1
Introdução

Vivemos hoje em um mundo onde as relações pessoais e profissionais são, mais do que nunca, virtuais. O uso de e-mails, *chats*, listas de discussões, Twitter, entre outras ferramentas de comunicação escrita, está em constante expansão, em alguns momentos coocorrendo com a comunicação oral face-a-face e, em outros, substituindo-a.

Em nosso trabalho, vamos nos concentrar na forma de escrita que surge tipicamente nos ambientes de comunicação digital, que, por apresentar características próprias, já foi nomeada como uma nova linguagem presente em todas as línguas: o internetês. Por conta dessa "revolução virtual", muito já se falou sobre o internetês, porém pouco se fez para descrevê-lo e aproveitá-lo como fonte de estudos sobre o português escrito atual.

Este trabalho dá continuidade à nossa pesquisa desenvolvida para dissertação de Mestrado e tem a intenção de aprofundar as investigações iniciadas naquele momento. Embora o internetês apresente características interessantes em todos os níveis de análise da língua, vamos nos concentrar no nível fonético-fonológico e na estrutura da sílaba no português, ou seja, naquilo que se reflete na escrita como ortografia variante. Além disso, apresentaremos as categorias de palavras mais sujeitas à modificação por abreviação, segundo nosso corpus de análise.

Nosso foco de interesse está especialmente nos recursos de abreviação típicos do texto que aparece em ferramentas de comunicação como chats, blogs e Twitter. Mostraremos, por meio de alguns exemplos, como o conhecimento sobre a grafia própria do internetês seria útil para áreas como a da Linguística de Corpus ou da Recuperação de Informação, que vêm deixando de fora de seu escopo milhões de textos por conta da dificuldade de processamento dessa fonte de dados. Embora se note recentemente a preocupação de incluir os novos gêneros digitais na agenda de pesquisa tanto de linguistas quanto de informatas, a descrição do texto digital informal ainda é escassa, especialmente para o português (cf. Park; Byrd, 2001).

Alguns artigos que utilizamos em nossa revisão bibliográfica ressaltam a pouca atenção dada ao reconhecimento das abreviaturas para o processamento automático de linguagem natural. Vejamos, como exemplo, alguns trechos de Park e Byrd (2001):

> Many organizations have a large number of on-line documents – such as manuals, technical reports, transcriptions of customer service calls or telephone conferences, and electronic mail – which contain information of great potential value. (p. 1)[1]

> While we were working on automatic glossary extraction, we noticed that technical documents contain a lot of abbreviated terms, which carry important knowledge about the domains. (p. 1)[2]

> We concluded that the correct recognition of abbreviations and their definitions is very important for understanding the documents and for extracting information from them. (p. 1)[3]

> The problem of abbreviation processing has attracted relatively little attention in NLP field. (p. 7)[4]

> the ability to find correct abbreviations and their definitions is very important to being able to utilize the information contained in those documents. (p. 7)[5]

Para os autores, a tendência de tornar as abreviaturas interessantes, únicas, está crescendo. Isso, sem dúvida, é um fator que dificulta a formalização de regras que as identifiquem. Também por conta dessa criatividade, seria inútil ou demasiadamente custoso criar uma lista fixa de abreviaturas, como comentam Terada, Tokunaga e Tanaka (2002):

> One may think that if there is a fixed list of abbreviations, there is no problem. However, in the real world, different people or even the same person may abbreviate the same word differently, so a fixed list of abbreviations has limited applications (e.g., both “A/C” and “ACFT” stand for “aircraft”). Also the use of such a fixed list of abbreviations requires some effort on the part of the user to look up an abbreviation list and on the part of administrators to maintain such a list.

[1] Muitas organizações têm um grande número de documentos on-line – como manuais, relatórios técnicos, transcrições de chamadas de serviço ao cliente ou conferências por telefone e correio eletrônico – que contêm informações de grande valor potencial.

[2] Enquanto estávamos trabalhando na extração de glossário automático, percebemos que os documentos técnicos contêm uma grande quantidade de abreviaturas, que carregam importantes conhecimentos sobre os domínios.

[3] Concluímos que o reconhecimento correto de abreviações e suas definições é muito importante para a compreensão dos documentos e para a extração de informações a partir deles.

[4] A dificuldade de processamento de abreviaturas tem atraído pouca atenção no campo da NLP.

[5] A capacidade de encontrar abreviações e definições corretas é muito importante para a utilização de informações contidas nos próprios documentos.

(p. 3)[6]

Além do fato de pessoas diferentes abreviarem de formas diferentes a mesma palavra, uma listagem de formas abreviadas teria de ser constantemente atualizada, de forma manual, o que não parece econômico. Pretendemos, então, analisar criteriosamente os padrões de abreviação, de modo que possamos entender quais seus propósitos, separando-as em classes.

