4
Metodologia

4.1
Origem dos dados

Para constituirmos o corpus de nossa pesquisa, selecionamos dados de fontes diversas, de modo que nossa análise pudesse abranger variadas formas de comunicação na Internet.

Foram coletadas informações de chats, do Orkut e do Twitter. Tal escolha foi feita pelo fato de o primeiro caracterizar-se como exemplo de comunicação síncrona; o segundo como comunicação assíncrona; e o terceiro, além de ser um exemplo de comunicação assíncrona, traz a limitação dos 140 caracteres. Ou seja, tais características podem influenciar os usuários a utilizarem diferentes estratégias de comunicação.

A escolha do Orkut como fonte de dados assíncronos foi feita pelo fato de ser um site em que os usuários utilizam uma linguagem mais informal, com maior possibilidade de inovações. Quando tratamos dos blogs em geral, temos uma grande variedade de assuntos, como mostramos no início deste trabalho, o que faz com que a linguagem utilizada neles varie conforme seu tema.

4.2
Seleção da Amostra

Para a análise nos chats, coletamos dados de cinco salas de bate-papo diferentes, de três sites: Ig Evangélicos, Ig Idade 20-30, Terra Rio de Janeiro, Terra Futebol e Uol Tema livre. Tal coleta foi feita nos meses de março e abril de 2010 e abrange um total de 32647 palavras.

Quanto aos dados do Orkut, contamos com a colaboração de Tadeu Rossato Bisognin, que nos cedeu o corpus utilizado em sua pesquisa de Mestrado, que gerou um livro (Bisognin, 2009). Foram coletados dados dos chamados *scraps*, ou recados, em que são enviadas geralmente mensagens curtas de um usuário a outro

e de *depoimentos* do Orkut, em que um amigo faz uma declaração para o outro. Tais depoimentos, antes de serem exibidos para todos os outros usuários, necessitam de aprovação da pessoa que o recebeu. Ou seja, essa forma de comunicação fica gravada na página pessoal do usuário, o que a afasta das características de texto oral escrito, como ocorre nos chats.

Como o corpus, coletado entre abril e julho de 2007, era bastante extenso (553.875 palavras), fizemos uma seleção de 40.802 palavras, de modo que analisássemos mais detalhadamente suas características e equilibrássemos o tamanho da amostra com a extensão dos dados provenientes dos outros contextos.

Para a análise das mensagens postadas no Twitter, também contamos com a preciosa ajuda de Pedro Asti, Mestre em informática pela PUC-Rio, que nos concedeu o corpus coletado no ano de 2010 para a etiquetagem de dados. Selecionamos 59.805 palavras de um corpus com 560.000, para os mesmos fins descritos anteriormente.

As seleções das amostras do Orkut e do Twitter foram feitas de maneira aleatória, porém seguindo a sequência temporal de postagem.

4.3
Tratamento dos Dados

Após a seleção dos dados, utilizamos o programa Unitex (Paumier, 2011) para a contabilização das palavras, os chamados *tokens*. Segundo definição encontrada na página 2 de seu manual de utilização, ele é...

> ...um conjunto de programas que possibilitam o tratamento de textos em língua natural utilizando recursos lingüísticos. Esses recursos encontram-se sob a forma de dicionários eletrônicos, gramáticas e tábuas de léxico-gramática...

Como tal software contabiliza como token até mesmo símbolos que não são palavras, como $, * etc., fizemos uma "limpeza" manual nos dados, levando em conta somente os tokens que representavam algum tipo de palavra.

Como passo seguinte, selecionamos, também manualmente, do total de palavras válidas, somente aquelas que representavam qualquer tipo de abreviação

ou palavra característica do internetês, como *vc*, *tc*, *q, add*, *alguem (por estar sem acento)* etc. Como há casos em que uma mesma abreviatura representa mais de uma palavra, foi feito um minucioso estudo, em que cada palavra era buscada no contexto em que aparecera no corpus (o que é permitido dentro do próprio Unitex, o qual conta com uma busca textual de corpus), de modo que ficasse claro seu significado. Foi o caso, por exemplo, da abreviatura *cm*, que ora representava a palavra *com*, ora a palavra *como*.

Descartamos todas as palavras que tiveram apenas uma ocorrência, o que poderia caracterizá-la apenas como uma inovação individual.

Foi feita, então, a contagem de ocorrências de tais palavras e uma posterior categorização de cada um delas. Outras estatísticas foram feitas, conforme veremos no capítulo seguinte deste trabalho.

Tais passos foram seguidos separadamente, levando em conta os dados de cada corpus (chats, Orkut e Twitter).

Tendo em mãos tais dados, analisamos ainda como se processavam as abreviaturas de cada sílaba, como veremos a seguir na exposição dos dados.