CONSTRUÇÃO DE CORPORA DE MENSAGENS ELETRÔNICAS PARA CONVERSÃO AUTOMÁTICA EM FALA

Monica Panigassi VICENTINI
Orientadora: Profa. Dra. Jussara Melo Vieira

RESUMO: O presente trabalho pretende apresentar pesquisa desenvolvida com o objetivo de compilar e tratar corpora de mensagens eletrônicas para a conversão desses textos em fala, por meio da tecnologia da Conversão de Texto em Fala (CTF). O corpus de mensagens eletrônicas construído nessa pesquisa permite aprimoramento de ferramentas computacionais, como o Normalizador de Texto, bem como contribui com outros pesquisadores por meio de glossário de abreviaturas, símbolos e códigos dessas mensagens. Ele ainda convida a iniciar discussões a respeito do "internetês", caracterizado por formas distintas na linguagem escrita para comunicação por meio da internet.
Palavras-chave: Linguística de Corpus; Mensagens Eletrônicas; Processamento Lingüístico; Conversão Texto em Fala; Internetês

Introdução

É cada vez maior o uso da internet como meio de comunicação. A forma como as pessoas se comunicam através da escrita pela internet tem se tornado cada vez mais peculiar, a ponto de vermos difundido, atualmente, o termo "internetês", quando há referência à linguagem escrita utilizada na internet. Os textos escritos da internet podem ser veiculados na forma de e-mail, *chat*, *blogs*, etc. Também nos telefones celulares é frequente o uso da escrita para a comunicação através do SMS (*Short Message Service*) – mensagens curtas enviadas por aparelhos celulares, cujo espaço para a escrita é de apenas 160 caracteres. É provável que esse número de caracteres seja um dos motivos para o frequente uso de abreviaturas e outras formas gráficas de escrita para "economizar" caracteres nas mensagens SMS.

Neste estudo analisam-se mensagens eletrônicas do tipo que se veiculam por e-mail e do tipo SMS. Para caracterizá-las, é necessário analisar o maior número possível dessas mensagens. O conjunto em grande volume, de cada tipo dessas mensagens, constitui um corpus. O estudo de um corpus ou de vários (corpora) insere-se na área de Linguística de Corpus, à qual compete a coleta e a exploração de corpora, coletados criteriosamente com a finalidade de análise/pesquisa de uma língua ou variedade linguística, como também, com o intuito de pesquisa e desenvolvimento em linguística e em áreas afins.

Uma dessas áreas para a qual interessa o estudo das mensagens eletrônicas é a área de Tecnologia da Fala e da Linguagem, especialmente pelas contribuições para a Conversão de Texto em Fala (CTF). A CTF refere-se à transformação de um texto escrito em um arquivo de áudio, utilizando-se uma voz virtual (sintetizada). Um sistema de CTF tem diversas aplicações: leitura automática de textos, livros, páginas de internet, etc., para deficientes/não-deficientes visuais, envio de mensagens faladas e personalizadas (avisos de recebimento, cobrança, lembretes, convites, confirmações, etc.), automaticamente, por telefone, para um grande número de pessoas, atendendo, por exemplo, os mercados de *Call Center*, telefonia (Telecom), empresas de VoIP, agências de *marketing*, bancos, governo, etc.

Para que um sistema de CTF tenha um bom desempenho e opere em tempo real, todas as suas etapas têm de ser automatizadas. Para que isso seja possível, de um lado é imprescindível saber sobre a língua que se está vocalizando e, de outro, é necessário construir ferramentas/programas computacionais capazes de processar a língua que se quer vocalizar. Para que um texto escrito seja vocalizado, são várias as etapas de processamento de um sistema de CTF, as quais podemos sumarizar como segue:

i. Processamento linguístico do texto escrito para vocalização;
ii. Processamento fonético-acústico-prosódico do texto processado linguisticamente;
iii. Síntese de fala.