A motivação para uma pesquisa sobre as abreviaturas do internetês surgiu pelo fato de percebermos, em nosso meio social e no ambiente de trabalho, uma atitude depreciativa a essa forma de se comunicar, tanto por parte de professores e estudiosos da língua quanto por parte de alunos. Contrariando essa postura, defendemos que essa linguagem apresenta padrões cujo estudo e descrição podem trazer benefícios para diversas áreas de conhecimento, principalmente para a educação. Acreditamos que o internetês não seja um desvio anárquico da língua padrão, como muitos o consideram.

Pudemos perceber também, ao longo de nossa prática docente, que as pessoas, por conta dessa grande expansão das comunicações on-line, parecem estar escrevendo bem mais do que antes, e, por consequência, lendo com mais frequência. Se aceitarmos essa hipótese, a intensa atividade textual (de escrita e leitura) na Internet deve causar algum efeito no processo de aquisição e desenvolvimento do letramento de jovens usuários. Esse efeito, por si só, justifica a necessidade de estudo dos padrões textuais da comunicação *on-line*.

Vale a pena investigar qual seria a influência que textos escritos em internetês podem causar na escrita e na leitura desses jovens. O que não podemos pretender, como professores e como linguistas, é que as pessoas simplesmente não usem esse recurso de comunicação, como se houvesse apenas um padrão de uso da língua, perfeito, ideal, que serviria para qualquer ocasião e como se tivéssemos o poder ou o direito de impô-lo. O que parece importante para nós, educadores, é entendermos esse tipo de linguagem para sabermos lidar com as novidades que a língua, viva que é, traz para nossa sociedade.

[6] Pode-se pensar que, se há uma lista fixa de abreviaturas, não há problema. No entanto, no mundo real, pessoas diferentes (ou a mesma pessoa) podem abreviar a mesma palavra de formas diferentes, por isso uma lista fixa de abreviaturas tem aplicações limitadas (por exemplo, tanto "A/C" quanto "ACFT" representam "aircraft"). Também o uso de tal lista fixa exige algum esforço por parte do usuário a procurar uma lista de abreviaturas e por parte dos administradores para mantê-la atualizada.

Tentaremos, com nossa tese, trazer um olhar novo a estudiosos de diversas áreas, que lidam com a linguagem da Internet, além de professores e alunos que têm uma visão, a nosso ver, equivocada do internetês. Mostraremos que há muito que se estudar sobre esse e outros fenômenos causados pela difusão da comunicação on-line.

No desenvolvimento deste trabalho, utilizaremos diversas fontes. Como não há pesquisa que trate especificamente da descrição do internetês da forma que pretendemos realizar, citaremos aqui alguns trabalhos para situar as pesquisas no campo do tratamento automático das abreviações. Para isso, procuramos descrever trabalhos de diferentes autores que, embora em campos diversos do nosso, podem contribuir para nossa pesquisa.

No que diz respeito ao material bibliográfico referente ao possível tratamento automático das abreviações, encontramos dificuldades devido ao pouco desenvolvimento de estudos desse tipo em língua portuguesa. Por tal motivo, a maioria dos autores que nos servirão como suporte em nossa pesquisa trata da língua inglesa.

Procuramos listar os resultados aos quais chegamos até o momento, destacando a importância de cada autor e de cada trabalho desenvolvido por eles para nossa pesquisa. Como são muitas as fontes nas quais nos baseamos, sua descrição será feita brevemente ao longo deste estudo, de modo a abordarmos somente o que for necessário ao melhor entendimento das hipóteses por nós desenvolvidas.

Este trabalho foi dividido da seguinte forma: primeiramente, no capítulo 2, mostraremos quando e onde habitualmente as abreviaturas são utilizadas, descrevendo antigos e novos gêneros em que ocorre tal fenômeno; logo após, no capítulo 3, destacaremos alguns estudos que demonstram a importância de um maior aprofundamento sobre o tratamento automático de dados advindos da Internet e o desenvolvimento atual de alguns processadores de texto; em seguida, no capítulo 4, será exposta a metodologia que adotamos em nossa pesquisa; após a descrição dos métodos, no capítulo 5, mostraremos os resultados aos quais chegamos após a análise de nossos corpora, descrevendo detalhadamente o internetês segundo nossa concepção e propondo passos a serem seguidos na criação de regras para um possível desenvolvimento de um programa próprio, para a língua portuguesa, que trate os dados em internetês; o capítulo 6 mostrará

as vantagens de se trabalhar em sala de aula com o internetês, explorando os resultados obtidos no capítulo anterior para sustentar nossa opinião de que não há por que ignorar a nova linguagem surgida com a internet. Por fim, faremos as considerações finais.

Vale ressaltar que utilizaremos as expressões *abreviatura* e *abreviação*, cada qual com seu significado. Entendemos *abreviação* como o processo de formação e *abreviatura* como o produto formado. Por exemplo, uma das abreviações da palavra *comigo* resulta na abreviatura *cmg*.