ETIQUETAGEM DE MICROMENSAGENS NO TWITTER: UMA ABORDAGEM LINGUÍSTICA

EVANDRO LANDULFO TEIXEIRA PARADELA CUNHA

ETIQUETAGEM DE MICROMENSAGENS NO TWITTER: UMA ABORDAGEM LINGUÍSTICA

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais - Departamento de Ciência da Computação como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Virgílio Augusto Fernandes Almeida

Belo Horizonte

Junho de 2012

Cunha, Evandro Landulfo Teixeira Paradela

C972e Etiquetagem de micromensagens no Twitter: uma abordagem linguística / Evandro Landulfo Teixeira Paradela Cunha. — Belo Horizonte, 2012
xxii, 66 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais - Departamento de Ciência da Computação
Orientador: Virgílio Augusto Fernandes Almeida

1. Computação - Teses. 2. Redes sociais on-line - Teses. 3. Folksonomia - Teses. I. Orientador. II. Título.

CDU 519.6*04(043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Etiquetagem de micromensagens no twitter: uma abordagem linguística

EVANDRO LANDULFO TEIXEIRA PARADELA CUNHA

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. VIRGÍLIO AUGUSTO FERNANDES ALMEIDA - Orientador
Departamento de Ciência da Computação - UFMG

PROF. CÉSAR NARDELLI CAMBRAIA
Faculdade de Letras - UFMG

PROF. MARCOS ANDRÉ GONÇALVES
Departamento de Ciência da Computação - UFMG

PROF. WAGNER MEIRA JÚNIOR
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 01 de junho de 2012.

Ao meu querido irmão, Rafael,

a quem dedico todas as minhas conquistas,

e

à Vovó Naná e ao Tio Lazito,

duas pessoas queridas que se foram durante a execução deste trabalho

e das quais nunca me esquecerei.

Agradecimentos

Ao fim desta jornada de dois anos, que em muitos momentos mostrou-se árdua e laboriosa, não posso deixar de externar a minha gratidão a quem, direta ou indiretamente, cooperou para a execução deste trabalho.

Em primeiro lugar, agradeço ao professor, orientador e amigo Virgílio Almeida, que me deu a oportunidade de realizar este mestrado. A ele, agradeço pelos ensinamentos, pelos esforços em tornar mais simples a minha adaptação a uma nova área do conhecimento, pela obtenção de recursos para a apresentação de trabalhos no exterior e, sobretudo, pela confiança em mim depositada. Por essas razões, lhe serei eternamente grato.

Agradeço também ao professor e coorientador Marcos André Gonçalves pelas ideias determinantes para a concretização deste trabalho, pela incessante disponibilidade em contribuir durante todo o período de elaboração da dissertação e por toda a atenção que sempre me dispensou. Sua conduta profissional tornou-se um exemplo que pretendo seguir na sequência da minha carreira.

Ao professor César Nardelli, que me apresentou à ciência da mais fascinante e intrigante dentre as faculdades humanas - a linguagem -, agradeço pela orientação sempre segura e pelas valiosas e acertadas sugestões ao longo dos últimos anos. Também agradeço imensamente ao professor Fabrício Benevenuto por ter cedido o dataset que permitiu a realização das análises aqui apresentadas. Devo gratidão ainda ao professor Wagner Meira Jr. por ter aceitado fazer parte da banca examinadora da defesa e pela precisão dos seus comentários e correções.

Deixo ainda registrados os meus sinceros agradecimentos aos muitos amigos que fiz no Centro de Análise e Modelagem de Performance de Sistemas (CAMPS), um verdadeiro celeiro de mentes brilhantes, os quais tornaram mais divertida a lida diária: Emanuel, Geraldo, Giovanni, Las Casas, Marisa, Pesce, Rapha, Rauber, Tat, Tiago e, principalmente, Gabriel, que exerceu um papel fundamental para o enriquecimento deste trabalho.

Sou grato aos demais amigos que, de uma forma ou de outra, contribuíram para

a realização deste mestrado: aos colegas de Fundação Torino, em especial aos *super cool road trippers* Adriano, Artur e Bruno, pelas aventuras; e aos outros amigos do Commando Desportivo Aminas ao Luar, pela distração. Como não poderia deixar de ser, agradeço ainda aos camaradas da esgrima, meus irmãos d'armas, pelos desafios diariamente propostos, e aos Mestres Leiria e França, por compreenderem a razão das minhas faltas aos treinos nos momentos mais difíceis do mestrado.

Mesmo ciente de que palavras não são suficientes para expressar o meu sentimento de gratidão, agradeço de todo o coração às pessoas mais importantes da minha vida: os meus familiares. Aos meus pais, Jorge e Heloísa, pelo amor, carinho e constantes conselhos, que me fizeram ser quem sou hoje. Ao meu irmão, Rafa, exemplo de força e superação que, mesmo em silêncio, sempre me apoiou incondicionalmente em todas as minhas decisões. Aos meus avós, Jacintho e Naná, Nêgo e Nilza, os quais são e sempre serão grandes exemplos na minha vida, estejam eles presentes ou ausentes. Aos meus tios, tias, primos e primas, por todo o apoio e torcida.

Por fim, agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela concessão da bolsa de mestrado e à Universidade Federal de Minas Gerais (UFMG), ao Instituto de Ciências Exatas (ICEx) e ao Departamento de Ciência da Computação (DCC) por terem oferecido as condições ideais para a realização de todas as atividades de pesquisa. Tenho muito orgulho de ter feito parte dessas instituições.

Saibam todos que, sem vocês, esta conquista não teria sido possível.

"Computer science is no more about computers than astronomy is about telescopes, biology is about microscopes or chemistry is about beakers and test tubes. Science is not about tools, it is about how we use them and what we find out when we do."

(Michael Fellows and Ian Parberry, 1993)

Resumo

Hashtags são etiquetas utilizadas pelos membros do Twitter a fim de classificar as micromensagens postadas nessa rede social. Elas são produzidas pelos próprios usuários sem nenhuma interferência da plataforma, o que gera interesse em estudá-las como elementos linguísticos, já que a designação de uma hashtag a uma mensagem apresenta-se como um processo dirigido por condicionadores linguísticos e sociais que interferem na criação de novas etiquetas e na aceitação das tags propostas pelos demais membros.

Nesta dissertação, é apresentado um estudo de inspiração sociolinguística acerca da utilização de hashtags pelos membros do Twitter, partindo do pressuposto de que a rede composta pelos usuários dessa mídia social possui feições comuns a comunidades de fala offline, ou seja, a grupos de pessoas cujos membros se influenciam linguisticamente. Inicialmente, são analisadas as motivações que levam os usuários do Twitter a inserir etiquetas em suas micromensagens. Verificou-se que as principais razões para a etiquetagem dos itens postados são o aumento da compreensibilidade da informação e a ampliação da possibilidade de efetivo compartilhamento do conteúdo. Em seguida, são examinados alguns fatores de ordem linguística que contribuem para o sucesso ou fracasso das tags. Finalmente, é investigado o papel desempenhado por um fator social - o gênero dos usuários - durante o processo de designação de hashtags. As análises realizadas indicam que algumas características presentes nas hashtags são capazes de contribuir para associá-las a usuários do gênero feminino ou masculino. Os resultados obtidos sugerem aspectos semelhantes aos encontrados em estudos do discurso offline, levando a crer que a livre etiquetagem em folksonomias possa servir como modelo para a caracterização da propagação de formas linguísticas em outros contextos.

As conclusões deste estudo complementam o conhecimento sobre o comportamento humano em ambientes de livre etiquetagem e podem ser úteis para o aumento da eficácia de algoritmos de busca em tempo real e de sistemas de recomendação de tags com base nas preferências coletivas dos membros das redes de informação.

Palavras-chave: redes sociais online, etiquetagem de conteúdo, folksonomias.

Abstract

Hashtags are labels used by Twitter members in order to classify messages posted in this social network. They are produced by the users themselves without any interference from the platform, which generates interest in studying them as linguistic elements since the appointment of a hashtag is driven by linguistic and social factors that influence the creation of new tags and the acceptance of labels proposed by other members.

In this work, we present a sociolinguistic-based study about the usage of hashtags on Twitter, assuming that its users' network has common features with offline speech communities, i.e., groups of people whose members linguistically influence each other. Initially, we analyze the motivations that lead Twitter users to insert tags in their tweets. We found that the main reasons for labeling on Twitter are to increase the comprehensibility of the information and to raise the possibility of effective content sharing. Then, we examine some linguistic factors that contribute to success or failure of tags. Finally, we investigate the role of a social factor - the user's gender - in the usage of hashtags. Our results indicate that characteristics of some groups of hashtags are able to contribute to genderize them. The outcomes show similar features to those found in studies of offline speech, that leads us to believe that free tagging in folksonomies can serve as a model for characterizing the propagation of linguistic forms in other contexts.

Our findings complement the knowledge about human behavior in free tagging environments and may be useful to increase the effectiveness of real-time streaming search algorithms and tag recommendation systems based on users' collective preferences.

Keywords: online social networks, content tagging, folksonomies.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Este capítulo tem como objetivo fornecer um panorama acerca da rede social e de informação online Twitter, do recurso de etiquetagem de micromensagens postadas nessa rede e do fenômeno da variação linguística, além de expor o problema a ser abordado e os objetivos do trabalho. Por fim, é apresentada a organização dos demais capítulos desta dissertação.

1.1 Twitter, Hashtags e Variação Linguística

1.1.1 O Twitter

O Twitter (www.twitter.com) é um serviço gratuito de rede social online e de postagem de micromensagens que permite aos seus usuários o envio e o recebimento de textos com até 140 caracteres, conhecidos como tweets[1]. Foi criado em 2006 por Jack Dorsey e em março de 2012 contava com 140 milhões de membros ativos publicando a expressiva média de 340 milhões de tweets por dia [Twitter, 2012].

A menos que os usuários definam os próprios perfis como protegidos (*protected accounts*), as mensagens publicadas por eles no Twitter são visíveis publicamente, inclusive por não-membros da rede. Há também a possibilidade do envio de mensagens privadas entre usuários. Além disso, os membros do Twitter podem assinar (*subscribe*) as contas de outros membros ou de grupos de membros (*lists*) e receber as suas atualizações diretamente na própria página de perfil, formando assim uma rede social de interesses [Rodrigues et al., 2011]. A Figura 1.1 mostra a configuração da página de

[1]Neste trabalho, por questões estéticas e de legibilidade, optou-se por não grafar em itálico alguns termos em língua estrangeira muito recorrentes aqui, como tweet, hashtag, tag, online, dataset, site, entre outros.

perfil de um usuário público, contendo os seus tweets mais recentes.

Figura 1.1. Exemplo de página de perfil de um usuário no Twitter

Assinar a conta de um usuário a fim de receber as suas atualizações é conhecido como seguir (*follow*) aquele usuário, em um procedimento que gera uma relação entre "seguidor" (*follower*) e "seguido" (*followee*). Assim, a estrutura da rede de membros do Twitter pode ser representada por um grafo orientado no qual os vértices descrevem os indivíduos e as arestas direcionadas indicam uma relação assimétrica entre seguidores e seguidos (*following relationship*) - isto é, um membro pode seguir outro sem que seja necessariamente seguido por este. A Figura 1.2 ilustra essa estrutura: o usuário A possui muitos seguidores, o que significa que ele conta com um alto grau de entrada (*indegree*). Por outro lado, ele segue poucos perfis e, consequentemente, possui um baixo grau de saída (*outdegree*). Isso pode sinalizar, por exemplo, que A seja uma celebridade ou uma fonte de informação. Enquanto isso, B, D e E formam uma comunidade em que todos se seguem mutuamente: são, possivelmente, amigos.

As atualizações, caracterizadas pelo envio de novos tweets, podem ser executadas de diversas maneiras, entre elas, em algumas localidades, por meio de mensagens

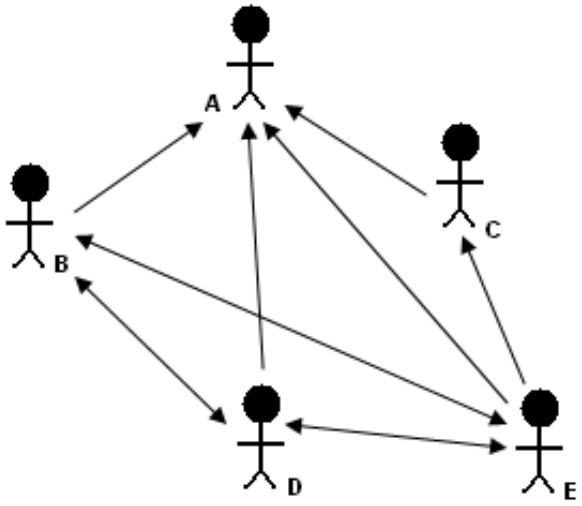

Figura 1.2. Esquema da estrutura da rede de conexões no Twitter

de texto via celular (SMS). Elas são mostradas instantaneamente na própria página inicial do usuário e também no perfil dos seguidores. A praticidade para postar e a instantaneidade da difusão do texto são fatores que tornam o Twitter uma plataforma de fácil acesso e que permite a rápida transmissão de informações. Inclusive, duas das características que têm sido apontadas para explicar a alta popularidade do Twitter no início dos anos 2010 são exatamente a sua simplicidade e a velocidade com que o conteúdo ali inserido se dissemina. Em diversas ocasiões, o Twitter foi utilizado para a propagação de notícias, trabalhando como uma plataforma de jornalismo participativo [Oliveira & Holanda, 2010; Lehmann, 2010], como no caso de desastres naturais em que muitas informações se disseminam nessa rede social antes mesmo que sejam divulgadas pela televisão e pelas demais fontes tradicionais de informação.

A utilização do Twitter é ainda caracterizada por picos durante eventos sociais populares, como competições esportivas, eleições, acontecimentos inesperados e fatos relacionados a celebridades. Um exemplo disso ocorreu no dia 25 de junho de 2009, na ocasião da morte do cantor Michael Jackson, momento em que os servidores da rede social caíram graças à alta quantidade de mensagens postadas em um curto período de tempo a respeito do acontecimento [ICMNews, 2009]. Convém ainda citar a grande utilização dessa rede social por ativistas em diversos protestos, como no caso da Primavera Árabe e do Occupy Wall Street, ambos em 2011, quando a divulgação das atividades via Twitter teve um importante papel nos processos de mobilização, fortalecimento e organização das massas de manifestantes [Huang, 2011; Santo, 2011].

Entretanto, de acordo com Kelly [2009], a maior parte do conteúdo que é postado

no Twitter pode ser descrito como conversação trivial ou *small talk*, isto é, mensagens dotadas de "comunhão fática", segundo conceito introduzido por Malinowski [1923] e definido por Lyons [1970]: mensagens, assim, que possuem um sentido ritualístico, mais do que informacional, e costumam servir apenas como oposição ao silêncio durante a interação. Boyd [2009] acrescenta que esse tipo de conteúdo é natural, visto que a grande maioria dos usuários do Twitter está interessada na rede apenas para manter relações sociais com amigos e conhecidos.

Johnson [2009] descreve da seguinte maneira o funcionamento básico do Twitter e as características das mensagens ali encontradas:

> *Como uma rede social, o Twitter gira em torno do princípio de seguidores. Quando você decide seguir outro usuário do Twitter, os tweets desse usuário aparecem em ordem cronológica inversa na sua página principal do Twitter. Se você seguir vinte pessoas, você verá uma mistura de tweets rolando na página: atualizações sobre o cereal do café da manhã, novos links interessantes, recomendações de músicas e até mesmo reflexões sobre o futuro da educação.* (tradução nossa[2])

1.1.2 As Hashtags

A inclusão de etiquetas textuais no corpo das mensagens é uma maneira utilizada pelos usuários para se categorizar os tweets. Tais etiquetas recebem o nome de "hashtags" e são definidas como todo conteúdo textual imediatamente precedido pelo símbolo cerquilha (#), conhecido em inglês como *hash sign*. Basicamente, as hashtags são cadeias de caracteres (apenas letras, números e traços inferiores/*underscores*) criadas livremente pelos membros da rede a fim de adicionar contexto e metadados às postagens, funcionando muitas vezes como palavras-chave dos tweets.

As hashtags, no entanto, não surgiram no Twitter. Messina [2007] informa que a utilização do sinal # como introdução a um metadado foi estabelecida nos anos 90 para categorizar canais de IRC (*Internet Relay Chat*). Em algumas linguagens de programação, especialmente em Perl, Python e Ruby, a utilização de # como indicador de comentário é contemporânea ou ainda anterior.

No Twitter, uma hashtag foi utilizada pela primeira vez em 2007, por Chris Messina. Segundo o próprio criador, havia um desejo entre os usuários do Twitter para

[2]Original: *As a social network, Twitter revolves around the principle of followers. When you choose to follow another Twitter user, that user's tweets appear in reverse chronological order on your main Twitter page. If you follow twenty people, you'll see a mix of tweets scrolling down the page: breakfast-cereal updates, interesting new links, music recommendations, even musings on the future of education.*

que existisse alguma maneira de grupos interessados nos mesmos tópicos se organizarem nessa rede social. Então, Messina idealizou a utilização do sinal # para identificar palavras-chave dos tweets e facilitar a busca diretamente por elas, aumentando assim a precisão das consultas [Messina, 2007]. Apenas dois anos depois, porém, o Twitter passou a inserir hiperlinks diretamente nas etiquetas, de maneira que um clique sobre uma hashtag tornou-se suficiente para efetuar uma busca pelos tweets mais recentes que a contivessem.

Um exemplo de tweet incluindo uma hashtag é:

O TME, principal competição em MG, será realizado entre 20 e 22/10!
#esgrima

A inclusão de uma hashtag nesse tweet sugere que o autor esteja conectando o conteúdo da mensagem a uma palavra-chave específica, a qual, além de complementar a informação contida no texto e de aumentar a sua compreensibilidade graças à adição de um metadado, permite o fácil acesso ao tweet por outras pessoas interessadas no mesmo tópico.

A Figura 1.3 mostra um exemplo anonimizado de busca pela hashtag #esgrima, a qual retorna os tweets mais recentes que contêm essa etiqueta. Na consulta, o Twitter não diferencia caracteres maiúsculos de minúsculos e tampouco inclui mensagens cujo termo buscado apareça unicamente sem a cerquilha.

Além de fornecerem metadados aos tweets, pode-se observar que hashtags têm sido frequentemente utilizadas com outros objetivos - por exemplo, como agregadoras de mensagens para a organização de fóruns de discussão não moderados no Twitter; ou ainda como forma de promoção de marcas e publicidade de produtos - como ilustra a Figura 1.4 -, campanhas, eventos e personagens; ou até mesmo com fins estritamente lúdicos, como jogos e brincadeiras. Há ainda o fenômeno dos "memes de Internet" (*Internet memes*), que está intimamente ligado às hashtags na medida em que a popularização de muitos deles é alimentada pela propagação das tags a eles associadas.

Graças às hashtags, o Twitter apresenta-se como um "ambiente de livre etiquetagem" (*free-tagging environment*), o que significa que a atribuição de tags aos itens não passa por controle do sistema, sendo de responsabilidade exclusiva dos usuários. O fato de que praticamente quaisquer cadeias de caracteres possam ser transformadas pelos próprios membros da rede em hashtags, sem nenhuma intervenção da plataforma, e, a partir daí, possam se disseminar pela rede, gera interesse no estudo das dinâmicas de criação, uso e propagação dessas etiquetas. Esse interesse se relaciona com a necessidade de compreensão do "comportamento humano de etiquetagem" (*human tagging*

Figura 1.3. Exemplo de busca pela hashtag #esgrima no Twitter

behavior), para que sejam oferecidos serviços melhores e mais adequados às exigências dos usuários de mídias sociais online.

1.1.3 A Variação Linguística

Como as hashtags, em grande parte, são criadas individualmente e isoladamente pelos usuários, um novo acontecimento social pode levar ao surgimento simultâneo de várias tags diferentes, que são ou não aceitas pelos demais membros da rede - isto é, seus seguidores. Dessa forma, algumas se propagam e obtêm sucesso, enquanto outras morrem imediatamente após o nascimento e ficam restritas a poucos tweets.