No que tange às etapas de processamento linguístico do texto escrito, podemos destacar/definir:

i. *Tokenização*: é a segmentação da sentença escrita em *tokens*. Um *token* corresponde a um caractere (um sinal de pontuação, por exemplo) ou a um conjunto de caracteres (uma palavra, por exemplo) que tem um significado dentro da sentença. Portanto, um *token* é cada um dos elementos de uma sentença que têm, entre si, um espaço em branco.
ii. Normalização: corresponde à escrita por extenso (também chamada "expansão") de formas gráficas abreviadas ou codificadas. Por exemplo, a forma expandida de "vc" é "você"; de "29/05/2009" é "vinte e nove de maio de dois mil e nove".
iii. Análise/etiquetagem Morfossintática: nessa etapa cada palavra que compõe a sentença escrita recebe uma classificação morfossintática (de acordo com a classe gramatical a que pertence e, também, de acordo com a estrutura sintática da sentença).
iv. Transcrição Fonética: é a transformação dos caracteres escritos ortograficamen te (grafemas) nos símbolos fonéticos correspondentes (fonemas), que serão a re presentação sonora dos grafemas.

O presente estudo se fixa nos três primeiros estágios do processamento linguístico do texto escrito. Ele pode contribuir para o desenvolvimento de sistemas de CTF para o português brasileiro, especialmente ao catalogar e propor soluções para o processamento linguístico de abreviações, símbolos e códigos, frequentemente encontrados nas mensagens eletrônicas.

Objetivos

Constitui o objetivo principal deste estudo: construir e tratar corpora de mensagens eletrônicas (SMS e e-mail) para contribuir nas tecnologias da fala e da linguagem, mais especificamente na conversão de texto em fala. Os objetivos secundários são: gerar/ disponibilizar um glossário específico de mensagens eletrônicas e aprimorar os procedimentos para a etapa de normalização no processamento linguístico da CTF.

Materiais e Métodos

Para cumprimento dos objetivos desse trabalho foram realizadas as seguintes etapas:

Coleta das mensagens eletrônicas

A coleta das mensagens eletrônicas ocorreu das seguintes formas: *website*, e-mail e grupo de discussão online.

Através do *website http://sites.google.com/site/corpusemailesm*s/ realizou-se a divulgação dessa pesquisa e uma das formas de coleta das mensagens eletrônicas. No *website* o colaborador da pesquisa depositava suas mensagens SMS por meio de formulário online e tomava conhecimento do e-mail *corpusemail@ymail.c*om para o qual deveria enviar suas mensagens de e-mail. Ele foi construído de forma a apresentar as informações de maneira clara, direta, em linguagem acessível e com aspecto gráfico agradável, além de facilitar ao máximo o envio das mensagens (de maneira descomplicada e que não tomasse muito tempo do (potencial) colaborador da pesquisa).

Também foi enviado um convite por e-mail para uma rede de contatos formada por amigos, colegas de trabalho, listas de alunos da UNICAMP e de outras faculdades solicitando a participação de voluntários por meio do envio de suas mensagens de e-mail. Solicitou-se um máximo de cinco e-mails de autoria do próprio participante.

Ao longo da coleta das mensagens, obteve-se acesso a um grupo de discussão online. Com a autorização de todos os participantes, coletaram-se 1200 mensagens trocadas entre eles. Como havia uma irregularidade na postagem de mensagens por parte de cada participante, foram utilizadas cinco mensagens de cada um deles, aleatoriamente.

A coleta das mensagens do grupo de discussão serviu como projeto-piloto para testar a aceitação dos potenciais participantes quanto ao "Termo de Consentimento Livre e Esclarecido" e sua consequente colaboração liberando as mensagens. Assim, cada colaborador que se interessasse em enviar suas mensagens deveria aceitar o termo de consentimento.

Tratamento das mensagens eletrônicas

As mensagens eletrônicas coletadas do grupo de discussão estavam em formato "HTML"14 o que significa que junto das mensagens havia códigos e símbolos referentes a essa linguagem de programação que deveriam ser eliminados para que as mensagens fossem analisadas. Para eliminação dos códigos e símbolos (*tags*) em "HTML" foi utilizado um programa computacional escrito na linguagem *Python*.

Todas as mensagens eletrônicas (SMS, e-mail, grupo de discussão) foram armazenadas em arquivos no formato ".txt" (conforme o tipo de mensagem eletrônica), para prosseguirem à etapa de análise. Após esse armazenamento, as mensagens eletrônicas foram reunidas em um único arquivo ".txt" utilizando-se comandos em *Linux*.

Análise das mensagens eletrônicas

As mensagens eletrônicas coletadas foram analisadas através da ferramenta computacional *Unitex*, conjunto de programas que possibilita o tratamento de um corpus linguístico através de recursos linguísticos (Paumier, 2002; *http://www-igm.univ-mlv.fr/~Unitex/*).

