1) visualizar o problema e (2) assumir postura não (re)ativa.

Em (1), queremos dizer que o modelo deve permitir uma clara visualização dos efeitos da Mudança Linguística num acervo documental considerado, estabelecendo as relações entre língua escrita, documentos e pessoas que utilizam os documentos, tanto no processo de representação como no de recuperação futural.

Em (2), defendemos o argumento de que as tentativas de contornar os efeitos da *Mudança Linguística* (como nos exemplos citados de material antigo digitalizado) é uma postura (re)ativa. É assim uma vez que, por meio dela, reage-se contra o problema dando-se conta de que certos documentos precisam ser recuperados porque a língua ali registrada está em um "estado" anterior. Consoante Saussure, "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" (2012, p. 146). Defendemos, portanto, agir no presente para evitar a necessidade dessa reação no futuro. A *figura 1* representa o *Modelo Diacrônico*[3].

Os elementos nesse Modelo são: (1) Processo de Representação, (2) Processo de Recuperação, (3) Econteúdo, (4) ErepresentaOriginal, (5) ErepresentaAtualizado e (6) Língua vernacular considerada.

Figura 1 - Modelo Diacrônico



Representação e *Recuperação* significam dois processos de trabalho (inter)dependentes, já que a tecnologia para RI atual depende de elementos que representem os documentos considerados mas também os processos que acontecem em momentos

[3] O termo diacrônico foi introduzido por Saussure em oposição a sincrônico: "Um fenômeno de linguagem é dito sincrônico, quando todos os elementos e fatores que emprega pertencem a um único e mesmo estado momento de uma única e mesma língua. É diacrônico quando faz intervir elementos e fatores pertencentes a estados de desenvolvimento diferentes de uma mesma língua" (DUCROT; TODOROV, 2010, p. 137).

temporais bem distanciados. O Modelo em questão aplica-se em documentos de longo prazo de guarda, tipicamente documentos arquivísticos com valor histórico-cultural.

Econteúdo é o elemento que reúne todos os elementos linguísticos presentes nos documentos textuais considerados. Em princípio, já que uma língua registrada não muda, esses elementos serão sempre iguais aos originais dos documentos.

ErepresentaOriginal e ErepresentaAtualizado, por outro lado, referem-se aos elementos linguísticos presentes nos produtos ou instrumentos de representação dos documentos considerados em determinado acervo (vocabulário controlado, tesauros, resumos etc.). ErepresentaOriginal é a preparação original desses elementos. Como já demonstramos, esses elementos linguísticos possuem um valor especial por terem sido produzidos próximos aos produtos dos documentos. Já ErepresentaAtualizado engloba os elementos linguísticos correspondentes aos originais, mas em novo estado da língua. Alguns poderão ser iguais aos originais ou podem até não possuir correspondente no léxico atual de uma língua, já que alguns conceitos de coisas podem desaparecer.

4 OBSERVAÇÕES FINAIS

O assunto que ora apresentamos e analisamos envolve várias áreas diferentes. Além da Ciência da Informação e da Ciência da Computação, é necessário dialogar com a Linguística e com a Filosofia da Linguagem. Nos limites do formato desse artigo, não foi possível um aprofundamento nos vários aspectos do *Modelo Diacrônico* apresentado.

Notadamente, é preciso abordar em mais detalhes o que são os elementos linguísticos, sua origem e relação com a Representação & Recuperação da Informação. Por ora, no entanto, é tudo que podemos apresentar.

REFERÊNCIAS

BAIR, Sheila A.; CARLSON, Sharon. **Where keywords fail**: using metadata to facilitate digital humanities scholarship. University Libraries Faculty & Staff Publications. 2008. Disponível em: <http://scholarworks.wmich.edu/cgi/viewcontent.cgi?article=1012&context=library_pubs>. Acesso em: 29 de julho de 2015.

DUCROT, Oswald; TODOROV, Tzvetan. **Dicionário enciclopédico das ciências da linguagem**. São Paulo: Perspectiva, 2010.

FARACO, Carlos Alberto. **Linguística histórica**: uma introdução ao estudo da história das línguas. São Paulo: Parábola Editorial, 2005.

FUNARI, Pedro Paulo Abreu. **Antiguidade clássica: a história e a cultura a partir dos**

documentos. 2. ed. Campinas: Editora da Unicamp, 2003.

FURNAS, G. W.; LANDAUER, T. K.; GOMEZ, L. M.; DUMAIS, S. T. The vocabulary problem in human-system communication. **Communications of ACM**, v. 30, n. 11, november 1987. Disponível em: <http://dl.acm.org/citation.cfm?id=32212>. Acesso em: 29 jul. 2015.

GARRETT, Jeffrey. KWIC and Dirty? human cognition and the claims of full-text searching. **Journal of Electronic Publishing**, 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.

GRETE, Seland. **User revealment revisited**. 2014. Tese (Doutorado)-Institut for Kommunikation, Aalborg Universitet. Disponível em: <http://vbn.aau.dk/files/201270036/PhdThesis_SelandGrete_20140622.pdf>. Acesso em: 29 jul. 2015.

JATOWT, Adam; DUH, Kevin. A framework for analyzing semantic change of words across time. **IEEE**. 2014. Disponível em: <http://dl.acm.org/citation.cfm?id=2740809&dl=ACM&coll=DL&CFID=531915645&CFTOKEN=32321217>. Acesso em: 29 jul. 2015.

LANCASTER, F. W. **Indexação e resumos: teoria e prática**. 2 ed. rev. atual. Brasília: Briquet de Lemos, 2004.

LIMA, Vania Mara Alves; 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**, 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.

MORSY, Sara; KARYPIS, George. **Accounting for language changes over time in document similarity search**. Technical Report. 2015. Disponível em: <https://www.cs.umn.edu/sites/cs.umn.edu/files/tech_reports/15-011.pdf>. Acesso em: 29 jul. 2015.

SOBEL, Karen; BEALL, Jeffrey. Humanities research, book digitization, and the problem of linguistic change. **Library innovation**. v. 2, n. 2, 2011. Disponível em: <http://www.libraryinnovation.org/article/view/99>. Acesso em: 29 jul. 2015.

SAUSSURE, Ferdinand de. **Curso de linguística geral**. 28. ed. São Paulo: Cultrix, 2012.