De maneira análoga, uma inovação lexical se dá quando uma nova forma é adicionada ao léxico de uma língua, seja por meio de: a) criação de novos termos (neologismos); b) reutilização de termos já existentes; c) importação de termos de outras bases

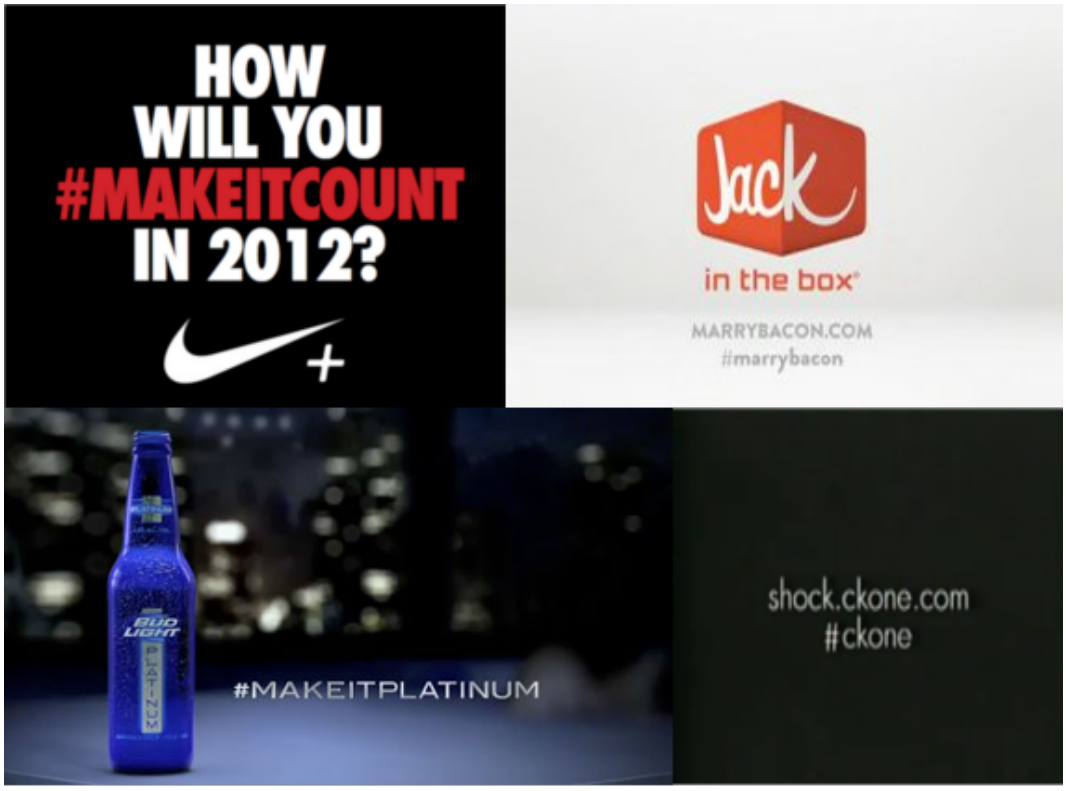

Figura 1.4. Exemplos de campanhas de marketing que utilizaram hashtags como maneira de incentivar a propagação de conteúdo sobre os produtos na Web

lexicais (estrangeirismos)[Rus, 2008]. A inovação nasce a partir do falante, que a propõe aos demais membros de sua comunidade de fala - ou seja, a quem está conectado a ele na sua rede de relacionamentos e contatos linguísticos -, os quais realizam uma seleção cultural dessa inovação, aceitando-a ou rejeitando-a, como mostram os grafos na Figura 1.5: o primeiro indica o momento inicial do processo de variação e mudança linguística; o segundo, um momento posterior, em que alguns membros da comunidade utilizam determinada forma inovadora, mesmo que não exclusivamente, enquanto outros, embora possivelmente a conheçam, não a utilizam (os vértices brancos indicam indivíduos que aderiram à inovação em um determinado instante; os pretos, aqueles que continuam utilizando a forma não inovadora). Segundo Easley & Kleinberg [2010], esse processo é similar àquele desencadeado em diversas situações nas quais ocorre a propagação de algum elemento inovador.

De acordo com a Teoria da Variação e Mudança Linguística, proposta por William Labov e outros linguistas a partir dos anos 60 [Weinreich et al., 1968; Labov, 1995, 2001] assim se propaga uma nova forma linguística: havendo uma forma inovadora - "uma variação, portanto -, e esta consiga algum prestígio, qualquer que seja a razão, pode ser

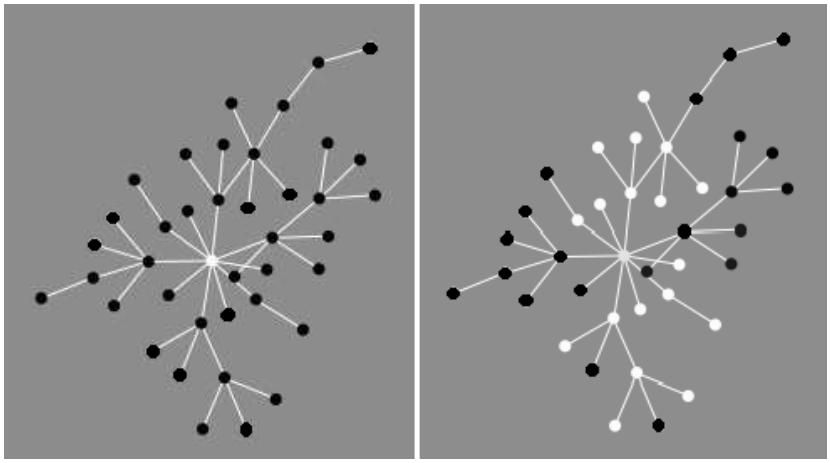

Figura 1.5. Dois momentos distintos no processo de propagação de uma inovação linguística (adaptado de Troutman et al. [2008])

que o falante comece a usá-la, adotando-a em certas circunstâncias" [Silva, 2006]. Nessa teoria, o termo "variação" é entendido como a utilização simultânea, pela sociedade ou pelo próprio indivíduo, de duas ou mais formas concorrentes ou "variantes" - isto é, formas indicando um mesmo valor semântico ou, segundo alguns autores, uma mesma função comunicativa. Já o termo "mudança" está relacionado à adoção categórica de uma forma em detrimento de outra. Assim, a mudança é sempre precedida por um período de variação linguística, embora nem toda variação gere necessariamente uma mudança.

Em princípio, já que essas formas variantes possuem o mesmo significado, elas poderiam ser utilizadas igualmente, da mesma maneira e nas mesmas situações. Entretanto, não o são, e não por acidente: há uma série de condicionadores sociais e linguísticos que regulam a escolha dos falantes em favor de uma ou de outra forma variante. Um dos objetivos da pesquisa sociolinguística é a identificação desses condicionadores, conhecidos como "fatores externos" (quando de natureza social) e "fatores internos" (quando de natureza estritamente linguística). Em outras palavras, busca-se conhecer quais são os fatores que direcionam os falantes a uma ou outra forma variante a fim de auxiliar a esclarecer a relação entre língua e sociedade.

1.2 Formulação do Problema

A partir da década de 2000, com o advento das redes sociais online e de outras plataformas interativas que compõem o que tem sido chamado de Web 2.0 [Oreilly, 2007], a participação do usuário no processo de geração e de disseminação de conteúdo na Web vem crescendo sistematicamente. Sites que promovem conexão e interação entre

membros, como Facebook, Youtube, Twitter e LinkedIn, estão entre os mais acessados no início dos anos 2010.

Um dos fenômenos surgidos nesse contexto é a livre etiquetagem de conteúdo digital. Trata-se de recurso oferecido aos usuários de serviços Web e que consiste em permitir que os próprios membros das redes categorizem o conteúdo que disponibilizam e compartilham nessas redes. Dessa maneira, passa a inexistir o controle sobre a indexação dos itens, que se torna aberta a toda a comunidade.

As etiquetas textuais designadas pelos usuários aos seus itens de conteúdo refletem, assim, características pessoais e coletivas desses indivíduos. Afinal, a escolha de cada etiqueta é resultado de processos mentais geradores de formas linguísticas, os quais se expressam nas próprias tags. Para que esses processos sejam conhecidos, tornam-se úteis a identificação e a caracterização dos fatores que influenciam as decisões de designação de etiquetas a itens de conteúdo compartilháveis nas mídias sociais. No caso específico do Twitter, em que a etiquetagem é um processo facultativo, é importante também conhecer os elementos que levam os usuários a inserir tags nas mensagens.

Duas tarefas, portanto, mostram-se relevantes: a descrição das motivações que fazem com que os membros das redes de informação etiquetem as suas postagens e a identificação de fatores que tornam as etiquetas mais ou menos produtivas em toda a comunidade ou em determinados grupos sociais.

1.3 Objetivos

Os objetivos gerais deste trabalho são identificar motivações para a etiquetagem no Twitter e descrever fatores linguísticos e sociais que influenciam as decisões dos usuários no momento em que designam hashtags às suas micromensagens.

Os objetivos específicos são:

- formular, aplicar e analisar questionários para a identificação das razões que levam os usuários do Twitter a inserir hashtags nas postagens;
- construir bases de dados compostas por tweets, hashtags e informações de membros do Twitter;
- caracterizar quantitativamente e qualitativamente os conjuntos de dados coletados;
- definir, com base na literatura, fatores linguísticos que possam ter relação com a alta ou com a baixa utilização de etiquetas e verificar essa relação;

- definir, com base na literatura, aspectos das hashtags que possam associá-las a usuários dos gêneros feminino ou masculino e verificar essas associações.

1.4 Organização da Dissertação

Esta dissertação é organizada da seguinte maneira: o Capítulo 2 apresenta trabalhos relacionados, que vão desde estudos de caracterização do Twitter até trabalhos sobre o recurso da livre etiquetagem de conteúdo digital em redes sociais online, além de publicações da área de linguística que dão suporte às hipóteses aqui levantadas; o Capítulo 3 discute o processo de etiquetagem textual, o conceito de folksonomia e apresenta pesquisa acerca das motivações encontradas pelos usuários do Twitter para etiquetar suas mensagens; o Capítulo 4 expõe a constituição dos conjuntos de dados utilizados nas análises experimentais; o Capítulo 5 apresenta os resultados dos estudos realizados; e, por fim, o Capítulo 6 conclui a dissertação, propondo caminhos para investigações futuras.

Capítulo 2

Trabalhos Relacionados

Neste capítulo, são apresentados e discutidos alguns estudos relacionados ao Twitter, à utilização de tags nessa e em outras redes sociais online e aos fundamentos linguísticos que serão abordados nas análises experimentais.

2.1 Twitter

Muito tem sido publicado, com diferentes abordagens e à luz de variados referenciais teóricos, acerca de redes sociais online e, em específico, do Twitter - seja nos campos tradicionais da Ciência da Computação e da Informação, seja em áreas como Antropologia, Sociologia, Linguística e Psicologia. Muitos desses estudos possuem características multi ou interdisciplinares, o que enriquece o debate sobre os temas tratados e oferece a possibilidade de se trabalhar as questões levantadas por meio de diferentes perspectivas. Afinal, a Web reflete os interesses e os valores das sociedades que a utilizam [Berners-Lee et al., 2006], funcionando como um espelho para o qual cientistas de diferentes áreas podem mirar a fim de analisar as comunidades que agem nesse espaço de informação [Sawyer & Rosenbaum, 2000]. Além disso, a compreensão dos padrões de comportamento dos indivíduos na Web pode ser útil para que lhes sejam oferecidos serviços mais personalizados de acordo com suas características, preferências e necessidades. É o caso de diversos estudos nos campos da Computação Social e da Sociologia da Web que abordam o comportamento dos usuários de redes sociais online e que serão mencionados nesta seção.

Algumas das primeiras caracterizações da utilização do Twitter e das propriedades topológicas e geográficas da sua rede de membros foram realizadas por Java et al. [2007] e Krishnamurthy et al. [2008], que identificaram ainda os interesses e as motivações dos usuários presentes na fase embrionária do Twitter, a exemplo de Zhao & Rosson

[2009]. Kwak et al. [2010] também estudaram de forma quantitativa as características topológicas do Twitter, além de terem investigado a difusão de informação na sua rede e seu poder como um novo meio de distribuição de conhecimento, tendo sido o primeiro trabalho a estudar o Twitter com um todo. Suas análises são, em certos momentos, similares a algumas daquelas realizadas aqui.

Benevenuto [2010] apresentou um amplo estudo sobre interações em mídias sociais, inclusive no Twitter, e cobriu aspectos do comportamento e da navegação dos usuários. Nesse estudo, foram revelados padrões de comportamento típico dos membros de redes sociais online e foram identificadas formas de conteúdo não solicitado (*spam*). Cha et al. [2010], por sua vez, estudaram o conceito sociológico de influência e o aplicaram para medir a influência online dos participantes do Twitter, concluindo que ter muitos seguidores nessa rede social não significa necessariamente ser influente sobre as ações desses membros. Bigonha et al. [2010] investigaram a polaridade das mensagens e dos usuários em relação a determinados tópicos, sendo capazes de determinar, em grande escala, apoiadores e opositores de certos conceitos. Comarela et al. [2012] analisaram a dinâmica humana no Twitter e se debruçaram sobre a tarefa de mensuração da importância relativa das mensagens postadas pelos usuários, identificando fatores que influenciam a taxa de resposta e a probabilidade de compartilhamento do conteúdo, além de terem oferecido um modelo eficaz para o ranqueamento de tweets baseado em relevância. Benevenuto et al. [2010] enfrentaram o problema da detecção de *spammers* nessa rede social, sugerindo uma estratégia que mostrou-se capaz de detectar grande parte do conteúdo não desejado com apenas um pequeno percentual de não-*spams* mal classificados. Rodrigues et al. [2011] apresentaram a questão da descoberta de conteúdo pelo processo conhecido como *word-of-mouth* e analisaram a propagação de URLs no Twitter em função da distância geográfica entre os usuários. Demonstrou-se que as árvores de propagação no Twitter são mais largas do que profundas e que usuários geograficamente próximos uns dos outros possuem maior probabilidade de compartilhar URLs em comum.

O'Connor et al. [2010] estudaram o Twitter como uma plataforma para aferir sentimentos das comunidades que a utilizam, enquanto Golder & Macy [2011] identificaram variações de humor entre diferentes culturas por meio da análise de sentimento de mensagens postadas no Twitter. Chew & Eysenbach [2010] conduziram um estudo que investiga a disseminação de tweets com as palavras-chave "swine flu" e "H1N1" durante a pandemia de gripe suína em 2009. Os objetivos desse trabalho foram monitorar o uso desses termos ao longo do tempo para analisar o conteúdo das mensagens e validar o Twitter como uma ferramenta de acompanhamento de eventos em tempo real. Gomide [2012] propôs ainda uma metodologia capaz de utilizar o conteúdo com-

partilhado no Twitter para a detecção e a previsão da ocorrência de eventos do mundo offline, como epidemias e desastres naturais. Gupta & Kumaraguru [2012] estudaram a credibilidade dos tweets postados acerca desses eventos e aplicaram estratégias para ranqueá-los de acordo com a sua credibilidade.

Verifica-se, portanto, que o rol de estudos envolvendo o Twitter é vasto e a natureza das pesquisas abrange várias áreas do conhecimento.

2.2 Livre Etiquetagem de Conteúdo em Redes Sociais Online

A livre etiquetagem de itens de conteúdo na Web também tem sido tema de diversas pesquisas. Brandt [2009] apresentou um amplo trabalho em que estudou etiquetagem e folksonomia sob a ótica dos processos de organização e recuperação de informação na Web, concluindo que a livre etiquetagem de conteúdo pode ser considerada um modelo válido para a classificação dos itens, dependendo da natureza da plataforma e dos objetos informacionais com os quais as etiquetas se relacionam. Gao et al. [2009] analisaram como a visualização da relevância das etiquetas aumenta a consistência da etiquetagem e reduz a exigência física decorrente da designação de tags, o que traz implicações no desenvolvimento e aperfeiçoamento de sistemas. Mejias [2004] investigou as práticas sociais e comunicativas surgidas em ambientes de livre etiquetagem, concluindo que usuários passam por dificuldades para se habituar a esquemas de classificação distribuída e que apenas a prática parece parece esclarecer a esses indivíduos os benefícios dessa nova forma de indexação de conteúdo. Mathes [2004] explorou os metadados criados pelos usuários, focando nas mídias sociais online Delicious e Flickr. Nov et al. [2008] apresentaram o comportamento de etiquetagem no Flickr, indicando que a motivação para o uso de tags nessa rede social está relacionada com o público alvo das postagens. Ivanov et al. [2012] pesquisaram o combate a ruídos e *spam* em ambientes de livre etiquetagem, comparando as abordagens já existentes para a detecção de tais conteúdos. Mistry & Sen [2012] propuseram um sistema de recomendação de tags que classifica as etiquetas baseando-se em similaridades semânticas, enquanto Eleta & Golbeck [2012] compararam os padrões de etiquetagem, em línguas diferentes, de coleções de imagens artísticas. Estes concluíram que perspectivas culturais diferentes podem ser encontradas por meio da análise das tags menos frequentes. Iofciu et al. [2011] investigaram se os usuários de redes sociais online podem ser identificados após a análise das suas práticas de etiquetagem, chegando à conclusão de que, mesmo com certas limitações, é possível realizar tal tarefa.

Com relação à utilização de hashtags especificamente no Twitter, convém destacar o trabalho de Romero et al. [2011], que estudaram o fenômeno da propagação de hashtags no Twitter com foco na variação dos padrões de difusão em função do tópico da mensagem. Esse artigo introduz os conceitos de *stickiness* e *persistence* de etiquetas. Tsur & Rappoport [2012] apresentaram uma abordagem que combina aspectos de conteúdo e características topológicas da rede para prever a propagação de hashtags. Carter et al. [2011] trabalharam com a questão da tradução de hashtags sobre o mesmo tópico. Poschko [2010] explorou as relações entre hashtags considerando as co-ocorrências e apresentou uma tentativa de classificá-las em diferentes classes utilizando uma abordagem baseada em aprendizagem de máquina. Davidov et al. [2010] propuseram uma análise de sentimento de tweets com base nas hashtags designadas e na utilização de *smileys* no corpo das mensagens. Bruns & Burgess [2011] analisaram a utilização de hashtags no debate político. Weng et al. [2010] estudaram quão interessantes são, aos olhos dos usuários, certas hashtags, para em seguida introduzir um método de ranqueamento de etiquetas de acordo com o interesse que elas são capazes de gerar na comunidade. Hong et al. [2011] investigaram as diferenças na utilização de diversos elementos no Twitter, inclusive hashtags, entre membros que publicam em línguas diferentes. Papacharissi & Oliveira [2011] realizaram um estudo baseado em análise do discurso a fim de traçar a história da hashtag #egypt durante os levantes populares de 2011. Wagner & Strohmaier [2010] aplicaram o modelo de *tweetonomies*, definido por eles mesmos e que, segundo os autores, vai além do conceito de folksonomias, para gerar redes semânticas de hashtags.

Porém, em nenhum desses trabalhos abordou-se a questão dos fatores, especialmente os sociais, que influenciam a decisão dos usuários do Twitter em favor de uma ou outra hashtag variante sobre o mesmo tópico, como é realizado neste estudo.

2.3 Variação e Mudança Linguística

A Teoria da Variação e Mudança Linguística, proposta por Weinreich et al. [1968] e Labov [1972, 1995, 2001], considera que a variação não é aleatória, mas regulada por fatores de natureza tanto linguística quanto social. Essa variação é constitutiva da linguagem humana e se manifesta como uma heterogeneidade ordenada, ou seja, seu comportamento é controlado por um conjunto de fatores de ordens diversas [Cambraia et al., 2008]. O objetivo dos estudos que seguem essa teoria, portanto, muitas vezes passa pela busca da descrição das variáveis intra e extralinguísticas que determinam a previsibilidade do fenômeno da variação [Gonçalves, 1993].

Nesta dissertação, é verificada a influência de alguns fatores linguísticos no processo de designação de hashtags por parte dos usuários. Um desses fatores é o comprimento das etiquetas. Zipf [1935] sugeriu que o comprimento de uma palavra tende a manter uma relação inversa, porém não necessariamente proporcional, com a sua frequência relativa. Sigurd et al. [2004] analisaram dados de diferentes gêneros textuais em inglês e sueco e corroboraram a hipótese, demonstrando que as palavras mais longas tendem a ser evitadas provavelmente por irem contra o princípio da economia linguística [Vicentini, 2003].

Com relação aos fatores sociais, neste trabalho é abordada a influência do gênero dos usuários na utilização das tags. Diversos estudos já indicaram que o gênero possui um papel importante no processo de variação linguística, pois homens e mulheres utilizam a língua de maneira diferente, de acordo com os padrões de comportamento associados às suas posições nas comunidades nas quais estão inseridos.

O primeiro estudo que correlacionou gênero à variação linguística examinou a pronúncia do *-ing* final no inglês falado em Boston [Fischer, 1958]. Verificou-se uma diferença significativa entre a pronúncia de falantes dos gêneros masculino e feminino: a variante padrão foi mais frequente entre mulheres do que entre os homens, que usaram principalmente a variante não-padrão. Esses resultados foram confirmados por estudos da mesma variante linguística em comunidades britânicas e australianas, com resultados semelhantes [Trudgill, 1974; Horvath, 1985].