Esses recursos são dicionários eletrônicos, gramáticas e tábuas de léxico-gramática, em 14 idiomas, incluindo o Português Brasileiro (Muniz, 2004; Muniz et al., 2005). Os dicionários permitem gerar uma lista das palavras contidas no corpus em análise (em português, p. ex.) e que não estão no dicionário, por estarem escritas de forma errada, serem neologismos ou em português arcaico.

Com o processamento linguístico das mensagens eletrônicas através do *Unitex* obteve-se uma lista de unidades lexicais não reconhecidas no seu dicionário (com suas frequências) e que podem corresponder às novas formas de escrita presentes no internetês.

Essa lista gerada pelo *Unitex* corresponde às abreviaturas, símbolos ou códigos que foram expandidos (escritos por extenso). Após a expansão, as abreviaturas, os símbolos e os códigos expandidos foram novamente submetidos ao *Unitex* a fim de se gerar a classificação morfossintática (categorias gramaticais) de cada unidade linguística expandida.

As mensagens eletrônicas e suas respectivas abreviaturas, símbolos e códigos foram analisados separadamente, e assim a quantidade de mensagens, a frequência das abreviaturas, dos símbolos e dos códigos, bem como suas categorias morfossintáticas foram computados.

Resultados

Nessa coleta obtiveram-se 178 mensagens de e-mail, 151 mensagens SMS e 105 mensagens do grupo de discussão. Para as mensagens de e-mail (incluindo-se o grupo de discussão), foram 19 colaboradores do sexo masculino e 43 colaboradores do sexo feminino. Para as mensagens SMS, foram 9 colaboradores do sexo masculino e 26 colaboradores do sexo feminino.

Após o processamento linguístico das mensagens eletrônicas através do *Unitex*, obteve-se uma lista de 322 palavras desconhecidas para as mensagens eletrônicas do tipo SMS; 817 palavras desconhecidas na lista oriunda dos e-mails e, por fim, das mensagens eletrônicas do grupo de discussão gerou-se uma lista de 272 palavras desconhecidas. Nas mensagens eletrônicas, as "palavras desconhecidas" pelo *Unitex* podem ser abreviaturas, símbolos, palavras com erros ortográficos, nomes próprios etc.

Cada palavra das listas foi analisada e então foram criadas dez categorias para classificação e interpretação das mesmas. As categorias criadas foram: "Abreviações"; "Palavras sem Acentuação", "Palavras Novas", "Estrangeirismos", "Siglas", "Palavras Simples" , "Erros", "*Emoticons*", "Pontuação" e "Nomes Próprios".

Na categoria “Abreviações” convencionou-se colocar as palavras usualmente abreviadas pelos usuários, tanto aquelas já conhecidas (utilização de letra maiúscula mais ponto final para abreviação de nomes próprios), quanto aquelas novas criadas por eles, como “refri” (de refrigerante), “qua” (de quarta), “imuno” (para imunológico), “fisio” (fisioterapia) e “rodo” (rodoviária). Foram encontradas 20 abreviações no total (e-mails, SMS e grupo de discussão).

As “Palavras sem Acentuação” foram aquelas escritas pelos usuários sem a preocupação com acentos ou marcas gráficas diferenciais. Através dos corpora foi possível notar que poucos usuários utilizam o acento gráfico ao escrever as mensagens eletrônicas ou não se preocupam com o mesmo ao escrever. Foram encontradas 140 palavras escritas sem a normatizada acentuação. Exemplos: manha (manhã), JUIZO (juízo).

Categorizaram-se em “Palavras Novas” aquelas palavras que não foram reconhecidas pela pesquisadora ou que são escritas com uma ortografia diferenciada de maneira consciente. Essas palavras foram as mais recorrentes nos três tipos de mensagens eletrônicas (653 ocorrências). É interessante apontar que a terminologia neologismo não foi escolhida devido ao fato de as palavras já existirem. Elas foram diferenciadas porque são escritas, mais especificamente, da mesma forma que são pronunciadas. Alguns exemplos são:

· “Bjinhoooos” ao invés de “beijinhos” (com uma representação enfática utilizando quatro letras “o”, além de representar o som de “bei” somente com a letra “b”); “respondeeeeendo” no lugar de “respondendo” (mostrando também uma ênfase com as 5 letras “e”); “muuuuuuuuiiiiiiittttttttooooooooooo” ao invés de “muito”, (com uma representação enfática através de replicações de várias de suas letras);

· “ateh” no lugar de “até” (com a utilização do “h” para representar o som da vogal acentuada); · “brinkdra” ao invés de “brincadeira” (com a utilização da letra “k” substituindo a sílaba “ca”);

· “axam” no lugar de “acham” (com a letra “x” substituindo “ch”).