A partir daí, muitos outros estudos [Laberge, 1977; Guy, 1981; Tannen, 1990; Cheshire, 2001; Macaulay, 1977] também indicaram diferentes padrões entre as maneiras como homens e mulheres falam, organizam o discurso e interagem usando a língua, inclusive em ambientes online [Soares & Peixoto, 2010]. Em geral, eles mostram que os falantes do gênero feminino são mais propensos a utilizar variantes não-estigmatizadas, ou até mesmo variantes de prestígio, do que os falantes do gênero masculino. Esse padrão foi identificado em uma série de línguas modernas ocidentais.

Outros trabalhos [Modaressi, 1978; Abdel-Jawad, 1987; Bakir, 1986; Haeri, 1987], porém, indicaram que esse padrão é diferente em comunidades islâmicas, onde as variantes de prestígio geralmente são predominantes entre os homens e não entre as mulheres. Resultados similares foram encontrados em comunidades hindus [Jain, 1973; Gambhir, 1981]. Por outro lado, foi demonstrado que em japonês [Hibiya, 1988] e no inglês camaronês [Ngefac, 2008] o uso de formas padrão não está relacionado ao gênero dos falantes. Esses resultados evidenciam que a correlação entre gênero e variação linguística está associada à organização social das comunidades estudadas.

Este trabalho difere dos anteriores na medida em que analisa o uso de hashtags do Twitter pelos usuários do sexo masculino e feminino, considerando o gênero como

um fator social capaz de influenciar na escolha de uma hashtag específica entre aquelas relacionadas a um determinado tema. Assim, é sugerido que hashtags possam ser estudadas como formas linguísticas e que as redes sociais online possam ser examinadas como redes nas quais o gênero desempenha um papel importante.

Encontrar características distintas entre os comportamentos de homens e mulheres ao utilizarem a Internet também tem sido um tópico de pesquisa. Alguns estudos analisaram a demografia da Internet e as diferenças existentes entre usuários do gênero feminino e masculino ao usar a Web e as redes sociais online [Bimber, 2000; Ono & Zavodny, 2003; Fallows, 2005; Ross et al., 2011]. Thelwall [2011] abordou a influência do gênero nas questões de privacidade na Web. Danescu-Niculescu-Mizil et al. [2012] mostraram que, em discussões na Web, alguns fatores linguísticos revelam diferenças de poder entre os membros, abordando inclusive as diferenças relativas a gênero.

Em uma outra perspectiva, diversos estudos já indicaram características próprias dos processos de variação e mudança linguística. Bailey [1973] e Kroch [1989] mostraram que a mudança no tempo tende a seguir uma curva no formato de S (*S-shaped curve*), com crescimentos lentos no início e no fim do processo. Weinreich et al. [1968] analisaram a existência de períodos de variação linguística interna no usuário, que não altera repentinamente o seu léxico, tornando assim a variação gradual não apenas no nível da sociedade, mas também no do indivíduo. Fischer [2007] verificou a multi-estabilidade, demonstrando que a direção da mudança não é fixa, pois, sob certas circunstâncias, podem ocorrer até mesmo movimentos reversos no processo.

Todos esses estudos buscam identificar as características que geram um fenômeno em princípio paradoxal - o que Nettle [1999] chama de *threshold problem*, interpretado por Troutman et al. [2008] da seguinte maneira:

> *Variantes inicialmente raras (...) conseguem se espalhar para inteiras comunidades de fala. Porém, isso é contra-intuitivo, pois os aprendizes deveriam adaptar suas falas para integrá-las ao ambiente. Se a maioria da população ainda está utilizando a forma antiga, um aprendiz deveria adotar essa forma também. Os aprendizes nunca deveriam usar mais da forma minoritária do que o resto da população.* (tradução nossa[1])

É o que pergunta Sapir [1921]: como pode uma variante inicialmente rara se espalhar para uma inteira comunidade de fala? Como se leva a cabo essa mudança

[1] Original: *Initially rare variants (...) manage to spread to entire speech communities. However, this is counterintuitive because learners should adapt their speech to match their environment. If the majority of the population is still using the older form, a learner should adopt that form as well. Learners should never use more of the minority form than the rest of the population.*

[Silva, 2006]? A mudança, consistindo na disseminação de variantes menos comuns para grande parte da rede ou até mesmo para toda a rede, apresenta-se, assim, como um fenômeno pouco esperado. No entanto, ocorre.

Dessa forma, conclui-se que: a) o processo de variação e mudança linguística possui uma série de características já descritas pela literatura; b) entretanto, as características relativas à propagação das formas variantes ainda não puderam ser vastamente descritas. Uma das razões da falta de descrições do fenômeno da propagação é a inexistência de mapeamentos dos caminhos percorridos pelas formas inovadoras nas redes de falantes, tarefa difícil ou impossível de se realizar em comunidades de fala offline. Uma das contribuições deste estudo é a elaboração da proposta de se analisar hashtags inovadoras como inovações linguísticas e a rede do Twitter como uma comunidade de fala, de maneira que, em trabalhos futuros, o processo de propagação das formas possa ser analisado com dados reais.

Capítulo 3

O Processo de Etiquetagem Textual

Etiquetas, ou *tags*, são, em sistemas de informação, palavras-chave ou termos associados a itens de conteúdo como imagens, textos, *bookmarks*, arquivos etc. Elas funcionam como metadados, isto é, informações sobre os objetos, na medida em que auxiliam e complementam a descrição dos itens e facilitam a busca posterior pelas informações relacionadas. Além disso, em alguns ambientes, as etiquetas parecem cumprir outras funções, tais como marcação de propriedade e de autoria, publicidade e indicação da identidade virtual dos usuários.

No mundo offline, o uso de etiquetas textuais para categorizar objetos não é um fenômeno recente. Cameron [2011] defende que elas têm sido usadas há séculos com funções comerciais e de catalogação, para identificação e classificação dos mais diversos itens e nas mais variadas situações, especialmente em museus e bibliotecas. Segundo Parry & Ortiz-Williams [2007], há pelo menos quatrocentos anos os museus utilizam etiquetas textuais contendo comentários e interpretações acerca do material disponível nas coleções e apresentado nas exposições - e, a despeito de toda a tecnologia disponível para arquitetar experiências multisensoriais de toda sorte, as etiquetas ainda sobrevivem. Hahn [2004] acrescenta ainda que, desde o século XVII, a marcação de animais por meio de etiquetas contendo determinados dados é uma técnica essencial para o estudo do comportamento das populações.

Portanto, a etiquetagem parece ser a maneira natural encontrada pelo ser humano para manter as informações desejadas próximas aos objetos referenciados. Smith [2011] complementa e considera que a etiquetagem é um sintoma da necessidade humana básica de criar uma percepção de ordem e de organização - muitas vezes até mesmo onde essa ordem não é fundamental.

A etiquetagem difere-se da *categorização*, recurso no qual as categorias surgem como pastas, ou seja, coleções de objetos relacionados a uma quantidade limitada de

tópicos geralmente pré-estabelecidos. Categorias tendem a estruturar de maneira mais eficiente o conteúdo do que as etiquetas, enquanto estas podem representar melhor as peculiaridades de cada objeto dada a maior liberdade concedida ao usuário no momento da sua atribuição.

3.1 Etiquetagem de Conteúdo Digital

A ciência da informação desenvolveu regras e esquemas elaborados para catalogação e categorização, os quais incluem processos precisos de classificação e vocabulários controlados para a descrição de tópicos [Mathes, 2004]. Afinal, tradicionalmente as etiquetas são designadas por catalogadores treinados, muitas vezes profissionais - é o caso, por exemplo, dos bibliotecários e dos arquivólogos.

Esses metadados criados profissionalmente possuem alta qualidade - são precisos, corretos e organizados - porém também geram custos muito elevados para serem produzidos, o que torna impossível a classificação profissional da grande quantidade de conteúdo novo que é postado e compartilhado diariamente na Web. Como informa Mejias [2004], a tarefa de processar e classificar todos esses itens é ainda mais difícil haja vista a rapidez com que novo conteúdo é produzido, tornando impossível o desenvolvimento e a manutenção de uma taxonomia capaz de dar conta, de maneira eficaz, de todo esse conteúdo.

Em função disso, para a classificação de conteúdo digital surgiu a alternativa da etiquetagem pessoal (*personal tagging*), ou etnoclassificação (*ethnoclassification*), isto é, a possibilidade da atribuição de etiquetas livremente escolhidas pelos próprios autores ou pelos usuários das ferramentas de compartilhamento. Dessa forma, a tarefa de classificar os itens é dividida entre o maior número possível de indivíduos interessados.

Os sistemas que oferecem essas alternativas são chamados de ambientes de livre etiquetagem (*free-tagging environments*) ou de classificação social/distribuída (*social/distributed classification*). Nesses ambientes, como as etiquetas não são criadas por especialistas, elas não seguem diretrizes formais. Não existe a imposição de uma taxonomia rígida, e sim a liberdade anárquica de uma *folksonomia*, termo cunhado por Wal [2007a] e que indica a participação das pessoas (*folks*) no processo taxonômico. Isso significa que os itens são etiquetados pelos próprios consumidores da informação e podem ser classificados por meio de quaisquer termos que definam uma relação entre o objeto e algum conceito na mente do usuário/etiquetador. Assim, algumas etiquetas são representações óbvias, enquanto outras acabam por fazer pouco sentido quando colocadas fora de um contexto conhecido apenas pelo autor ou pela sua comunidade.

O recurso da livre etiquetagem de conteúdo digital relaciona-se intimamente com o advento da consulta mediada por computadores e com o desenvolvimento de ambientes virtuais de interação e de colaboração, como redes sociais online, blogs e *wikis*. Wal [2007b] afirma que a etiquetagem pessoal de objetos digitais pode ter sua origem em 1988, com o software Lotus Magellan, o qual fornecia aos usuários a possibilidade de etiquetar livremente documentos e outros itens armazenados no disco rígido a fim de facilitar buscas posteriores.

No universo da Web, esse recurso se desencadeou a partir do lançamento do site Delicious (www.delicious.com), em 2003, que permite a adição de etiquetas aos *bookmarks* armazenados e compartilhados pelos seus membros [Keller, 2007]. Em 2004, o site Flickr (www.flickr.com) passou a oferecer o mesmo recurso para a catalogação de imagens, obtendo grande sucesso na tarefa de aumentar a "pesquisabilidade" desses itens [Garrett, 2005]. Nos anos seguintes, surgiram muitos outros sites oferecendo o mesmo recurso, entre eles o Youtube (www.youtube.com, para compartilhamento de arquivos de vídeo), o Last.fm (www.lastfm.com, para compartilhamento de arquivos de áudio) e, por meio das hashtags, o Twitter (www.twitter.com, para compartilhamento de micromensagens).

Segundo Sinha [2005], o crescimento no número de sistemas virtuais que possibilitam a livre etiquetagem de conteúdo parece estar vinculado à simplicidade desse processo no plano cognitivo em comparação com o processo de categorização, pois, basicamente, a etiquetagem elimina a difícil fase de decisão presente durante a designação de uma das categorias fixas. Isso é particularmente relevante quando se trata de objetos digitais, em relação aos quais a autora defende a existência de pouco *consenso cultural* [Weller, 2007] acerca das categorias associadas aos itens. Nov et al. [2008] acrescentam que a popularidade da etiquetagem de objetos digitais pode ser atribuída, pelo menos em parte, aos benefícios que os usuários recebem com a organização de grandes quantidades de informação e, muitas vezes, com o aumento do efetivo compartilhamento dos itens etiquetados.

De acordo com Wal [2005], os ambientes de livre etiquetagem na Web podem ser divididos em folksonomias abertas (*broad folksonomies*) e folksonomias restritas (*narrow folksonomies*)[1]. Segundo o autor, em uma folksonomia aberta, muitos usuários têm autonomia para etiquetar o mesmo objeto, mesmo que esse item tenha sido compartilhado ou postado por outrem. É o que acontece no Delicious, por exemplo, onde um usuário gera a informação e a torna acessível aos demais, os quais podem etiquetá-la utilizando-se de terminologia pessoal. A Figura 3.1 ilustra a estrutura de uma folkso-

[1]A tradução de *broad* e *narrow folksonomies* como folksonomias abertas e restritas foi proposta por Brandt [2009].

nomia aberta: grupos de usuários com o mesmo vocabulário etiquetam o objeto (ação indicada pelas setas na direção das etiquetas) com os próprios termos (representados pelos números). Esses usuários, então, encontram a informação (ação indicada pelas setas apontando na direção dos grupos de usuários) por meio das etiquetas que fazem parte do seu vocabulário.

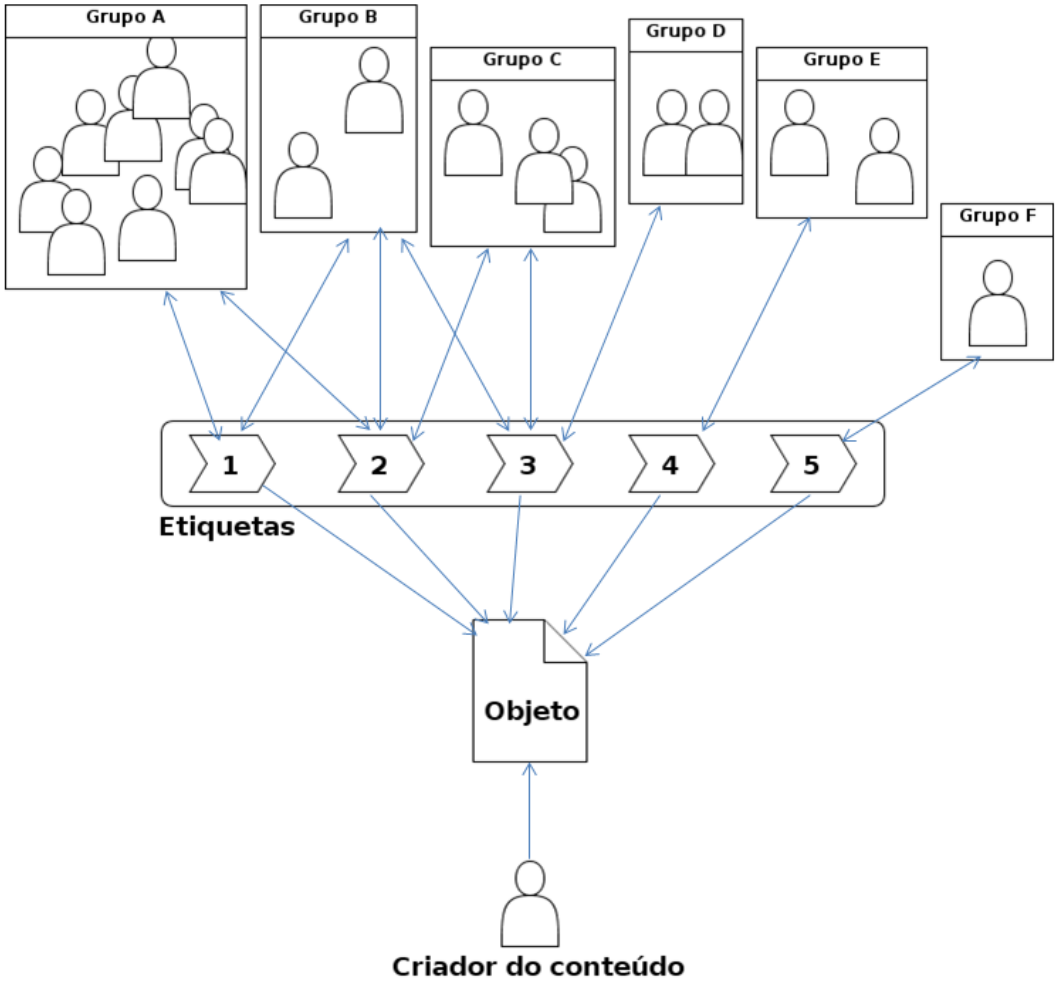

Figura 3.1. Estrutura de uma folksonomia aberta (adaptado de Wal [2005])

Em folksonomias restritas, porém, o usuário que compartilha o objeto é o responsável inicial pela sua etiquetagem. Os demais membros da rede podem recuperar o item utilizando a etiqueta designada pelo criador do conteúdo ou criar novas tags para fazer referência ao mesmo objeto. Assim se estrutura o Twitter com relação à atribuição de hashtags: o autor da micromensagem realiza a etiquetagem do objeto durante a postagem e os demais membros somente são capazes de atribuir novas hashtags caso *re-tweetem* o texto.

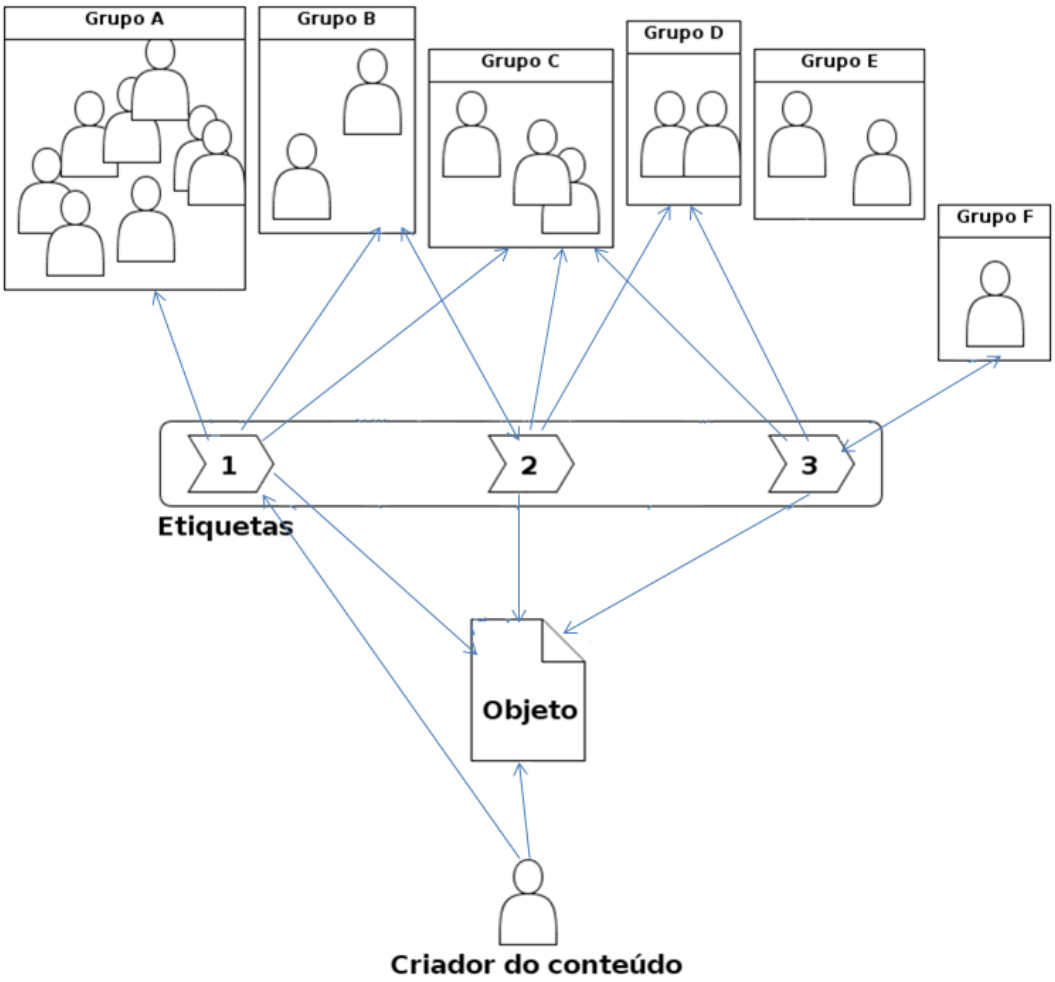

Figura 3.2. Estrutura de uma folksonomia restrita, como o Twitter (adaptado de Wal [2005])

3.1.1 Motivação dos Usuários do Twitter para a Etiquetagem

A fim de compreender melhor o comportamento dos usuários de redes sociais online, especialmente o seu comportamento de etiquetagem (*tagging behavior*), é importante examinar as suas motivações quando adicionam etiquetas aos itens de conteúdo, as razões que os levam a decidir em favor de determinadas etiquetas, a forma como os mesmos itens são classificados por diferentes usuários e como esses usuários modificam o próprio comportamento de etiquetagem de acordo com o comportamento dos seus amigos, entre diversas outras questões. Nesta seção, é apresentado um estudo realizado por meio de questionários aplicados a membros do Twitter para a obtenção de informações a respeito das motivações para a etiquetagem do conteúdo postado especificamente nessa rede social online, o que torna este trabalho inovador.

De acordo com Golder & Huberman [2006], as razões primárias encontradas pelos

usuários de mídias sociais para realizar a etiquetagem do conteúdo compartilhado são de cunho tanto organizacional quanto social e extrapolam as principais motivações apresentadas por Ames & Naaman [2007], a saber: a) fornecer informações contextuais adicionais acerca dos itens, aumentando assim a compreensibilidade da informação pelos seus amigos ou seguidores; e b) facilitar a recuperação posterior dos objetos. Golder & Huberman [2006] ainda acrescentam razões como a atração de atenção para o conteúdo, a participação em jogos, competições e promoções publicitárias, a expressão de opinião e a referência a si mesmos por parte dos etiquetadores. Zollers [2007] adiciona também o ativismo, ou seja, a utilização das etiquetas para a realização de algum tipo de campanha. Brandt [2009], em seu estudo acerca das motivações dos usuários para a etiquetagem de recursos no Delicious e no Flickr, concluiu que as motivações mais citadas nas entrevistas foram a recuperação dos itens, a organização dos objetos e o compartilhamento do conteúdo.

Entretanto, a utilização de hashtags no Twitter parece se apresentar como uma variação da etiquetagem tradicional realizada em sites de compartilhamento de *bookmarks* e fotos. Isso se deve ao fato de que micromensagens não tendem a ser buscadas posteriormente, ao contrário das imagens, por exemplo, que são a todo o tempo recuperadas pelos usuários. Além disso, a prática no Twitter tem mostrado que a maioria das hashtags possui uma vida curta a curtíssima, o que aumenta a variabilidade das tags e vai de encontro a uma das ideias básicas da indexação tradicional, que é a homogeneização das etiquetas.

Com o objetivo de complementar a bibliografia sobre a motivação para o uso de etiquetas na Web, foi realizada uma pesquisa com usuários do Twitter sobre a utilização de hashtags nas mensagens postadas especificamente nessa rede social online. Foram elaboradas duas versões de um mesmo questionário, as quais foram disponibilizadas em um website cuja URL foi divulgada entre usuários do Twitter a partir de tweets postados no perfil do autor e retweetados por vários de seus seguidores. As perguntas constantes de cada uma das versões do questionário eram as mesmas, estando as únicas diferenças presentes nas respostas, as quais, na versão 1, eram exclusivamente de múltipla escolha (questionário estruturado) e, na versão 2, eram divididas entre abertas e de múltipla escolha (questionário semi-estruturado). O objetivo da diferenciação entre as estruturas das duas versões jaz na necessidade de verificação da influência da presença de opções de respostas no questionário estruturado sobre o raciocínio do entrevistado. Cada versão foi também traduzida e disponibilizada em inglês. Para a elaboração dos questionários, procurou-se seguir as orientações fornecidas por Sensorpro.net [2012].

Os questionários estiveram disponíveis para preenchimento entre os dias sete de janeiro e doze de abril de 2012. Entretanto, apenas um questionário era disponibi-

lizado a cada dia, pois as versões se alternavam a cada 24 horas para uma melhor homogeneização das amostras entre os dois grupos. A aplicação dos questionários cessou quando foram alcançados, em cada grupo, duzentos indivíduos válidos, ou seja, que responderam a todas as indagações solicitadas, totalizando assim uma amostra composta precisamente por quatrocentos sujeitos.

Os questionários foram divididos em três seções, sendo que apenas a terceira delas - composta por apenas uma questão - é diferente entre as duas versões. A seção inicial diz respeito às informações demográficas: gênero e faixa etária dos indivíduos. A Tabela 3.1 indica a distribuição dos sujeitos com relação a essas características, que se apresentam homogêneas entre os dois grupos.

Tabela 3.1. Distribuição dos sujeitos da amostra com relação às características de idade e gênero

Grupo	Faixa etária	Gênero	
		Feminino	Masculino
1 (questionário estruturado)	15-24	21 (10,5%)	37 (18,5%)
	25-34	40 (20%)	51 (25,5%)
	35-44	14 (7%)	19 (9,5%)
	45 ou mais	5 (2,5%)	13 (6,5%)
2 (questionário semi-estruturado)	15-24	24 (12%)	39 (19,5%)
	25-34	36 (18%)	46 (23%)
	35-44	9 (4,5%)	22 (11%)
	45 ou mais	7 (3,5%)	17 (8,5%)

A segunda seção traz as seguintes questões acerca do comportamento dos usuários no Twitter e do uso de hashtags nas postagens:

- Questão 1) Em média, com que frequência você posta no Twitter?

  ( ) mais de uma vez por dia

  ( ) uma vez por dia

  ( ) pelo menos uma vez por semana

  ( ) pelo menos uma vez a cada duas semanas

  ( ) menos de uma vez a cada duas semanas

- Questão 2) Você já deve ter visto, no Twitter, termos que se iniciam com o sinal #. Nós os chamamos de *hashtags*. Em média, com que frequência você insere hashtags nos seus tweets?

  ( ) em todos ou praticamente todos os tweets que posto

( ) na maioria dos tweets que posto

( ) em alguns tweets que posto

( ) utilizei as hashtags poucas vezes

( ) nunca utilizei uma hashtag

A fim de relacionar as duas questões da segunda seção, procedeu-se da seguinte maneira: para cada item da questão 2 foi atribuída uma pontuação de 1 a 5, sendo 1 o item "nunca utilizei uma hashtag" e 5 o item "insiro hashtags em todos ou praticamente todos os tweets que posto". Em seguida, para cada item da questão 1, calculou-se a média aritmética das pontuações associadas aos itens da questão 2. Verificou-se que a frequência de postagem no Twitter influencia na utilização de hashtags nas mensagens: quanto mais tweets um usuário posta, maior a pontuação média da questão 2, ou seja, maior a probabilidade de que ele use hashtags regularmente. Esse resultado não é necessariamente esperado, visto que as opções de resposta da questão 2 não se diferenciam pela temporalidade como as opções de resposta da questão 1. Por exemplo, um indivíduo pode postar pouco frequentemente no Twitter, mas mesmo assim ser um usuário ativo de hashtags. Dessa forma, conclui-se que as hashtags são um recurso preferido pelos usuários mais experientes. O gráfico apresentado na Figura 3.3 ilustra essa relação.

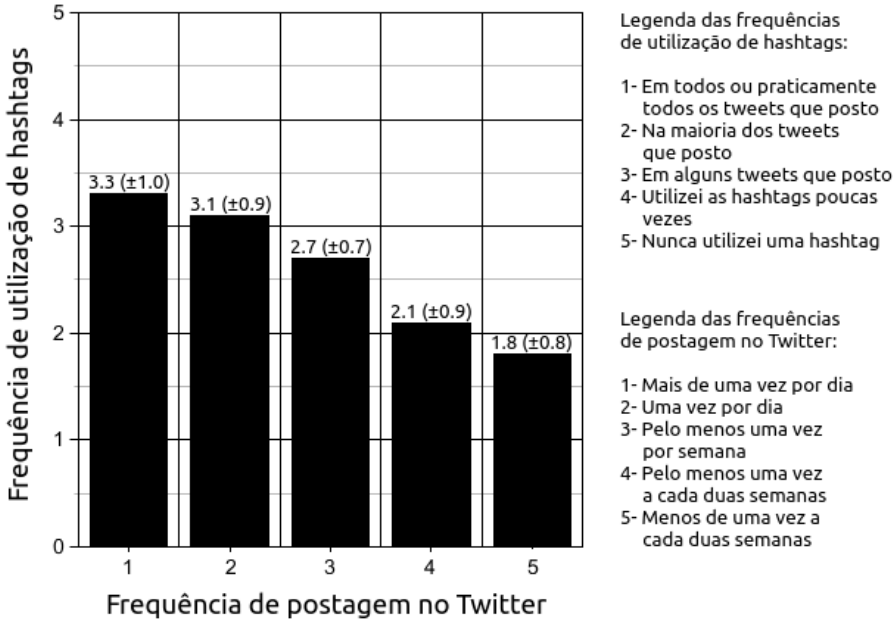

Figura 3.3. Relação entre a frequência de postagem no Twitter e a média (± desvio padrão) da frequência de utilização de hashtags nos tweets

Verificou-se ainda que a variável gênero não exerce influência na frequência das postagens e nem mesmo na utilização de hashtags. Contudo, a idade parece ser um fator importante para a caracterização dos usuários mais ou menos frequentes nessa rede social: 45,2% dos sujeitos com mais de 45 anos de idade declararam que postam no Twitter menos de uma vez a cada duas semanas e 71,4% informaram que nunca utilizaram ou que utilizam as hashtags em poucas ocasiões. Os índices correspondentes relativos aos usuários dos dois grupos mais jovens (de 15 a 24 anos e de 25 a 34 anos) são, respectivamente, 13,6% e 24,1%.

A terceira seção, finalmente, indaga sobre a principal motivação para a utilização de hashtags. Para o grupo 1, foram disponibilizadas opções de acordo com o que a literatura consultada considera como motivações para a utilização de etiquetas em rede sociais e, mais especificamente, de hashtags no Twitter. Para o grupo 2, essas opções não foram mencionadas, ficando a cargo dos entrevistados a redação da motivação principal. A questão proposta foi a seguinte:

- Questão 3) Se você utiliza ou já utilizou alguma hashtag, qual o principal motivo que te leva/levou a isso? [na versão 2 do questionário, esta pergunta não traz múltiplas escolhas na resposta, e sim um espaço de duas linhas para livre reflexão do entrevistado]

  ( ) Para que meus tweets sejam acessados e lidos por mais pessoas

  ( ) Para participar de grupos de discussão via Twitter

  ( ) Para facilitar a compreensão das minhas mensagens

  ( ) Para participar de jogos e brincadeiras

  ( ) Para participar de promoções

  ( ) Para que eu possa recuperar os meus tweets no futuro com mais facilidade

A Figura 3.4 mostra as motivações indicadas pelos entrevistados do grupo 1. Mais de 70% deles informaram que etiquetam seus tweets a fim de aumentar a compreensibilidade das mensagens ou para que estas tenham mais possibilidades de serem acessadas e, como consequência, de serem compartilhadas na rede.

A análise das motivações indicadas livremente pelos entrevistados do grupo 2 mostra um resultado ligeiramente diferente, como indica a Figura 3.5. A definição da categoria da motivação foi realizada manualmente para cada resposta. Motivações como "explicar o tema do tuíte" e "ajudar os meus seguidores a entender o que eu queria dizer" foram categorizadas em "compreensibilidade"; outras como "para o tweet aparecer no alto da lista de quem procura a hash tag" e "acho que as pessoas retweetam

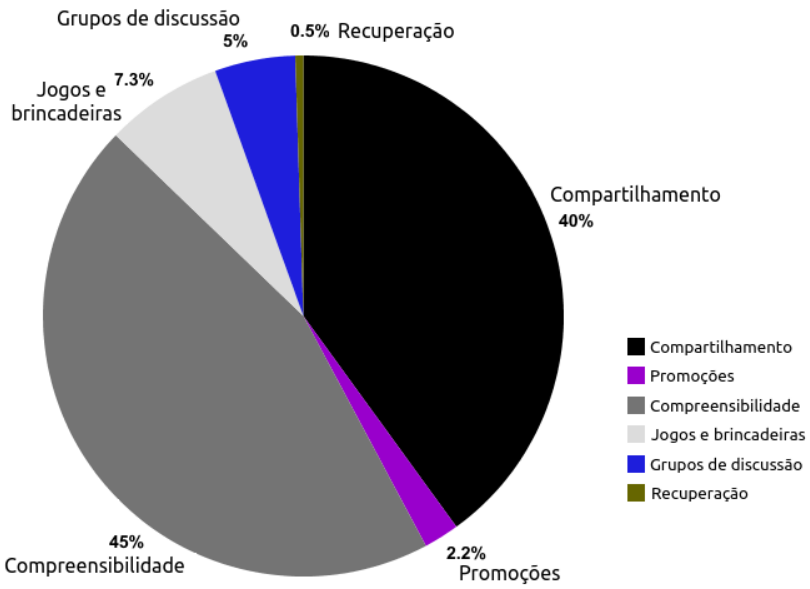

Figura 3.4. Motivações para o uso de hashtags no Twitter, segundo usuários do grupo 1

mais quando tem hashtag" foram categorizadas em "compartilhamento"; "falar sobre memes", por exemplo, foi categorizado em "participar de jogos"; e, finalmente, "poder postar em discussões sobre um tema específico" foi a única motivação apresentada pelos entrevistados do grupo 2 categorizada como "grupos de discussão".

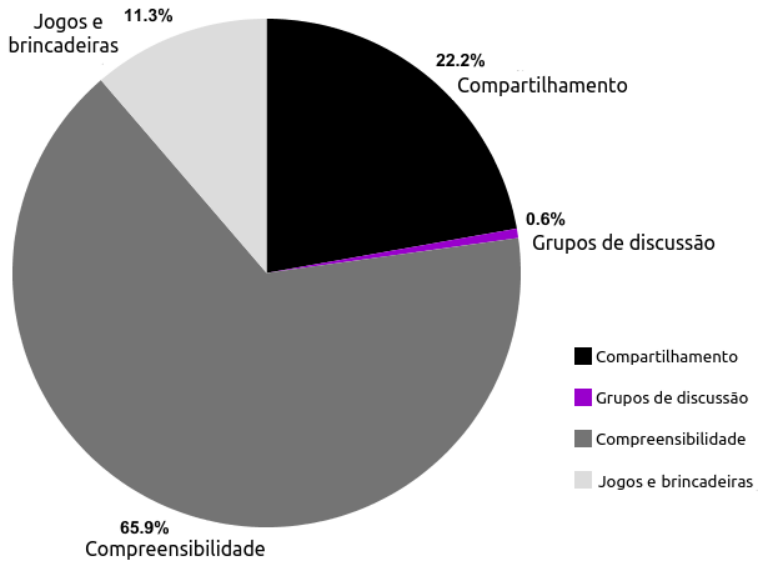

Figura 3.5. Motivações para o uso de hashtags no Twitter, segundo usuários do grupo 2

É interessante notar como algumas motivações parecem ser lembradas apenas quando são citadas pelo questionário, o que se torna mais curioso dada a homogeneidade dos grupos de entrevistados. É o caso da participação em promoções e da recuperação futura dos tweets, que sequer apareceram nas respostas da questão não estruturada. De forma similar, o compartilhamento é mais citado no grupo que respondeu à questão estruturada, enquanto que o aumento da compreensibilidade das mensagens é a principal motivação entre aqueles que não tiveram sugestões de resposta.

Capítulo 4

Apresentação dos Dados

Este capítulo trata da constituição dos conjuntos de dados utilizados na parte experimental do trabalho: sua composição, coleta e características gerais.

4.1 Constituição dos Datasets

Neste estudo, foram utilizados dois grandes conjuntos de dados. O primeiro deles consiste em todas as mensagens postadas por todos os usuários donos de perfis públicos do Twitter no período que se inicia com a criação do site, em julho de 2006, até o mês de agosto de 2009. Para a realização da coleta, empregou-se a Interface de Programação de Aplicativos (API) oficial do Twitter, em um processo que durou cerca de um mês utilizando 58 servidores no Max Plank Institute for Software Systems, na Alemanha [Cha et al., 2010]. O Twitter permitiu a coleta de dados sobre cada usuário, incluindo as suas conexões na rede, e de todos os tweets postados por eles. No total, foram coletadas informações de perfil de 54.981.152 membros do Twitter, 1.923.263.821 conexões entre membros e todos os 1.755.925.520 tweets postados por perfis públicos durante o período. Aproximadamente 8% dos perfis foram definidos pelos usuários como privados, de maneira que autorizam apenas os seguidores - e não o API - a acessarem os seus tweets. Esses usuários são ignorados em todas as análises apresentadas nesta dissertação. Mais informações a respeito do dataset podem ser obtidas na página oficial do Twitter Project, em http://twitter.mpi-sws.org/.

A respeito desse conjunto de dados, Rodrigues et al. [2011] informam que a topologia da rede é similar àquelas de outras mídias sociais online, como o Facebook. Enquanto uma fração muito pequena de usuários possui um alto número de vizinhos, a maioria deles possui apenas poucas conexões: 99% participam de menos de vinte conexões, entre seguidos e seguidores. Os perfis mais populares - isto é, com maior grau

de entrada - incluem figuras públicas como Barack Obama, celebridades como Oprah Winfrey e fontes de notícias como a BBC. É interessante notar que, diferentemente de outras redes sociais, a reciprocidade entre seguidos e seguidores é pequena: apenas 23% dos links são bi-direcionais, isto é, com usuários que se seguem mutuamente. O restante das conexões é uni-direcional e representa uma relação em que o usuário A segue B, porém não o contrário. Cha et al. [2010] acrescentam que a rede coletada é composta por um grande componente conectado que contém 94,8% dos usuários - e agrega 99% de todas as conexões e mensagens postadas -, além de 5% de vértices isolados e uma pequena parcela, correspondente a 0,2% da rede, formada por componentes menores.

O segundo dataset foi construído a partir de dados obtidos pelo Instituto Nacional de Ciência e Tecnologia para a Web (InWeb), o qual, em seu projeto Observatório da Web, também utiliza um API do Twitter para coletar tweets sobre tópicos específicos com o objetivo de monitorar eventos importantes e de criar indicadores visuais destinados a apresentar que tipos de conteúdo e de informação estão circulando na Web [Santos et al., 2010]. Nesta dissertação, são utilizados os dados relativos às eleições brasileiras de 2010, coletados entre 02 de março e 17 de dezembro daquele ano. Mais informações a respeito dos dados coletados pelo InWeb podem ser obtidos nos sites do Instituto (http://www.inweb.org.br/) e do Observatório da Web (http://observatorio.inweb.org.br/).

Ambos os datasets contêm, além dos tweets propriamente ditos e, consequemente, das hashtags utilizadas, algumas informações pessoais dos membros da rede, inclusive os seus nomes. Esses dados pessoais são utilizados nesta dissertação na seção em que se analisa a relação entre a preferência por certas categorias de etiquetas e o gênero dos usuários. É importante frisar que todas as informações pessoais que compõem os datasets aqui apresentados foram coletadas a partir de perfis definidos como públicos pelos próprios usuários, de maneira que a sua utilização não configura nenhum tipo de violação de privacidade.

4.2 Constituição dos Subdatasets

Já que em alguns dos estudos propostos neste trabalho é necessário analisar as características do fenômeno da variação de hashtags, tornou-se fundamental encontrar etiquetas intercambiáveis, ou seja, tags concorrentes usadas com o objetivo de categorizar mensagens sobre o mesmo tema. Isso corresponde à característica básica das formas linguísticas variantes, que, embora tenham aspectos diferentes, são usadas pelos falantes para nomear os mesmos elementos. Analisam-se, assim, situações que admi-

tem variação linguística, ou seja, o uso de diferentes formas linguísticas - neste caso, de hashtags - mesmo quando os valores semânticos e funcionais são equivalentes.

Com o objetivo de encontrar essas hashtags intercambiáveis, foram coletados tweets sobre temas específicos. Foi possível verificar a existência de hashtags diferentes usadas para categorizar mensagens que poderiam ser agrupadas em uma só categoria. Por exemplo, hashtags como #michaeljackson, #mj, #jackson, entre muitas outras, referem-se ao mesmo assunto e, em um ambiente de etiquetagem controlada, provavelmente seriam condensadas em apenas uma tag.

A partir do primeiro dataset, foram selecionados três temas relevantes, a saber: "Michael Jackson" (a morte do cantor foi amplamente divulgada e comentada nas redes sociais), "Gripe Suína" (a epidemia de Influenza A H1N1 foi um grande tópico de 2009, em especial no período correspondente ao inverno no hemisfério norte) e "Music Monday" (relacionado a uma campanha bem sucedida em favor de se postar tweets associados a música às segundas-feiras). Foram construídos, então, subconjuntos de dados contendo tweets e hashtags sobre cada um desses tópicos. Os subconjuntos foram construídos após a filtragem de tweets que incluíssem pelo menos uma hashtag e pelo menos um dos termos considerados relacionados aos temas. Dessa forma, no subconjunto "Michael Jackson", por exemplo, foram reunidos todos os tweets incluindo o termo "michael jackson" e que contivessem pelo menos uma hashtag. A Tabela 4.1 apresenta dados de cada subconjunto: número de tweets publicados, número de usuários que postaram tweets, número de conexões entre os usuários do subconjunto e número de hashtags diferentes presentes nesse subconjunto.