A categoria “Estrangeirismos” foi necessária, uma vez que se notou o uso frequente de termos em língua estrangeira nas mensagens eletrônicas. A presença deles pode mostrar a influência de outras línguas, em especial do inglês, em nossa própria língua. Foram encontradas 78 palavras estrangeiras. Exemplos: “kisses”, “hello”, “banner”; “download”, “sorry”; “welcome”.

A categoria “Erros” conta com os erros de digitação dos usuários, como em: “biliografia” (bibliografia); “relacinamento” (relacionamento), “litga” (liga), “trabalhat” (trabalhar) etc. Foram encontrados 70 palavras com erros de digitação.

A lista de “Nomes Próprios” deu-se devido ao fato da ferramenta computacional não reconhecer muitos dos nomes próprios mencionados nas mensagens, além de que os autores das mensagens eletrônicas não costumam diferenciar maiúsculas de minúsculas em nomes próprios. Foram encontradas 73 palavras para essa categoria. Exemplos são: “pucc” (PUCC); “lu” (Lu); “meebo” (Meebo), “bh” (para Belo Horinzonte); “gabi” (Gabi).

Para a categoria “Siglas”, foram encontradas 18 ocorrências, 3 para “Palavras Simples” (convencionou-se “Palavras Simples” aquelas não reconhecidas pelo *Unitex*, mas que reconhecemos como pertencentes ao vocabulário em geral, como “agendinha”), 37 “Emoticons” e 65 diferentes formas de utilização dos sinais de “Pontuação” (frequentemente multiplicados).

Dentro de cada lista selecionaram-se as 5 palavras mais frequentes e as 5 menos frequentes. A abreviação "vc" foi a mais utilizada nas mensagens de texto e nas de e-mail, assim como obteve a segunda posição nas mensagens do grupo de discussão. As abreviações menos frequentes são "ctz" para as mensagens SMS, "qto" para as mensagens de e-mail e "ateh" para as mensagens do grupo de discussão. Os exemplares mais e menos frequentes apresentaram todas as categorias morfossintáticas (gramaticais), à exceção das categorias "artigo" e "numeral".

Reunindo-se as abreviações, símbolos e codificações obtidas a partir de todas as mensagens eletrônicas foi possível gerar um glossário de aproximadamente 1150 termos.

Discussão

O primeiro tema que se coloca como discussão neste estudo é a dificuldade que se encontra ao compilar corpora em geral, uma vez que quando se necessita da colaboração de outrem, como no caso desse projeto que contava única e exclusivamente com a participação de usuários de celulares (SMS) e internet (e-mail), é extremamente necessária uma divulgação maciça para se conquistar o maior número possível de participantes e, por conta disso, a coleta de dados (mensagens, por exemplo) pode tornar-se lenta. Isso se dá uma vez que a internet hoje promove tanta informação que um simples *website* deve fazer uso de muitas estratégias para que se torne de grande conhecimento. É por esse motivo que foi necessário o estudo de algumas estratégias de *marketing* e publicidade, pois foi devido a esse estudo que chegamos em estratégias como registro em sites de busca e em redes sociais diversas. A estratégia de *marketing* e publicidade que proporcionou uma maior abrangência foi a criação do *blog*, o qual podia ser adicionado a vários *websites* que visam a sua divulgação.

Uma segunda questão é que nenhum dos estudos levantados na literatura realizou o procedimento de criação de *website* para coleta de mensagens eletrônicas.

Outro ponto importante é o consentimento dos colaboradores com relação à utilização de suas mensagens. A atitude diante das mensagens coletadas foi de solicitação de autorização dos autores das mesmas, solicitando-lhes o aceite do "Termo de Consentimento Livre e Esclarecido".

Além dessas considerações, verificou-se que as mensagens apresentaram marcas da oralidade, ou seja, eventos/ marcas gráficas que permitem fazer uma relação com os textos falados.

À luz de estudos como Hilgert (2000), pôde-se notar que a oralidade é uma marca forte nas conversações de internet, o que também foi observado nas mensagens eletrônicas analisadas (SMS, e-mail, do grupo de discussão). As marcas de oralidade presentes na escrita das mensagens eletrônicas analisadas puderam ser verificadas pelo uso excessivo de sinais de pontuação como apresentado em "Resultados", os alongamentos vocálicos como em "bjinhoooos", a utilização de pausas como no uso excessivo de reticências.