Tabela 4.1. Informações sobre os subdatasets "Michael Jackson", "Gripe Suína" e "Music Monday"

Tópico	Tweets	Usuários	Conexões	Hashtags diferentes
Michael Jackson	221.128	91.176	3.171.118	19.679
Gripe Suína	295.333	83.211	5.806.407	17.196
Music Monday	835.883	196.411	7.136.213	16.005

A partir do segundo dataset, que inclui apenas tweets sobre as eleições brasileiras de 2010, foram obtidos quatro subconjuntos de dados, relacionados com as posições políticas dos membros da rede nas eleições em questão: a) apoiadores de Dilma Rousseff; b) apoiadores de José Serra; c) opositores de Dilma Rousseff; e d) opositores de José Serra[1]. Esses subconjuntos de dados foram construídos de acordo com o conteúdo das hashtags, as quais foram manualmente associadas a uma das quatro posições políticas

[1]Dilma Rousseff e José Serra foram os candidatos mais votados na corrida presidencial de 2010.

citadas. Hashtags consideradas neutras, isto é, não expressando suporte ou oposição a nenhum candidato (como #eleições e #votabrasil) ou expressando outras posições (por exemplo, apoiando outros candidatos, como #votemarina e #plinio50) foram excluídas das análises. Esses casos representam 62,4% da totalidade de hashtags presentes no dataset. A Tabela 4.2 apresenta alguns exemplos de etiquetas que fazem parte dos quatro subdatasets construídos.

Tabela 4.2. Exemplos de hashtags que formam os subdatasets construídos a partir dos dados obtidos de tweets acerca das eleições brasileiras de 2010

Apoiadores de Dilma Rousseff	Apoiadores de José Serra
#dilma13	#serra45
#votodilma	#votoserra
#dilmapresidenta	#br45il
#soudilma	#45confirma
Opositores de Dilma Rousseff	**Opositores de José Serra**
#forapt	#forapsdb
#dilmamente	#serramilcaras
#dilmanao	#serranao
#dilmafujona	#serracaluniador

Os dados relativos às etiquetas que têm como tópico as eleições brasileiras foram utilizados apenas na análise da influência do fator gênero no processo de designação das tags.

Capítulo 5

Análise dos Dados

Este capítulo torna presentes as análises dos dados que constituem as amostras coletadas. Em um primeiro momento, é exposta uma caracterização geral de alguns aspectos relativos ao uso de hashtags no Twitter. Em seguida, são analisados fatores linguísticos que influenciam na aceitação de tags pelos membros dessa rede social online. Por fim, é apresentado um estudo inovador que relaciona um fator social - o gênero dos usuários - à utilização de hashtags. Partes deste capítulo foram publicadas por Cunha et al. [2011, 2012].

5.1 Caracterização Geral

Nesta seção, são apresentadas algumas informações de caracterização obtidas a partir das análises efetuadas nos subconjuntos de dados sobre os tópicos "Michael Jackson", "Gripe Suína" e "Music Monday".

5.1.1 Frequência de Utilização das Hashtags

Nos gráficos apresentados na Figura 5.1, pode-se observar a evolução do número de tweets etiquetados sobre diferentes tópicos em determinados intervalos de tempo e, em especial, as diferentes dinâmicas de utilização de tags entre esses tópicos. Os gráficos (a), (b), (c) e (d) dizem respeito à frequência absoluta das hashtags sobre, respectivamente: "Michael Jackson", "Gripe Suína", "Music Monday" e "Empregos"[1].

[1] Os dados sobre "Empregos" são compostos por todas as hashtags coletadas em mensagens que continham também a etiqueta #job ou #jobs. Esses dados não são utilizados em nenhuma outra análise nesta dissertação e surgem aqui apenas com o objetivo de ilustrar uma dinâmica de utilização de hashtags diferente das demais.

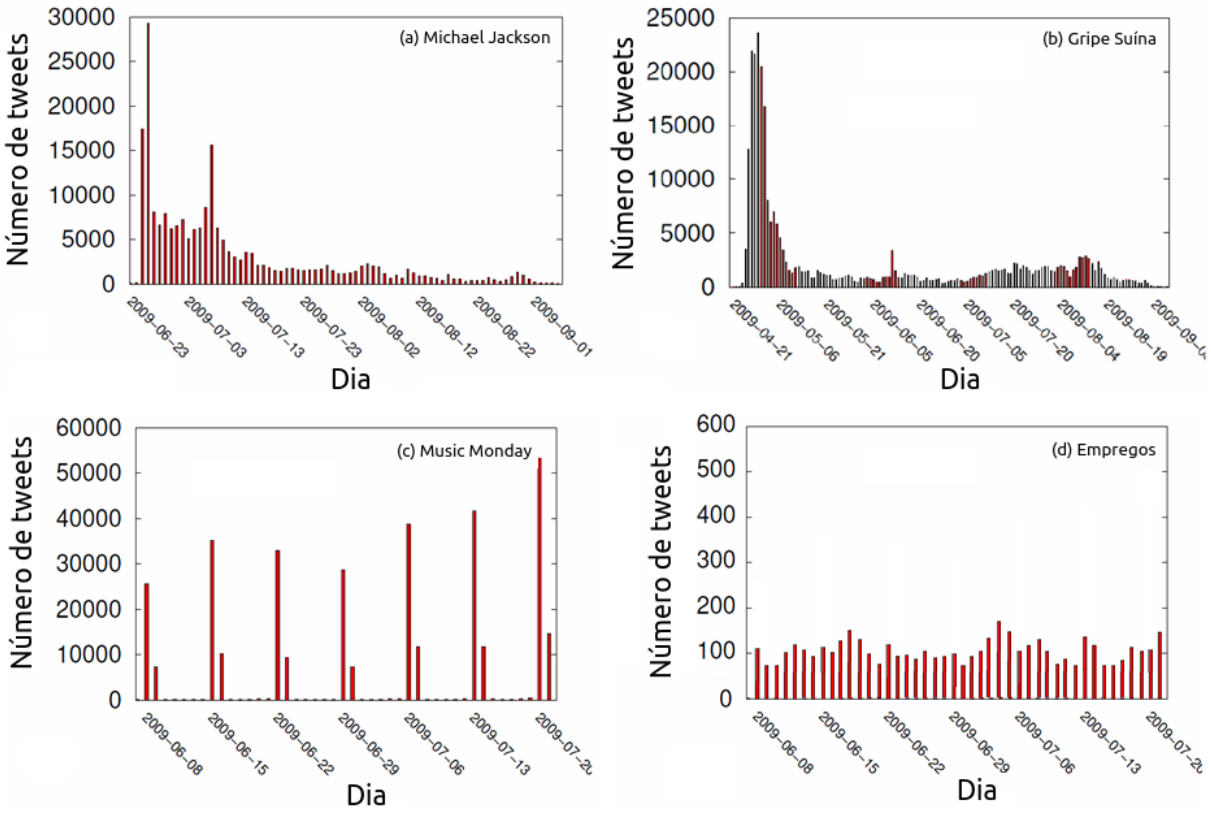

Figura 5.1. Frequência absoluta da utilização hashtags sobre determinados tópicos em função do tempo

Os picos nos dois primeiros gráficos mostram a alta utilização das hashtags em curtos períodos. Nos casos em questão - tópicos "Michael Jackson" e "Gripe Suína"-, essa dinâmica reflete a existência de eventos repentinos, ou seja, morte e funeral de Michael Jackson e início da disseminação da gripe suína. No terceiro gráfico, no entanto, os picos são sazonais, pois a hashtag #musicmonday é utilizada como um categorizador de mensagens que fazem referência a música postadas apenas às segundas-feiras. Nesse gráfico, a existência de picos menores também às terças-feiras pode ter duas explicações: a) as diferenças de fuso-horário entre os usuários ao redor do mundo; e b) a repercussão que mensagens do dia anterior podem ter ainda no dia seguinte, gerando *retweets* (reenvios da mesma mensagem por algum usuário que a tenha recebido) e comentários. No quarto gráfico, não há picos, mas apenas uma utilização discreta e constante das hashtags.

Estes são os três padrões encontrados para a dinâmica de utilização de hashtags no Twitter: picos repentinos (gráficos a e b), sazonalidade (gráfico c) e constância (gráfico d). Outros tópicos e hashtags foram analisados e todos eles parecem seguir um desses padrões, de acordo com características alheias à própria etiqueta e intimamente

relacionadas aos seus elementos geradores.

É interessante registrar ainda como as curvas de ocorrência de consultas na Web se relacionam com as frequências de utilização de hashtags no Twitter. As quantidades de consultas pelos termos "michael jackson" e "swine flu" no motor de busca do Google, obtidas graças à ferramenta Google Trends (www.google.com/trends), oferecem um bom *fitting* para os gráficos de frequência de utilização das hashtags, como pode ser verificado nos gráficos da Figura 5.2 relativos a períodos de um mês de consultas na Web sobrepostos à frequência de utilização, no mesmo período, de hashtags acerca dos mesmos tópicos. Considerando que a ocorrência de consultas nos sistemas de busca possa ser um indicador de interesse coletivo por um determinado assunto, supõe-se que a frequência de postagens de mensagens e de uso de hashtags sobre um tema específico também indique algo semelhante.

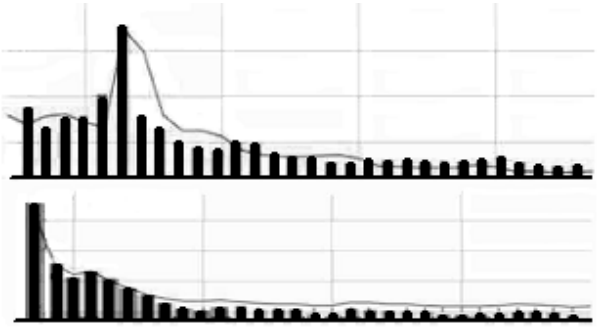

Figura 5.2. Frequência de utilização de hashtags x frequência de consultas no Google

5.1.2 Frequência de Hashtags Distintas

Nos gráficos da Figura 5.3, podem ser visualizadas as evoluções do número de hashtags distintas pertencentes aos subconjuntos de dados em função do tempo. Mais uma vez, os dois primeiros gráficos (respectivamente referentes a "Michael Jackson" e a "Gripe Suína") apresentam comportamento similar caracterizado por picos repentinos, enquanto o terceiro, referente aos dados de "Music Monday", possui características diferentes, mantendo a sazonalidade. Verifica-se que a taxa de criação de novas hashtags acompanha o interesse geral pelo tópico, ou seja: quanto mais se fala sobre um assunto, não apenas as etiquetas já criadas passam a ser mais utilizadas, mas também mais etiquetas distintas sobre o tema surgem.

É importante notar a linha que indica a fração de novas hashtags criadas naquele dia, isto é, de hashtags que nunca haviam sido utilizadas anteriormente. Pode-se observar que, nos picos de maior atividade, a taxa de novidade também foi alta e

mesmo depois se manteve significante, variando entre 10% e 40%, ou ainda mais, no caso da base "Music Monday". Isso parece indicar um alto índice de inovação no que tange as hashtags, fato que pode se relacionar também a seus curtos ciclos de vida já identificados em estudos relacionados.

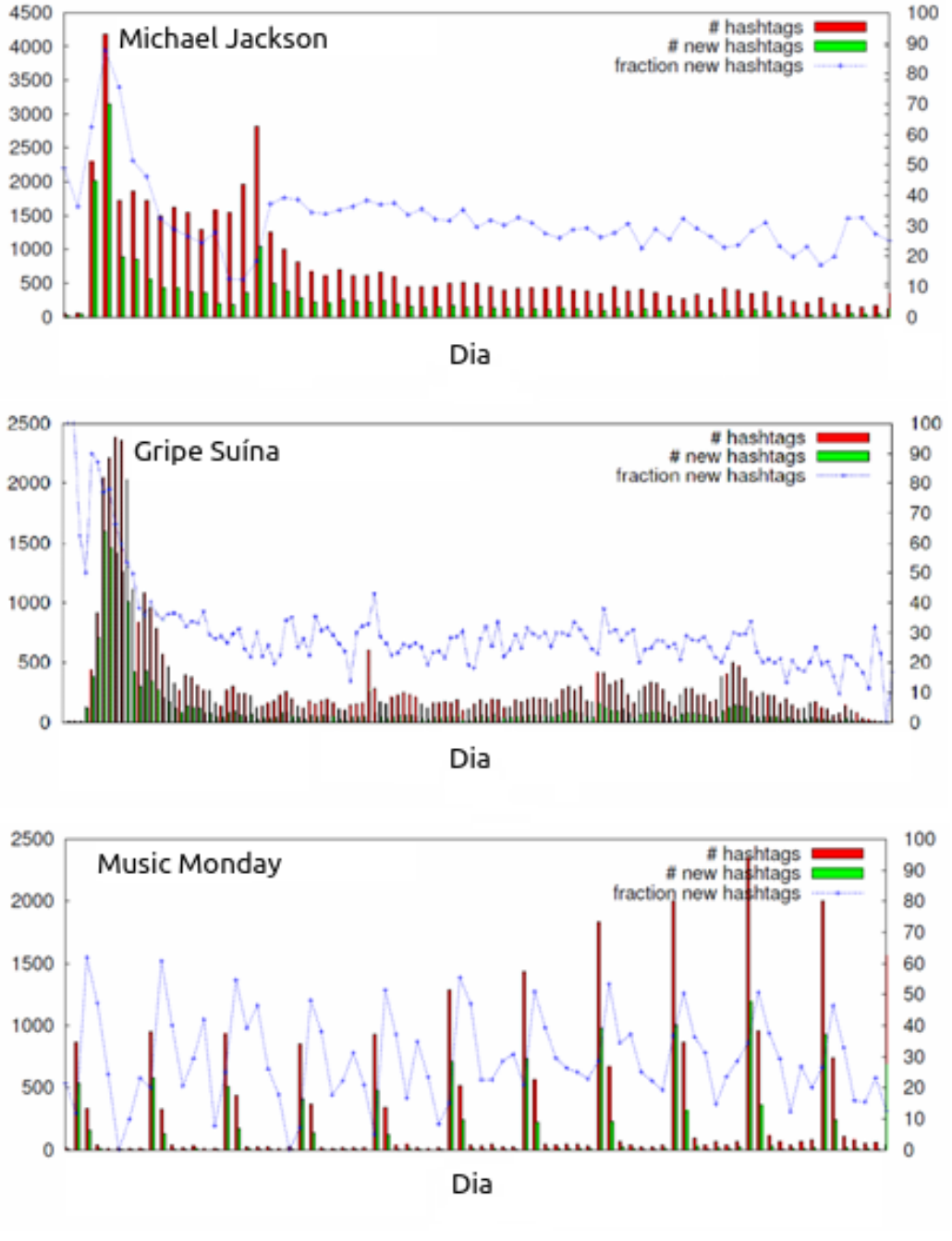

Figura 5.3. Frequência de hashtags distintas (*#hashtags*) e de hashtags novas (*#new hashtags*) por dia, além da fração de hashtags novas no total de ocorrências diárias (*fraction new hashtags*)

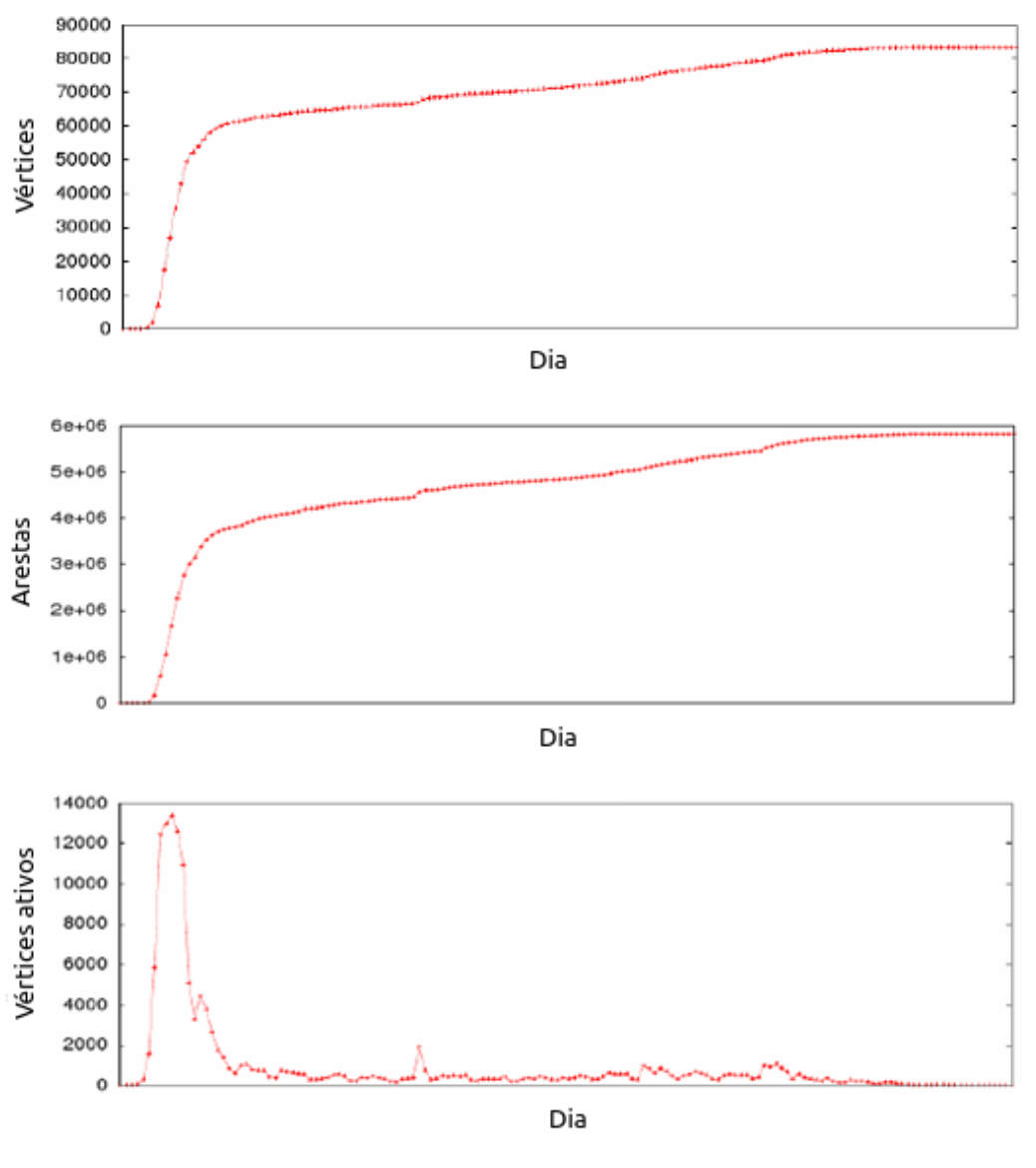

Figura 5.4. Total de vértices, de arestas e de vértices ativos no conjunto de dados "Gripe Suína", em função do tempo

Os gráficos da Figura 5.4 fazem referência apenas aos dados relativos ao subconjunto da "Gripe Suína". Pode-se observar a evolução cumulativa do número de vértices - representando os usuários - que utilizaram hashtags sobre o tema e o número de arestas entre esses vértices - representando as conexões entre os usuários existentes na base de dados -, respectivamente. Os números de vértices e de arestas possuem crescimentos logarítmicos, resultados do aumento mais rápido do volume de atividades de usuários no período inicial das postagens. O terceiro gráfico da figura, que mostra o número de vértices ativos por dia, representa a quantidade de usuários que etiquetaram suas mensagens em cada dia. A dinâmica desse gráfico segue o mesmo padrão, inclusive com

os mesmos picos, dos gráficos apresentados anteriormente. Isso sugere que o número de usuários que postam sobre um tema está relacionado com o interesse gerado pelo tópico naquele dia, como seguramente já era de se esperar.

5.1.3 Subgrafos Representativos da Propagação

Foram construídos subgrafos representativos da rede para que as diferentes dinâmicas de propagação encontradas pudessem ser ilustradas. Nesses subgrafos, os vértices indicam usuários que, em algum momento, passaram a integrar a rede de indivíduos que utilizaram hashtags sobre o tópico e as arestas indicam a relação do tipo seguidos/seguidores entre eles. A janela temporal entre cada subgrafo é de um dia.

Os subgrafos apresentados na Figura 5.5 mostram os quatro primeiros dias de utilização de hashtags sobre os temas "Gripe Suína" (a) e "Music Monday" (b), respectivamente. É fundamental notar as diferenças entre as duas sequências, que indicam um comportamento distinto dos usuários nos dois tópicos: os subgrafos representativos da base "Gripe Suína" possuem poucas arestas entre os nós, sugerindo que não houve influência de usuários uns sobre os outros para o compartilhamento das hashtags, pois cada membro postou a sua mensagem etiquetada isoladamente. Essa parece ser uma tendência de tópicos que possuem como gatilho inicial um acontecimento social offline - como é o caso da gripe suína e da morte de Michael Jackson -, já que, em um determinado instante, diferentes vértices não necessariamente conectados entre si postam mensagens acerca desses temas. Esse fenômeno não ocorre com os tweets sobre "Music Monday": afinal, a hashtag #musicmonday foi criada por um usuário e, essa sim, se propagou por influência para os seus seguidores, como mostra a segunda sequência de subgrafos, que indica claramente o processo de disseminação dessa tag. Essa dinâmica se repete com as hashtags menos intuitivas, enquanto a dinâmica dos subgrafos da sequência (a) é o padrão entre as hashtags mais intuitivas, que surgem simultaneamente em diversos pontos da rede.

5.1.4 Processo de Conexão Preferencial

Easley & Kleinberg [2010] e Vera [2011] caracterizam o que é conhecido como *rich-get-richer phenomenon* ou "processo de conexão preferencial": em alguns ambientes, a popularidade dos itens mais comuns tende a crescer mais rapidamente do que a popularidade dos itens menos comuns. Esse fenômeno gera uma propagação ainda maior das formas que alcançam um determinado prestígio e uma estagnação daquelas que não o alcançam. Zipf [1949] testou e confirmou que a frequência das palavras em

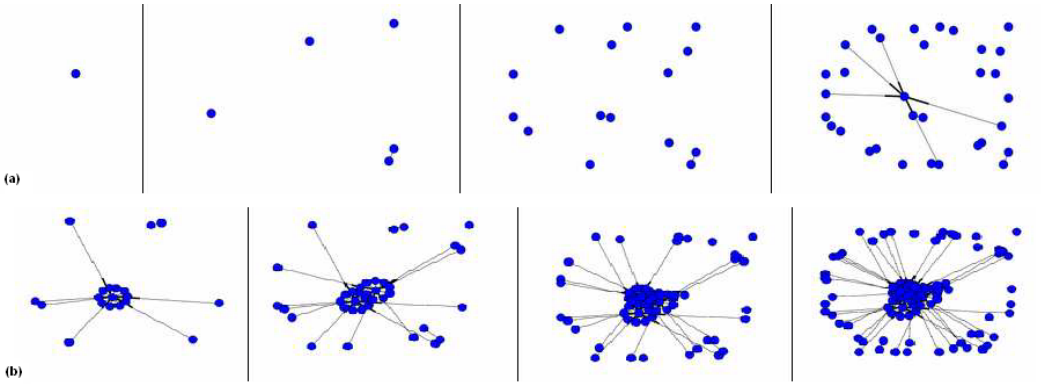

Figura 5.5. Subgrafos representativos da propagação de hashtags nas bases "Gripe Suína" (a) e "Music Monday" (b)

inglês e em outras línguas segue uma lei de potência, de forma que poucas delas são muito frequentes, enquanto a grande maioria é utilizada em poucas ocasiões. Com o objetivo de verificar se algum tipo de padrão é seguido na distribuição das hashtags, foram analisados os dados do Twitter. As Tabelas 5.1 e 5.2 mostram informações sobre a distribuição de hashtags em cada um dos subdatasets.

Tabela 5.1. Distribuição das hashtags menos utilizadas em cada base

Tópico	% de hashtags com até i utilizações		
	i=1	i=2	i=10
Michael Jackson	59%	72%	88%
Gripe Suína	59%	73%	92%
Music Monday	60%	74%	91%

Tabela 5.2. Distribuição das hashtags mais populares em cada base

Tópico	Número de hashtags com mais de j utilizações		
	j=10.000	j=5.000	j=1.000
Michael Jackson	3	6	28
Gripe Suína	3	4	14
Music Monday	2	3	28

O percentual de hashtags em relação ao número de tweets em que elas são utilizadas é consideravelmente similar em cada uma das três bases. Isso parece confirmar a possível existência de um padrão *rich-get-richer*: poucas etiquetas - as mais populares - são utilizadas na maioria dos tweets, enquanto a grande maioria delas surge em apenas poucas postagens. A Tabela 5.1 mostra que em torno de 60% das hashtags são usadas

apenas uma vez, isto é, elas não se propagam para o resto da rede; aproximadamente 90% delas não são usadas mais do que dez vezes, o que mostra que grande parte das hashtags está restrita a apenas um usuário ou a uma comunidade muito pequena de usuários.

Por outro lado, da mesma forma que Zipf [1949] demonstrou para línguas naturais, as hashtags mais utilizadas possuem frequências de uso muito altas. A Tabela 5.3 mostra dados das três hashtags mais usadas em cada uma das bases e sugerem que, também no Twitter, o comportamento de etiquetagem de um usuário depende das escolhas realizadas pelos outros membros da rede que o influenciam [Easley & Kleinberg, 2010].

Tabela 5.3. Dados das hashtags mais usadas em cada base

Tópico	Mais popular	2a. mais popular	3a. mais popular
Michael Jackson	#michaeljackson 35.861 ocorrências 12,3% do total	#michael 27.298 ocorrências 9,3% do total	#mj 16.758 ocorrências 5,7% do total
Gripe Suína	#swineflu 230.457 ocorrências 51,5% do total	#h1n1 70.693 ocorrências 15,8% do total	#swine 12.444 ocorrências 2,8% do total
Music Monday	#musicmonday 824.778 ocorrências 79,7% do total	#musicmondays 11.770 ocorrências 1,1% do total	#music 5.106 ocorrências 0,5% do total

Complementarmente, a Figura 5.6 associa a posição de uma hashtag em um ranking de popularidade, baseado no número de vezes em que uma tag foi utilizada, com o volume de tweets em que ela aparece. O gráfico foi plotado em coordenadas log-log, em que x é uma colocação no ranking de frequências e y é o número total de ocorrências da tag. Pode-se observar que a distribuição de hashtags também segue a tendência geral de uma distribuição zipfiana, aparecendo linearmente em coordenadas log-log - e, consequentemente, como um gráfico de cauda longa em uma plotagem realizada a partir dos dados brutos.

5.2 Análise de Fatores Condicionadores da Variação

Como mencionado anteriormente, o termo "variação", em linguística, é entendido como o fenômeno de duas ou mais formas diferentes ocorrerem, em um certo contexto linguístico, com o mesmo valor de verdade - ou seja, com o mesmo significado. Para que haja variação, portanto, as formas envolvidas devem necessariamente ser intercambiáveis em uma dada situação. O que define quando uma ou outra variante é usada

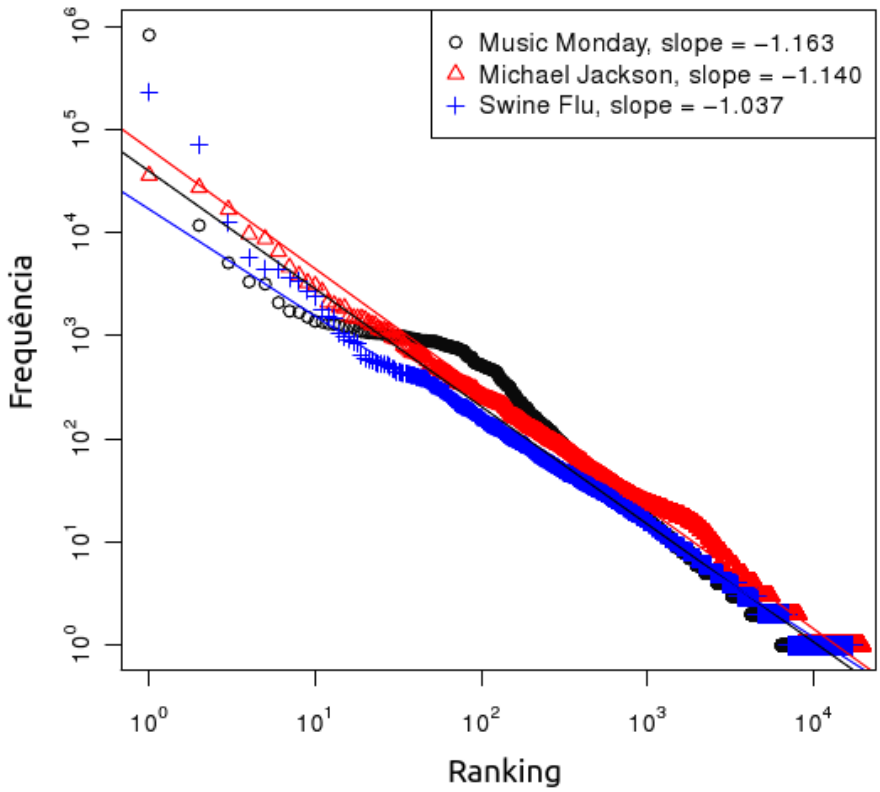

Figura 5.6. Ocorrências de hashtags *versus* suas posições em um ranking de popularidade

pelo falante são fatores linguísticos e sociais subjacentes. De acordo com Coelho et al. [2010], é a descrição desses fatores, sejam eles internos ou externos ao sistema linguístico, que permite ao linguista sugerir em que tipo de ambiente uma variante tem maior probabilidade de ser escolhida em detrimento de outra.

5.2.1 Fatores Internos

Os "fatores internos" ou "linguísticos", descritos por Labov [1995], são aqueles inerentes ao sistema linguístico, atuando, portanto, nos níveis sintático, fonológico, morfológico etc. Alguns exemplos de fatores internos são a ordem dos elementos de uma frase e a categoria das palavras, entre muitos outros.

5.2.1.1 Comprimento das hashtags

Em algumas línguas naturais, um dos fatores estritamente linguísticos capazes de alterar as dinâmicas de utilização de uma determinada forma parece ser o comprimento das palavras, como observado por Zipf [1935] e analisado por Sigurd et al. [2004]: aquele

sugere que o comprimento de uma palavra tende a manter uma relação inversa, porém não necessariamente proporcional, com a sua frequência relativa; estes analisam dados de diferentes gêneros textuais em inglês e sueco e corroboram a hipótese, demonstrando que as palavras mais longas tendem a ser evitadas provavelmente por irem contra o princípio da economia linguística[2] [Vicentini, 2003].

Diante dessas evidências e considerando a preocupação dos usuários do Twitter em economizar espaço, já que cada tweet tem um tamanho máximo de apenas 140 caracteres, foi investigado se o comprimento de uma hashtag é um dos fatores estritamente linguísticos que influenciam para o seu sucesso ou fracasso.

A fim de realizar essa análise, os comprimentos das hashtags mais populares de cada conjunto de dados foram comparados aos das menos populares. Uma análise qualitativa revela que as etiquetas mais comuns parecem simples, diretas e curtas; por outro lado, entre aquelas com pouca utilização, muitas são formadas por longas cadeias de caracteres. A Tabela 5.4 mostra informações preliminares sobre o comprimento das hashtags e a popularidade, indicando que tags formadas por quinze ou mais caracteres não estão presentes entre as mais comuns em nenhum dos conjuntos de dados analisados.

Tabela 5.4. Comparação entre as hashtags mais populares e as hashtags mais populares com 15 ou mais caracteres em cada uma das bases

Hashtags mais comuns (número de tweets)	Hashtags mais comuns com 15 ou mais caracteres (número de tweets)
#michaeljackson (35.861)	#nothingpersonal (962)
#michael (27.298)	#iwillneverforget (912)
#mj (16.758)	#thankyoumichael (690)
#swineflu (230.457)	#swinefluhatesyou (1.056)
#h1n1 (70.693)	#crapnamesforpubs (145)
#swine (12.444)	#superhappyfunflu (124)
#musicmonday (824.778)	#musicmondayhttp (540)
#musicmondays (11.770)	#fatpeoplearesexier (471)
#music (5.106)	#crapurbanlegends (23)

A Tabela 5.5 lista o comprimento médio, em número de caracteres, de diferentes grupos de hashtags divididas de acordo com suas posições no ranking ordenado por frequências em cada subconjunto de dados. As amostras das tags menos populares

[2]De acordo com Vicentini [2003], o conceito de economia linguística varia na literatura: para um formalista, diz respeito mais à organização do sistema, enquanto para um funcionalista, faz referência mais à estratégia comunicativa. No entanto, as ideias de simplicidade dos itens linguísticos e de menor esforço dos falantes permeia todas as definições do conceito.

foram formadas por cinquenta etiquetas selecionadas aleatoriamente entre aquelas que foram utilizadas em apenas um tweet do respectivo subconjunto.

Tabela 5.5. Comprimento médio das hashtags mais e menos populares acerca de cada um dos tópicos tratados

Tópico	Comprimento médio das...					
	...k hashtags mais populares					...hashtags
	k=10	k=20	k=30	k=40	k=50	menos populares
Michael Jackson	7,10	6,85	7,80	8,02	7,74	10,16
Gripe Suína	5,30	7,35	7,17	7,20	7,04	10,30
Music Monday	9,50	8,40	7,27	6,40	5,92	11,66

Em todos os subdatasets, o comprimento médio das hashtags mais populares é consideravelmente inferior àquele das menos populares. A Figura 5.7 compara os dados da Tabela 5.5, incluindo ainda informação sobre o desvio padrão. É evidente que as diferenças entre os comprimentos das etiquetas dos grupos mais populares não são relevantes, já que os comprimentos médios das k hashtags mais populares, com $k = \{10, 20, 30, 40, 50\}$, são mais ou menos semelhantes e não seguem um padrão fixo. No entanto, a comparação com as hashtags utilizadas apenas uma vez - as menos populares - mostra diferenças consideráveis que levam a acreditar que o comprimento de uma hashtag pode ser um fator interno - ou um fator de ordem estritamente linguística - que colabora para determinar o sucesso ou o fracasso de tags no Twitter.

A baixa popularidade das hashtags longas reflete o pequeno número de tags compostas por sentenças completas, como #mileycometobrazil, #herewegoagain e muitas outras, ocupando boas posições nos rankings de popularidade. Seu pequeno sucesso pode ser atribuído a algumas razões que vão além do seu comprimento, tais como: a) sentenças podem admitir um alto índice de variação graças às diferentes configurações possíveis de serem utilizadas durante as suas elaborações - por exemplo, #thankyoumichael, #thanksmj, #michaeljacksonthanks - o que reduz a frequência individual de cada uma das formas concorrentes; b) sentenças podem ser mais difíceis de se memorizar - e, consequentemente, de se reproduzir -, graças à possibilidade de se utilizar diferentes ordens de palavras para indicar o mesmo conteúdo. Um exemplo é a utilização das duas hashtags variantes #maiorqueissotudo e #maiorquetudoisso durante uma campanha publicitária no ano de 2010. Enquanto a estratégia publicitária era promover a primeira tag, muitos membros utilizaram a segunda devido ao fato de essa sentença aceitar um ordenamento diferente dos itens lexicais, o que certamente confundiu os usuários e levou cada hashtag variante a posições mais baixas no ranking de frequências; e (c) em sentenças, os usuários parecem estar mais propensos a erros or-

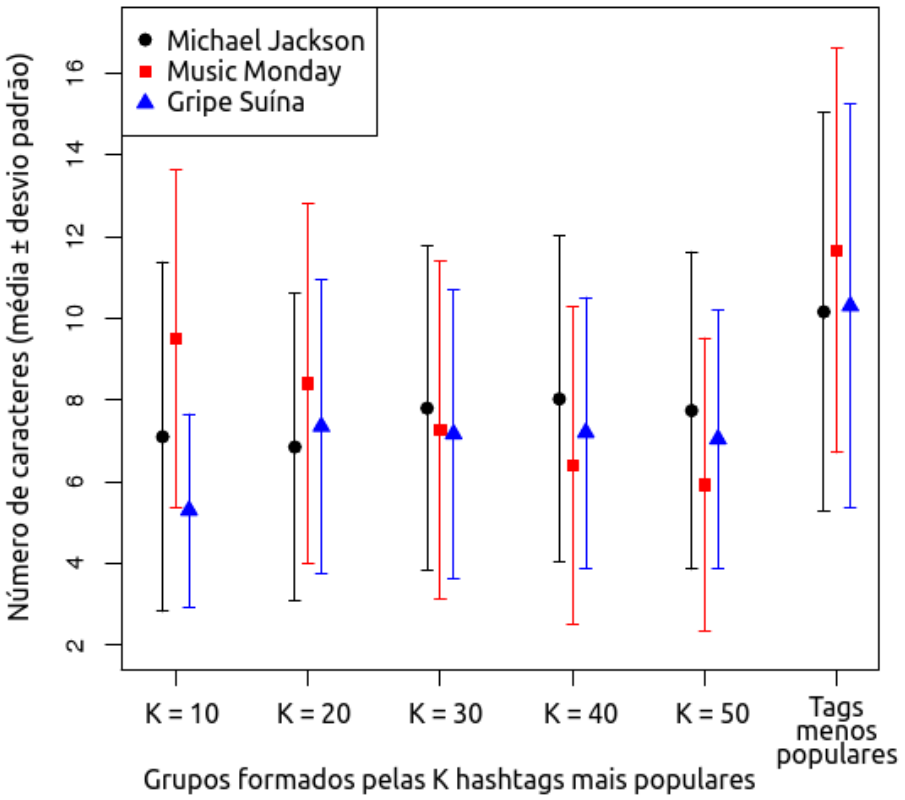

Figura 5.7. Número médio de caracteres das hashtags mais populares e de amostras selecionadas aleatoriamente entre as tags pouco populares (com apenas uma ocorrência)

tográficos, como em #thanktyoumichael, talvez pela maior dificuldade de se visualizar os termos sem os tradicionais espaços entre itens. Intuitivamente, parece ser mais fácil perceber o erro ortográfico em "thankt you michael" do que em "thanktyoumichael", embora esta seja apenas uma hipótese a ser verificada no futuro.

Entretanto, o fato de a limitação a 140 caracteres - um fator condicionante no comprimento das hashtags - ser de natureza diferente do que condicionaria o comprimento de uma forma linguística natural gera a necessidade de se realizar uma análise mais detalhada das idiossincrasias das hashtags, possivelmente levando-se em consideração o princípio da quantidade [Dirven & Verspoor, 2004], de base funcionalista, que relaciona a quantidade de informação ao tamanho das formas linguísticas.

5.2.1.2 Presença do elemento *underscore*

Verificou-se a influência do único sinal aceito na formação de hashtags além de letras e números: o *underscore* ou traço inferior (_).

Em todos os subconjuntos de dados, a utilização do sinal _ levou as hashtags

a ranqueamentos baixos: #michael_jackson alcançou a posição 248 na sua base, com 128 tweets; #swine_flu alcançou a posição 67 na sua base, com 246 tweets; #music_monday sequer foi utilizada. A Tabela 5.6 mostra a utilização do sinal _ nas hashtags analisadas.

Tabela 5.6. Distribuição das hashtags contendo o sinal *underscore* (_)

Tópico	Número de hashtags contendo _	% de hashtags contendo o sinal _ entre as hashtags usadas até i vezes	
		i=2	i=10
Michael Jackson	251 (1,2%)	89%	97%
Gripe Suína	155 (0,9%)	87%	97%
Music Monday	143 (0,9%)	89%	98%

Percebe-se que quase a totalidade das hashtags com o sinal _ encontra-se nas posições de ranqueamento inferiores: pelo menos 97% delas, em todos os subconjuntos, foram utilizadas em até 10 tweets, o que parece indicar que existe uma certa rejeição dos usuários às hashtags com esse sinal.

5.2.2 Fator Externo: Gênero dos Usuários

Além dos fatores estritamente linguísticos que influenciam a forma como o ser humano se expressa, há também fatores externos ao sistema linguístico, chamados ainda de "fatores sociais", que realizam tal influência, como Labov [2001] apresenta.

Cada palavra ou frase proferida por alguém conta uma história ao refletir características desse indivíduo e de seu grupo. As escolhas linguísticas são o resultado de uma série de interações sociais que compõem e formam, pouco a pouco, a língua dos indivíduos, de maneira tão sutil que dificilmente os próprios falantes são capazes de percebê-las e identificá-las. O modo como a língua é usada em diversas situações do cotidiano reflete, assim, o gênero, a idade, a naturalidade, o papel social, a posição hierárquica em uma organização, entre inúmeras outras características dos falantes. Compreender quais dessas características influenciam na maneira de utilização da língua e nas escolhas - inclusive lexicais - dos falantes é um dos objetivos da sociolinguística.

Diferenças de comportamento entre homens e mulheres têm sido estudadas em muitos campos do saber. Conhecê-las pode permitir uma melhor compreensão não apenas das características dos indivíduos, mas também de propriedades das comunidades das quais eles fazem parte e, especialmente, das dinâmicas sociais entre os dois gêneros. Afinal, identidades de gênero são social e culturalmente construídas [Weeks et al., 2003], de modo que, embora bases biológicas para legitimar diferenças existam,

muitas das percepções de gênero são produtos de relações sociais baseadas em processos históricos [Hacking, 1999].