Na escrita das mensagens eletrônicas, assim como nas conversações de internet estudadas por Hilgert (2000) e Crystal (2005) também encontraram-se abreviações e representações de sons produzidos durante a fala (onomatopeias). Observaram-se reduções gráficas para partículas como "te", "de", "que", "com", "para", "sem", que foram escritas, respectivamente como: "t", "d", "q", "c", "p" ou "p/", "s" ou "s/". Essas reduções podem ser relacionadas àquelas das conversações na internet, cuja opção pela abreviação mostra-se de grande utilização. Como exemplo de onomatopeias, é possível citar a reprodução de risos ou gargalhadas com em: "kkkkk", "hehehehe", "hahahaha", "uahuahuahuahuahau".

Características como o desprezo ao uso de maiúsculas em textos de internet apontadas por Crystal (2005) também são pertinentes. Nas mensagens analisadas, a utilização de maiúsculas apareceu nas mensagens eletrônicas analisadas denotando "gritos" ou muita ênfase em alguma pronúncia.

É importante ainda ressaltar o intenso uso de *emoticons* dos quais os autores se valem para expressar suas emoções ou estados de humor. As mensagens eletrônicas analisadas demonstraram uma relação entre a língua falada e a língua escrita, pois foram encontradas manifestações exclusivas da fala (Hilgert, 2000), o que vem a caracterizar o que está se convencionando chamar de "internetês".

Por fim, não foi possível encontrar um estudo que tenha se ocupado exclusivamente de prover expansões de abreviaturas, símbolos e codificação para o desenvolvimento de tecnologias da fala e da linguagem, notadamente para o português brasileiro.

Para que tecnologias da fala e da linguagem obtenham um bom desempenho com as abreviações, símbolos, codificações, destacando-se a peculiaridade do uso dos sinais de pontuação e dos *emoticons*, será importante que os provedores desse tipo de tecnologia considerem a necessidade de transformarem, o mais fielmente possível, as marcas da oralidade dos autores das mensagens eletrônicas escritas (SMS e email) no momento de sua vocalização.

Conclusão

Cumpriram-se os objetivos dessa pesquisa:

i. Foram construídos e tratados corpora de mensagens eletrônicas (SMS, e-mail e grupo de discussão) visando a conversão de texto em fala;
ii. Foi elaborado um glossário com as abreviações, símbolos e códigos encontrados nas mensagens eletrônicas analisadas;
iii. A análise da especificidade das abreviações, símbolos e códigos encontrados nas mensagens eletrônicas analisadas contribuirá para o aprimoramento da cons trução da ferramenta computacional "normalizador automático de texto" e, por con seguinte, também contribui para o desenvolvimento da tecnologia de Conversão de Texto em Fala (CTF).

Referências Bibliográficas

ALUÍSIO, S. M. ; Oliveira, L. H. M. ; Pinheiro, G. M. (2004). Os tipos de anotações, a codificação, e as interfaces do Projeto Lácio-Web: Quão longe estamos dos padrões internacionais para córpus? In: Anais do II TIL - Workshop em Tecnologia da Informação e da Linguagem Humana, Salvador, Centro de Convenções, 5 a 6 de agosto de 2004. Disponível em *http://www.nilc.icmc.usp.br/lacioweb/publicacoes.htm.*

ALUÍSIO, S. M. ; Almeida, G. M. de B. (2006). O que é e como se constrói um corpus? Lições aprendidas na compilação de vários corpora para pesquisa lingüística. *Calidoscópio*, v. 4, n. 3, p. 155-177, set/dez.

BODOMO, A. (2002). "Linguistics Features of Mobile Phone Communication".

CÂNDIDO Jr., A. (2008). Criação de um ambiente para processamento de córpus de português histórico. São Carlos: O autor.

CARMONA, J. ; Cervell, S. ; Màrquez, L. ; Martí, M. A. ; Padrí, L. ; Placer, R. ; Rodríguez, H. ; Taulé, M. ; Turmo, J. (1998). An Environment for Morphosyntactic Processing of Unrestricted Spanish Text. Proceedings of the 1st International Conference on Language Resources and Evaluation, LREC, pg. 915—922. Granada, Spain. May, 1998.