Esta seção, amparada por estudos que indicam que homens e mulheres tendem a lidar com elementos linguísticos e com inovações de maneiras diferentes, investiga a conduta de usuários de ambos os gêneros no que tange a utilização de hashtags no Twitter.

Para executar as análises apresentadas aqui, foi necessário inicialmente definir o gênero dos membros da rede a partir das informações de perfil presentes nos datasets. Essa tarefa foi cumprida comparando-se os nomes próprios dos usuários disponíveis em seus perfis a listas de nomes masculinos e femininos, em português e inglês, disponíveis em sites de registros pessoais na Internet. Nomes considerados neutros - que aparecem em listas tanto masculinas quanto femininas - totalizam cerca de 0,04% do total de nomes surgidos nos datasets e foram ignorados nesta fase.

A Figura 5.8 mostra as análises temporais dos conjuntos de dados "Michael Jackson" e "Gripe Suína", respectivamente. Os gráficos indicam claramente diferentes dinâmicas no uso de hashtags pelos diferentes gêneros ao longo do tempo: enquanto algumas hashtags são muito populares entre os usuários de um determinado gênero, elas parecem não ter sucesso entre outros usuários. Analogamente, alguns picos de utilização de certas etiquetas estão presentes apenas entre usuários do sexo masculino ou feminino. Um aspecto interessante é que, nos períodos iniciais, isto é, logo após os acontecimentos sociais que desencadeiam a difusão das hashtags - que são, nesses casos, a morte de Michael Jackson e o início da epidemia de gripe suína -, os gráficos são similares, indicando uma influência do tempo sobre a aceitação e sobre a consequente difusão das hashtags, tal qual ocorre com formas linguísticas em comunidades de fala no mundo offline [Labov, 2001].

Nesta seção, a principal questão endereçada é se usuários do Twitter de diferentes gêneros escolhem as mesmas hashtags quando falam sobre o mesmo tema. Espera-se responder a essa questão e, no caso de ser encontrada a existência de formas neutras, masculinas e femininas, o objetivo será identificar alguns aspectos e características que distinguem esses três grupos de etiquetas.

Para cada hashtag, calculou-se o percentual de ocorrências geradas por usuários de cada gênero. No entanto, como o percentual total de ocorrências de hashtags usadas por homens e mulheres é diferente para cada conjunto de dados, os escores brutos (*raw scores*) entre os gêneros não são diretamente comparáveis. A fim de determinar se uma tag em particular é mais comum entre usuários de um dado gênero, os escores brutos foram convertidos para a mesma unidade de medida utilizando escores z (*z-scores*). Assim, nesta abordagem, o uso de escores z não funciona como um teste de

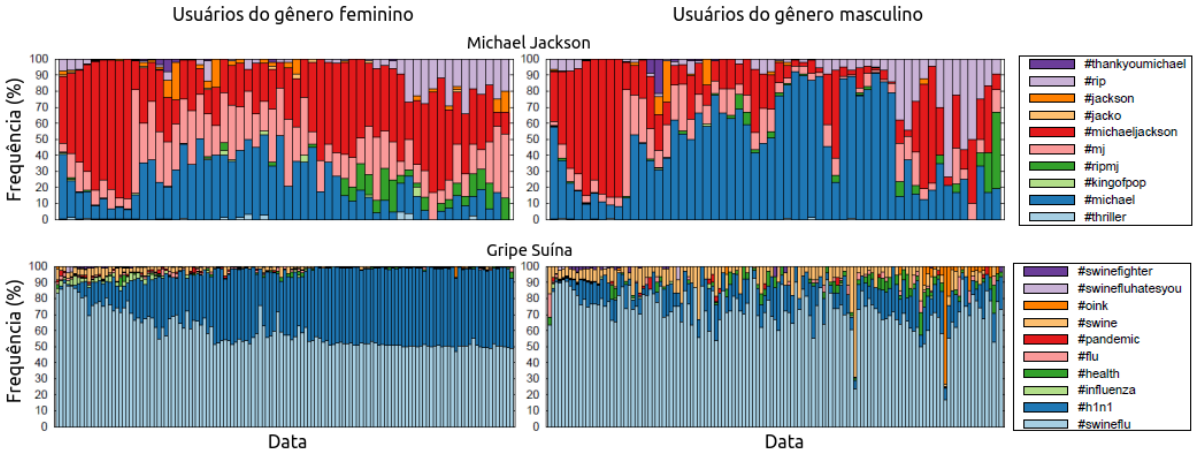

Figura 5.8. Percentual de uso das hashtags mais populares de cada tópico por usuários femininos e masculinos

significância estatística, mas como um fator de escala para que as comparações entre as ocorrências provenientes de usuários de cada gênero possam ser mensuradas a partir de um único parâmetro.

Escores z representam a distância, em função do número de unidades de desvio padrão, que o escore bruto está acima ou abaixo da média. Escores brutos acima da média geram escores z positivos; escores brutos abaixo da média, por sua vez, geram escores z negativos. Tais escores são calculados de acordo com

$$z = (x - \mu)/\sigma \tag{5.1}$$

em que x indica o escore bruto (percentual de ocorrências geradas por usuários femininos ou masculinos para cada hashtag), μ designa a média (percentual de ocorrências geradas por usuários femininos ou masculinos no total do subdataset) e σ simboliza o desvio padrão.

Dessa forma, para uma determinada hashtag, são obtidos dois escores z: o "escore z feminino", que determina o peso correspondente ao uso da etiqueta em questão por mulheres; e o "escore z masculino", que indica o peso correspondente à utilização daquela hashtag específica por homens. Os "escores z femininos" e os "escores z masculinos" são, para cada hashtag, complementares, de forma que a sua soma é sempre igual a zero. Por questão de conveniência, todos os escores z apresentados neste estudo serão relativos ao uso das hashtags por usuários do sexo feminino ("escores z femininos"). Portanto, neste trabalho, escores z positivos indicam uma prevalência de utilização por usuários do sexo feminino e escores z negativos sempre indicam uma prevalência de utilização por usuários do sexo masculino.

Para verificar a existência de formas mais prevalentes entre usuários de um determinado gênero, as hashtags pertencentes aos subdatasets "Michael Jackson", "Gripe Suína", "Eleições - apoiadores de Dilma", "Eleições - opositores de Dilma", "Eleições - apoiadores de Serra" e "Eleições - opositores de Serra" foram divididas em cinco grupos de acordo com os escores z obtidos, conforme apresentado na Tabela 5.7. Os valores dos escores z de referência (1 e 1,96) representam, respectivamente, os valores críticos para coeficientes de confiança de 68% e 95% em uma distribuição normal padrão.

Tabela 5.7. Grupos de hashtags de acordo com o escore z calculado

Hashtag fortemente feminina (HFF)	Hashtag feminina (HF)	Hashtag neutra (HN)	Hashtag masculina (HM)	Hashtag fortemente masculina (HFM)
$z > 1{,}96$	$1{,}96 \geq z > 1$	$1 \geq z \geq -1$	$-1 > z \geq -1{,}96$	$z < -1{,}96$

A Tabela 5.8 mostra que, em todos os conjuntos de dados, as hashtags neutras correspondem a grande parte do *corpus*, mas que também existe uma presença significativa de etiquetas mais associadas a algum dos gêneros. Nota-se ainda que as hashtags fortemente femininas ocorrem significativamente mais do que as fortemente masculinas, enquanto que as moderadamente masculinas são ligeiramente mais frequentes do que as moderadamente femininas.

Tabela 5.8. Presença de hashtags neutras e associadas a um gênero nos conjuntos de dados

Tópico	HFF	HF	HN	HM	HFM
Total	5,4%	10,8%	68,8%	14,0%	1,1%
Eleições - apoiadores de Dilma	0,0%	20,0%	70,0%	10,0%	0,0%
Eleições - apoiadores de Serra	0,0%	22,2%	55,6%	22,2%	0,0%
Eleições - opositores de Dilma	10,0%	0,0%	80,0%	10,0%	0,0%
Eleições - opositores de Serra	7,4%	7,4%	70,4%	14,8%	0,0%
Michael Jackson	6,7%	6,7%	66,7%	20,0%	0,0%
Gripe Suína	4,5%	13,6%	68,2%	9,1%	4,5%

O objetivo é encontrar aspectos que tornem as hashtags, ou grupos de hashtags, associadas a algum dos gêneros - ou seja, que as tornem "hashtags femininas" ou "hashtags masculinas". Foram analisadas, então, quatro diferentes categorias opostas de hashtags e foram observados os escores z médios obtidos para aquelas pertencentes a essas categorias.

5.2.2.1 Formas mais frequentes x Formas menos frequentes

No Capítulo 2, foram mencionados vários estudos clássicos demonstrando que, em geral, as mulheres ocidentais tendem a usar mais formas linguísticas padrão e não-estigmatizadas do que os homens, que comumente sentem-se mais confortáveis do que os falantes do sexo feminino para usar variantes não-padrão e muitas vezes linguisticamente inovadoras, pelo menos no início do processo de variação.

A definição do que é uma forma linguística padrão ou não-padrão não é trivial. No caso específico das hashtags, talvez seja impossível definir o estigma que cada forma carrega, caso carregue algum. Contudo, foi identificada uma diferença qualitativa, de certa forma relacionada a essa discussão, entre as hashtags mais frequentes e muitas daquelas menos frequentes: as primeiras tendem a ser mais transparentes acerca do tema a que se referem, enquanto as segundas, em diversas ocasiões, apresentam-se de maneira mais opaca. Por exemplo, as hashtags mais utilizadas nos conjuntos de dados "Michael Jackson" e "Gripe Suína" são justamente as bem transparentes #michaeljackson e #swineflu, mas muitas daquelas com frequência de utilização mais baixa são mais inovadoras e menos intuitivas (como #jacko e #swinefluhatesyou).

Nesta seção, buscou-se identificar se a oposição entre hashtags muito e pouco frequentes - e, como consequência indireta, entre hashtags mais transparentes e mais opacas com relação aos seus referentes - possa ser um fator que afete a sua aceitação por usuários de determinado gênero.

Para cada conjunto de dados, foram calculados os escores z médios das hashtags 20% mais e menos comuns. Os resultados estão mostrados na Tabela 5.9.

Tabela 5.9. Escores z médios das hashtags mais e menos frequentes

Tópico	Escores z	
	Formas mais frequentes	Formas menos frequentes
Eleições - apoiadores de Dilma	0,974	-0,145
Eleições - apoiadores de Serra	0,450	-0,215
Eleições - opositores de Dilma	1,024	-1,512
Eleições - opositores de Serra	0,885	0,031
Michael Jackson	1,467	-0,024
Gripe Suína	0,002	0,079

Descobriu-se que, em todos os conjuntos de dados, os usuários do sexo feminino são mais propensos a utilizar as hashtags mais populares do que aqueles do sexo masculino. Com exceção do subdataset "Gripe Suína", as mulheres também usam mais frequentemente as formas mais comuns do que as formas menos comuns, em princípio

mais inovadoras.

5.2.2.2 Envolvimento pessoal x Persuasão clara

Como descreveu-se na subseção anterior, foram notadas algumas diferenças durante o processo de designação de hashtags por homens e mulheres. Verificou-se basicamente que usuários femininos tendem a se sentir mais confortáveis ao designar etiquetas mais frequentes, e geralmente mais transparentes, aos seus tweets, enquanto usuários masculinos costumam ser os principais usuários das tags menos utilizadas e, muitas vezes, mais opacas.

Porém, o fato mais interessante encontrado ao se analisar as diferenças de gênero na escolha de hashtags para mensagens postadas no Twitter diz respeito às estratégias discursivas adotadas por homens e mulheres na rede. A análise da formação linguística das hashtags que fazem referência a algum tipo de apoio a um dos candidatos das eleições brasileiras de 2010 sugere a existência de uma diferença em como usuários homens e mulheres expressam as suas preferências e buscam convencer os seus seguidores no campo político.

Ao analisar os subconjuntos das etiquetas de apoiadores dos candidatos, pode-se distinguir claramente algumas dessas hashtags entre duas categorias: (1) aquelas em que os usuários buscam informar a própria opção pessoal por um determinado candidato; e (2) aquelas em que os usuários focam em sugerir, de maneira imperativa, um candidato para os seus seguidores. No grupo 1, foram incluídas hashtags contendo verbos conjugados na primeira pessoa do singular do modo indicativo, como #votodilma/#votoserra e #euquerodilma/#euqueroserra. No segundo grupo, por outro lado, foram incluídas tags contendo verbos conjugados na segunda pessoa do singular do modo imperativo, como se os usuários estivessem expressando comandos para que os seus seguidores ajam de uma determinada maneira, tal qual em #vote13/#vote45 e #sejamais1dilma. Nesta seção, não foram analisadas as hashtags indicando oposição a algum dos candidatos, já que, entre elas, aquelas que abertamente buscam persuadir os leitores a não votarem em algum dos candidatos usando uma das estratégias linguísticas descritas acima pouco aparecem no dataset.

Essas duas diferentes estratégias discursivas, embora possuam o mesmo objetivo - tentar convencer os leitores a votarem em um candidato específico -, pretendem atingi-lo de maneiras indubitavelmente diferentes. O uso da primeira pessoa do modo indicativo sugere uma conexão mais íntima entre o autor e o leitor, como se o primeiro dissesse "Eu votarei no candidato X, por que você também não faz isso?". Já a utilização de formas imperativas indica que o autor situa-se em uma posição hierárquica superior,

como se ele tivesse algum tipo de poder sobre o leitor. Naturalmente, essas relações - conexão íntima entre usuários e poder do autor sobre o leitor - não são necessariamente reais: elas podem ser simples reflexos dos papéis esperados de serem desempenhados por certos indivíduos em situações offline.

O cálculo dos escores z médios dos grupos de hashtags associados aos gêneros mostrou diferenças significativas no comportamento de homens e mulheres que buscaram persuadir os seus seguidores. As formas pertencentes ao primeiro grupo, que traz etiquetas com verbos na primeira pessoa do singular, tendem a ser mais comuns entre mulheres. Contudo, as hashtags imperativas do segundo grupo são mais frequentes entre usuários do gênero masculino. A Tabela 5.10 e a Figura 5.9 ilustram essas diferenças.

Tabela 5.10. Média dos escores z femininos do grupo 1 ("tags pessoais") e do grupo 2 ("tags imperativas")

Tópico	Média dos escores z femininos	
	Grupo 1: tags pessoais (1a. pessoa do singular, modo indicativo)	Grupo 2: tags imperativas (2a. pessoa do singular, modo imperativo)
Eleições - apoiadores de Dilma	0,601	-1,894
Eleições - apoiadores de Serra	1,477	-0,957

Tais resultados não são inteiramente inesperados, já que estudos anteriores nos campos de psicologia, antropologia, comunicação e análise do discurso mostraram diferenças em como homens e mulheres ocidentais buscam convencer os demais e são persuadidos por eles [Brunel & Nelson, 2003; Chung & Trivedi, 2003], inclusive em ambientes mediados por computador [Guadagno & Cialdini, 2002]. Estudos anteriores da área de comunicação indicam que os homens ocidentais são mais confiantes em relação a sua capacidade de persuadir [Andrews, 1987], o que pode ser um motivo para deixá-los mais à vontade para usar estratégias de convencimento diretas e claras, inclusive no Twitter. Outros estudos também sugerem que as gerentes do sexo feminino, ao tentar convencer os subordinados, confiam mais frequentemente no altruísmo do que os gerentes do sexo masculino [Harper & Hirokawa, 1988]. Considerando que a estratégia de "envolvimento pessoal", utilizada nas tags do grupo 1, reduz a distância entre o autor e o leitor, pode-se também sugerir que essa estratégia esteja relacionada ao comportamento altruísta de mulheres gerentes.

Outros estudos indicaram que, dadas algumas condições, mulheres ocidentais são mais facilmente influenciadas e menos influentes do que homens [Eagly, 1978], o que leva a questões como "que tipos de comportamentos as pessoas usam quando

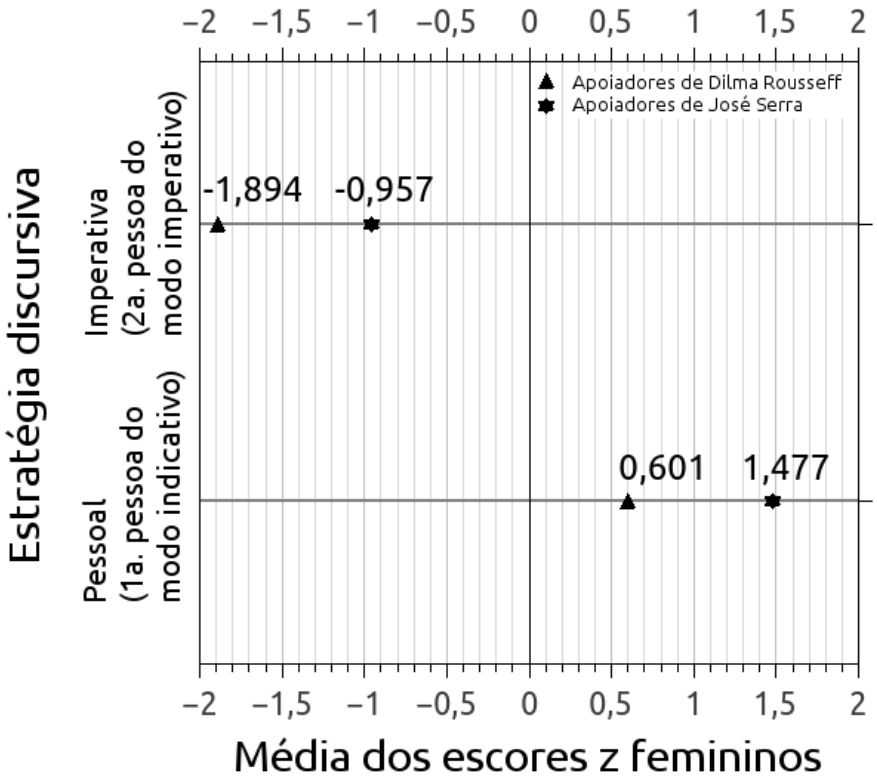

Figura 5.9. Média dos escores z femininos do grupo 1 ("tags pessoais") e do grupo 2 ("tags imperativas")

tentam influenciar homens ou mulheres?" [Carli, 1978]. Neste estudo, foi identificado um desses comportamentos em uma situação até então inexplorada: embora esses resultados não sejam novos no estudo do processo de comunicação humana, esta é a primeira vez em que eles foram observados no domínio da comunicação virtual e relacionados ao uso de etiquetas em um ambiente de etiquetagem completamente livre. Porém, é sempre importante deixar claro que comportamentos diferentes com relação a estratégias persuasivas não são diretamente ligados a sexo, mas a poder e status, de forma que as diferenças de gênero no comportamento devem ser compreendidas dentro de um contexto mais amplo de relações sociais [Sagrestano, 1992].

Esses resultados também podem ser analisados a partir de uma perspectiva política. O valor negativo mais acentuado para os escores z femininos médios entre as hashtags imperativas e o valor positivo mais discreto entre as hashtags pessoais postadas por apoiadores de Dilma Rousseff indicam que seus eleitores são mais propensos a usar as etiquetas imperativas - ou, equivalentemente, menos propensos a usar aquelas mais pessoais - do que os apoiadores de José Serra, que preferem, em geral, estratégias discursivas mais íntimas.

Capítulo 6

Conclusões e Trabalhos Futuros

Esta dissertação analisa, em alguns momentos por meio de uma abordagem linguística, questões relativas à formação e ao uso de hashtags no Twitter, merecendo crédito por propor uma análise que busca associar conhecimento de duas áreas distintas e por adotar uma base empírica robusta para a análise de fenômenos. Foi proposto que a teoria sociolinguística pudesse ser usada para formular hipóteses sobre a linguagem em sistemas online como o Twitter e os resultados mostraram semelhanças não apenas qualitativas, mas também quantitativas, entre comunidades de fala offline e online. Concedeu-se especial atenção a questões estruturais e sociais na análise linguística, mostrando, assim, comprometimento com uma visão socio-histórica da linguagem humana.

Verificaram-se as motivações encontradas pelos usuários do Twitter para etiquetar as suas mensagens e concluiu-se que as principais razões que levam os membros dessa rede de informação a inserir hashtags em seus tweets são os aumentos da compreensibilidade da mensagem e da possibilidade de disseminação do conteúdo postado.

Foram revelados aspectos interessantes sobre a distribuição de hashtags com relação às suas popularidades, associando-as à distribuição de palavras em rankings de frequência. Foram também estudados fatores linguísticos de natureza formal que distinguem hashtags que se disseminam muito de outras que não conseguem atrair a atenção dos usuários: o comprimento da tag, por exemplo, é um desses fatores.