CRYSTAL, D. (2006) A revolução da linguagem, tradução: Ricardo Quintana, consultoria: Yonne Leite. Rio de Janeiro: Jorge Zahar.

DANTAS, R.A. (2006). Pedidos de ajuda acadêmica em lista de discussão digital: um estudo do gênero. Niterói: O autor.

DUTOIT, T. ; Stylianou, Y. (2003). Text-to-Speech Synthesis. In: Mitkov, R. (Ed.). "The Oxford Handbook of Computational Linguistics". USA: Oxford University Press, p. 323-338.

HILGERT, J. G. (2000). A construção do texto falado por escrito: a conversação na internet. In: Dino Preti. (Org.). "Fala e escrita em questão". 1 ed. São Paulo: Humanitas - FFLCH/USP, 2000, v. 4, p. 17-55.

KOCK e Oesterreicher (1985, 1990, 1994) *apud* Hilgert, J. G. (2000) A construção do texto falado por escrito: a conversação na internet. In: Dino Preti. (Org.). Fala e escrita em questão. 1 ed. São Paulo: Humanitas - FFLCH/USP, v. 4, p. 17-55.

MARCUSCHI. L. A. (1997) *apud* Hilgert, J. G. (2000) A construção do texto falado por escrito: a conversação na internet. In: Dino Preti. (Org.). Fala e escrita em questão. 1 ed. São Paulo: Humanitas - FFLCH/USP, v. 4, p. 17-55.

MARCUSCHI, L. A. (2004) Gêneros textuais emergentes no contexto da tecnologia digital. In: Marcuschi, L. A.; Xavier, A.C. dos S. (Org.) (2004). "Hipertexto e Gêneros Digitais: novas formas de construção de sentido". 1.ed. Rio de Janeiro: Lucerna, v., p. 13-67.

McENERY, T. Corpus Linguistics. In: Mitkov, R. (Ed.) (2003). "The Oxford Handbook of Computational Linguistics". USA: Oxford University Press, p. 448-463.

RICH, L. (2005). The Sociolinguistics of SMS: Na analysis of SMS use by a random sample of norwegians. In: Ling, R. (2005). "Mobile communications*":* Renegotiation of the social sphere. New York: Springer.

SARDINHA, T.B. (2004). Visão geral da Lingüística de Corpus. In: Sardinha, T.B. (2004). "Lingüística de Corpus". Barueri, SP: Manole.

SARDINHA, T.B. (2004). Coleta, armazenamento e pré-processamento de corpora. In: Sardinha, T.B. (2004). "Lingüística de Corpus". Barueri, SP: Manole.

SHEFERD, 2006 *apud* Dantas, R.A. (2006). Pedidos de ajuda acadêmica em lista de discussão digital: um estudo do gênero. Niterói: O autor.

SILVA, F. de S. (2006). Uma abordagem diacrônico-comparativa da abreviação em diferentes gêneros, suportes e tecnologias. Recife: O autor.

SINCLAIR, J. (2005). Corpus and Text – Basic Principles. In*:* Wynne, M. (ed.), "Developing Linguistic Corpora: a Guide to Good Practice". Oxford, Oxbow Books, p. 1-16. Disponível em *http://ahds.ac.uk/linguistic-corpora*/. Acesso em: 22/04/2008.

TEIXEIRA, J. (2003). O q é q é + importt n1 msg? (Mensagens SMS e novos usos da escrita), Diacrítica Série Ciências da Linguagem, nº 17/1. Braga: Universidade do Minho.

TORRUELLA, J.; Llisterri, J. (1999). Diseño de corpus textuales y orales. In: Blecua, J. M.; Clavería, G. ; Sánchez, C. ; Torruella, J. (Eds.). "Filologia e informática. Nuevas tecnologias em los estudios filológicos". Barcelona: Seminário de Filologia e Informática, Departamento de Filologia Española, Universidad Autónoma de Barcelona – Editorial Milenio, p. 45-77. Disponível em *http://liceu.uab.es/~joaquim/publications/Torruella_Llisterri_99.pdf*. Acesso em: dezembro/2008.

TRASK, R. L. (2004). "Dicionário de Linguagem e Lingüística". São Paulo: Contexto. VOCALIZE, 2008-2009. Empresa apoiadora dessa pesquisa. Período de discussões diárias e compartilhamento de conhecimentos.

XAVIER, A.C. (2006). Reflexões em torno da escrita nos novos gêneros digitais da internet. "Investigações". Recife, v. 18, p. 115- 129, 2006.