Foi ainda apresentada a análise inovadora da influência de um fator social no processo de designação de etiquetas: o gênero. O objetivo principal dessa seção do trabalho é verificar se e como o comportamento de usuários do sexo masculino e feminino difere no uso desses elementos, tal qual ocorre com outros elementos linguísticos. A motivação para enfrentar esse problema surge a partir da necessidade de se caracterizar as preferências coletivas dos usuários, a fim de compreender as dinâmicas sociais

entre homens e mulheres em comunidades online e contribuir para o desenvolvimento de serviços mais personalizados na Web. Foram fornecidas evidências de que, embora a maioria das hashtags pareçam ser neutras, algumas deles são, em certa medida, mais associadas a um dos gêneros. Analisaram-se também diferentes categorias de hashtags e descobriu-se que certos papéis sociais ocupados por cada um dos sexos nas comunidades offline são igualmente desempenhados em redes sociais online. Esses resultados são interessantes pois podem ser correlacionados com aqueles obtidos por estudos nos campos da sociolinguística, da psicologia e das ciências sociais.

Trabalhos futuros deverão investigar outros fatores que poderiam atuar como condicionadores linguísticos e sociais capazes de influenciar como os usuários empregam certas hashtags. Afinal, conhecer a dinâmica de etiquetagem dos usuários e as características das hashtags de sucesso é útil não apenas para um estudo do comportamento dos usuários em redes, como também para a otimização de sistemas de recomendação de tags em diversos ambientes.

Referências Bibliográficas

Abdel-Jawad, H. (1987). Cross-dialectal variation in arabic: Competing prestigious forms. *Language and Society*, 16:359–367.

Ames, M. & Naaman, M. (2007). Why we tag: Motivations for annotation in mobile and online media. Em *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*.

Andrews, P. H. (1987). Gender differences in persuasive communication and attribution of success and failure. *Human Communication Research*, 13(3):372–385.

Bailey, C. J. (1973). *Variation and Linguistic Theory*. Center for Applied Linguistics, Washington DC.

Bakir, M. (1986). Sex differences in the approximation to standard arabic: a case study. *Anthropological Linguistics*, 28(11):3–10.

Benevenuto, F. (2010). *An Empirical Analysis of Interactions in Online Social Networks*. Tese de doutorado, Universidade Federal de Minas Gerais.

Benevenuto, F.; Magno, G.; Rodrigues, T. & Almeida, V. (2010). Detecting spammers on twitter. Em *Proceedings of the 7th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference (CEAS)*.

Berners-Lee, T.; Hall, W.; Hendler, J. A.; O'Hara, K.; Shadbolt, N. & Weitzner, D. J. (2006). A framework for web science. *Foundations and Trends in Web Science*, 1(1):1–130.

Bigonha, C.; Cardoso, T. N.; Moro, M. M.; Almeida, V. & Gonçalves, M. A. (2010). Detecting evangelists and detractors on twitter. Em *Anais do Simpósio Brasileiro de Sistemas Multimídia e Web - Webmedia*, pp. 107–114.

Bimber, B. (2000). Measuring the gender gap on the internet. *Social Science Quarterly*, 81:868–876.

Boyd, D. (2009). Twitter: Pointless babble or peripheral awareness plus social grooming? Disponível em http://www.zephoria.org/thoughts/archives/2009/08/16/twitter_pointle.html.

Brandt, M. B. (2009). Etiquetagem e folksonomia: Uma análise sob a Óptica dos processos de organização e recuperação da informação na web. Dissertação de mestrado, Universidade de Brasília.

Brunel, F. & Nelson, M. (2003). Message order effects and gender differences in advertising persuasion. *Journal of Advertising Research*, 43:330–341.

Bruns, A. & Burgess, J. (2011). The use of twitter hashtags in the formation of ad hoc publics. Em *Proceedings of the European Consortium for Political Research conference.*

Cambraia, C.; Cunha, E.; Bezerra, V. & Ramalho, V. (2008). Variação, mudança e estilística: Demonstrativos. Em Lima-Hernandes, M. C., editor, *A Língua Portuguesa no Mundo*. Faculdade de Filosofia de Ciência Humanas da Universidade de São Paulo, São Paulo, Brasil.

Cameron, M. (2011). The history of tagging: It's what you make of it that counts. Disponível em http://thehistoryof.net/the-history-of-tagging.html.

Carli, L. (1978). Gender differences in interaction style and influence. *Journal of Personality and Social Psychology*, 85:86–116.

Carter, S.; Tsagkias, M. & Weerkamp, W. (2011). Twitter hashtags: Joint translation and clustering. *Human Factors*, pp. 1–3.

Cha, M.; Haddadi, H.; Benevenuto, F. & Gummadi, K. P. (2010). Measuring user influence in twitter: The million follower fallacy. Em *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM).*

Cheshire, J. (2001). Sex and gender in variationist research. Em Chambers, J.; Trudgill, P. & Schilling-Estes, N., editores, *The Handbook of Language Variation and Change*. British Library, Oxford, UK.

Chew, C. & Eysenbach, G. (2010). Pandemics in the age of twitter: Content analysis of tweets during the 2009 h1n1 outbreak. *PLoS ONE*, 5(11).

Chung, J. & Trivedi, V. (2003). The effect of friendly persuasion and gender on tax compliance behavior. *Journal of Business Ethics*, 47:133–145.

Coelho, I.; Gorski, E.; May, G. & Souza, C. (2010). *Sociolinguística*. LLV/CCE/UFSC, Florianópolis, SC, Brasil.

Comarela, G.; Crovella, M.; Almeida, V. & Benevenuto, F. (2012). Understanding factors that affect response rates in twitter. Em *Proceedings of the 23rd ACM Conference on Hypertext and Social Media (HT)*.

Cunha, E.; Magno, G.; Almeida, V.; Gonçalves, M. A. & Benevenuto, F. (2012). A gender based study of tagging behavior in twitter. Em *Proceedings of the 23rd ACM Conference on Hypertext and Social Media (HT)*.

Cunha, E.; Magno, G.; Comarela, G.; Almeida, V.; Gonçalves, M. & Benevenuto, F. (2011). Analyzing the dynamic evolution of hashtags on twitter: a language-based approach. Em *Proceedings of the Workshop on Language in Social Media (LSM)*.

Danescu-Niculescu-Mizil, C.; Lee, L.; Pang, B. & Kleinberg, J. (2012). Echoes of power: Language effects and power differences in social interaction. Em *Proceedings of the International World Wide Web Conference (WWW)*.

Davidov, D.; Tsur, O. & Rappoport, A. (2010). Enhanced sentiment learning using twitter hashtags and smileys. *Universiteit van Amsterdam*, 2:241–249.

Dirven, R. & Verspoor, M. (2004). *Cognitive Exploration of Language and Linguistics*. John Benjamins Publishing, Philadelphia, PA.

Eagly, A. (1978). Sex differences in influenceability. *Psychological Bulletin*, 85:86–116.

Easley, D. & Kleinberg, J. (2010). *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press, Cambridge, MA.

Eleta, I. & Golbeck, J. (2012). A study of multilingual social tagging of art images: Cultural bridges and diversity. Em *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, pp. 695–704, New York, NY, USA. ACM.

Fallows, D. (2005). How men and women use the internet. Em *The Pew Internet and American Life Project*. Pew Research Center, Washington, DC.

Fischer, J. (1958). Social influences on the choice of a linguistic variant. *Word*, 14:47–56.

Fischer, O. (2007). *Morphosyntactic Change: Functional and Formal Perspectives*. Oxford University Press, Oxford.

Gambhir, S. (1981). *The East Indian Speech Community in Guyana: a Sociolinguistic Study with Special Reference to Koine Formation.* Tese de doutorado, University of Pennsylvania.

Gao, Q.; Dai, Y. & Fu, K. (2009). Improving personal tagging consistency through visualization of tag relevancy. Em *Proceedings of the 3d International Conference on Online Communities and Social Computing (OCSC*, pp. 326–335.

Garrett, J. J. (2005). An interview with flickr's eric costello. Disponível em http://www.adaptivepath.com/ideas/e000519.

Golder, S. & Huberman, B. (2006). Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208.

Golder, S. A. & Macy, M. W. (2011). Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 333(6051).

Gomide, J. S. (2012). Mineração de redes sociais para detecção e previsão de eventos reais. Dissertação de mestrado, Universidade Federal de Minas Gerais.

Gonçalves, C. A. V. (1993). Aférese e prótese: Verso e reverso morfológico. Dissertação de mestrado, Universidade Federal do Rio de Janeiro.

Guadagno, R. & Cialdini, R. (2002). Online persuasion: An examination of gender differences in computer-mediated interpersonal influence. *Group Dynamics: Theory, Research, and Practice*, 6:38–51.

Gupta, A. & Kumaraguru, P. (2012). Credibility ranking of tweets during high impact events. Em *Proceedings of the Workshop on Privacy and Security in Online Social Media (PSOSM).*

Guy, G. (1981). *Linguistic Variation in Brazilian Portuguese: Aspects of the Phonology, Syntax and Language History.* Tese de doutorado, University of Pennsylvania.

Hacking, I. (1999). *The Social Construction of What?* Harvard University Press, Cambridge, MA.

Haeri, N. (1987). Male/female differences in speech: an alternative interpretation. Em Dennig, K.; Inkelas, S.; McNair-Knox, F. & Rickford, J., editores, *Variation in Language*, pp. 173–182. Stanford University.

Hahn, L. (2004). *Padrões de Migração de Peixes no Alto Rio Uruguai e Capacidade de Transposição de Obstáculos.* Tese de doutorado, Universidade Estadual de Maringá.

Harper, N. L. & Hirokawa, R. Y. (1988). A comparison of persuasive strategies used by female and male managers: An examination of downward influence. *Communication Quarterly*, 36(2):157–168.

Hibiya, J. (1988). *A Quantitative Study of Tokyo Japanese*. Tese de doutorado, University of Pennsylvania.

Hong, L.; Convertino, G. & Chi, E. H. (2011). Language matters in twitter: A large scale study characterizing the top languages in twitter characterizing differences across languages including urls and hashtags. *Artificial Intelligence*, 91(1):518–521.

Horvath, B. (1985). *Variation in Australian English*. Cambridge University Press, Cambridge, UK.

Huang, C. (2011). Facebook and twitter key to arab spring uprisings: Report. Disponível em http://www.thenational.ae/news/uae-news/facebook-and-twitter-key-to-arab-spring-uprisings-report.

ICMNews (2009). Google and twitter crash at news of jackson's death. Disponível em http://news.icm.ac.uk/technology/google-twitter-crash-at-news-of-jackson80

Iofciu, T.; Fankhauser, P.; Abel, F. & Bischoff, K. (2011). Identifying users across social tagging systems. Em *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM)*, pp. 522–525.

Ivanov, I.; Vajda, P.; Lee, J.-S. & Ebrahimi, T. (2012). In tags we trust: Trust modeling in social tagging of multimedia content. *Signal Processing Magazine, IEEE*, 29(2):98–107.

Jain, D. (1973). *Pronominal Usage in Hindi: a Sociolinguistics Study*. Tese de doutorado, University of Pennsylvania.

Java, A.; Song, X.; Finin, T. & Tseng, B. (2007). Why we twitter: Understanding microblogging usage and communities. Em *Proceedings of the 13th International Conference on Knowledge Discovery and Data Mining (KDD)*.

Johnson, S. (2009). How twitter will change the way we live. Disponível em http://www.time.com/time/magazine/article/0,9171,1902818,00.html.

Keller, P. (2007). Tag history and gartners hype cycles. Disponível em http://www.pui.ch/phred/archives/2007/05/tag-history-and-gartners-hype-cycles.html.

Kelly, R. (2009). Twitter study. Disponível em http://www.pearanalytics.com/blog/wp-content/uploads/2010/05/Twitter-Study-August-2009.pdf.

Krishnamurthy, B.; Gill, P. & Arlitt, M. (2008). A few chirps about twitter. Em *Proceedings of the 1st Workshop on Online Social Networks*.

Kroch, A. (1989). Reflexes of grammar in patterns of language change. *Language Variation and Change*, 1:199–244.

Kwak, H.; Lee, C.; Park, H. & Moon, S. (2010). What is twitter, a social network or a news media? Em *Proceedings of the International World Wide Web Conference (WWW)*.

Laberge, S. (1977). *Étude de la Variation des Pronoms Définis et Indéfinis dans le Français Parlé à Montreal*. Tese de doutorado, Université de Montreal.

Labov, W. (1972). *Sociolinguistic Patterns*. University of Philadelphia Press, Philadelphia, USA.

Labov, W. (1995). *Principles of Linguistic Change: Internal Factors*. Blackwell, Oxford/Cambridge.

Labov, W. (2001). *Principles of Linguistic Change: Social Factors*. Blackwell, Malden, MA.

Lehmann, M. (2010). Participatory journalism: Risks and opportunities for newspaper companies to grow with user-generated content. *Detecon International GmbH*, pp. 1–20.

Lyons, J. (1970). *New Horizons in Linguistics*. Penguin, Harmondsworth.

Macaulay, R. (1977). *Language, Social Class and Education: a Glasgow Study*. Edinburgh University Press, Edinburgh.

Malinowski, B. (1923). The problem of meaning in primitive languages. Em Ogden, C. K. & Richards, I. A., editores, *The Meaning of Meaning*, pp. 146–152. Routledge and Kegan Paul, London.

Mathes, A. (2004). Folksonomies - cooperative classification and communication through shared metadata. Em *Computer Mediated Communication, Graduate School of Library and Information Science, University of Illinois Urbana-Champaign*.

Mejias, U. (2004). Bookmark, classify and share: A mini-ethnography of social practices in a distributed classification community. Disponível em http://blog.ulisesmejias.com/2004/12/27/a-delicious-study/.

Messina, C. (2007). Groups for twitter; or a proposal for twitter tag channels. Disponível em http://factoryjoe.com/blog/2007/08/25/groups-for-twitter-or-a-proposal-for-twitter-tag-channels/.

Mistry, O. & Sen, S. (2012). Probabilistic approaches to tag recommendation in a social bookmarking network. Em Desai, N.; Liu, A. & Winikoff, M., editores, *Principles and Practice of Multi-Agent Systems*, volume 7057 of *Lecture Notes in Computer Science*, pp. 270–287. Springer Berlin / Heidelberg. 10.1007/978-3-642-25920-3_19.

Modaressi, Y. (1978). *A Sociolinguistic Analysis of Modern Persian*. Tese de doutorado, University of Kansas.

Nettle, D. (1999). Using social impact theory to simulate language change. *Lingua*, 108:95–117.

Ngefac, A. (2008). *Social Differentiation in Cameroon English: Evidence from Sociolinguistic Fieldwork*. Peter Lang Publishing, New York, NY.

Nov, O.; Naaman, M. & Ye, C. (2008). What drives content tagging: The case of photos on flickr. Em *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*.

O'Connor, B.; Balasubramanyan, R.; Routledge, B. R. & Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. Em *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Oliveira, N. & Holanda, A. F. C. (2010). Jornalismo participativo e informação hiperlocal: O papel de mashups e hashtags na construção da notícia em redes sociais. *Revista Brasileira de Iniciação Científica em Comunicação Social*, 2(1):1–17.

Ono, H. & Zavodny, M. (2003). Gender and the internet. *Social Science Quarterly*, 84:111–121.

Oreilly, T. (2007). What is web 2.0: Design patterns and business models for the next generation of software. *Communications and Strategies*, (65).

Papacharissi, Z. & Oliveira, M. (2011). The rhythms of news storytelling on twitter: Coverage of the january 25th egyptian uprising on twitter. Em *Proceedings of the World Association for Public Opinion Research Conference*.

Parry, R. & Ortiz-Williams, M. (2007). How shall we label our exhibit today? applying the principles of on-line publishing to an on-site exhibition. Em *Proceedings of the International Conference for Culture and Heritage On-line.*

Poschko, J. (2010). Exploring twitter hashtags. Disponível em http://www.kdnuggets.com/2010/12/exploring-twitter-hashtags.html.

Rodrigues, T.; Benevenuto, F.; Cha, M.; Gummadi, K. P. & Almeida, V. (2011). On word-of-mouth based discovery of the web. Em *Proceedings of the International Measurement Conference (IMC).*

Romero, D.; Meeder, B. & Kleinberg, J. (2011). Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter. Em *Proceedings of the International World Wide Web Conference (WWW).*

Ross, M.; Mansson, S. & Daneback, K. (2011). Prevalence, severity, and correlates of problematic sexual internet use in swedish men and women. *Archives of Sexual Behavior.*

Rus, M. L. (2008). Lexical innovation. *Onomastica Lexicologie.*

Sagrestano, L. (1992). Power strategies in interpersonal relationships: The effects of expertise and gender. *Psychology of Women Quarterly*, 16:481–495.

Santo, A. (2011). Occupy wall street's media team: A day in the life. Disponível em http://www.cjr.org/the_news_frontier/occupy_wall_streets_media_team.php.

Santos, W.; Pappa, G.; Meira, W.; Guedes, D.; Veloso, A.; Almeida, V.; Pereira, A.; Guerra, P.; Silva, A.; Mourão, F.; Magalhães, T.; Machado, F.; Cherchiglia, L.; Simões, L.; Batista, R.; Arcanjo, F.; Brunoro, G.; Mariano, N.; Magno, G.; Ribeiro, M.; Teixeira, L.; Silva, A.; Reis, B. & Silva, R. (2010). Observatório da web: Uma plataforma de monitoração, síntese e visualização de eventos massivos em tempo real. Em *Anais do XXXVII Seminário Integrado de Hardware e Software (SEMISH)*, pp. 110–120.

Sapir, E. (1921). *Language: An Introduction to the Study of Speech.* Harcourt, Brace and World, New York, NY, USA.

Sawyer, S. & Rosenbaum, H. (2000). Social informatics in the information sciences: Current activities and emerging directions. *Informing Science*, 3(2):89–95.

Sensorpro.net (2012). Survey guidelines. Disponível em http://www.sensorpro.net/SurveyGuidelines.pdf.

Sigurd, B.; Eeg-Olofsson, M. & de Weijer, J. V. (2004). World length, sentence length and frequency - zipf revisited. *Studia Linguistica*, 58(1):37–52.

Silva, L. G. (2006). A dimensão sociolingüística do atlas lingüístico do brasil. Em *Anais da VIII Semana de Letras da Universidade Federal de Ouro Preto*. Universidade Federal de Ouro Preto, Ouro Preto.

Sinha, R. (2005). A cognitive analysis of tagging. Disponível em http://rashmisinha.com/2005/09/27/a-cognitive-analysis-of-tagging/.

Smith, J. E. (2011). The pageantry of tagging. Disponível em http://www.practicallyefficient.com/2011/08/05/tagging/.

Soares, A. & Peixoto, T. (2010). Variação linguística em textos de homens e mulheres em fóruns de comunidades do orkut. Em *III Simpósio Hipertexto e Tecnologias na Educação: Redes Sociais e Aprendizagem*.

Tannen, D. (1990). *You Just Don't Understand: Women and Men in Conversation*. William Morrow, New York, NY.

Thelwall, M. (2011). Privacy and gender in the social web. Em Trepte, S. & Reinecke, L., editores, *Privacy Online: Perspectives on Privacy and Self-Disclosure in the Social Web*. Springer, New York, NY, USA.

Troutman, C.; Clark, B. & Goldrick, M. (2008). Social networks and intraspeaker variation during periods of language change. Em *Proceedings of the 31st Annual Penn Linguistics Colloquium*, pp. 325–338. University of Pennsylvania, Philadelphia.

Trudgill, P. (1974). *The Social Differentiation of English in Norwich*. Cambridge University Press, Cambridge, UK.

Tsur, O. & Rappoport, A. (2012). What's in a hashtag? content based prediction of the spread of ideas in microblogging communities. Em *Proceedings of the International Conference on Web Search and Data Mining (WSDM)*.

Twitter (2012). Twitter turns six. Disponível em http://blog.twitter.com/2012/03/twitter-turns-six.html.

Vera, A. M. (2011). Propriedades de redes complexas de telecomunicações. Dissertação de mestrado, Universidade de São Paulo.

Vicentini, A. (2003). The economy principle in language: Notes and observations from early modern english grammars. *Mots Palabras Words*, 3.

Wagner, C. & Strohmaier, M. (2010). The wisdom in tweetonomies: Acquiring latent conceptual structures from social awareness streams. Em *Proceedings of the Semantic Search Workshop (SemSearch), in conjunction with the 19th International World Wide Web Conference (WWW)*.

Wal, T. V. (2005). Explaining and showing broad and narrow folksonomies. Disponível em http://www.vanderwal.net/random/entrysel.php?blog=1635.

Wal, T. V. (2007a). Folksonomy. Disponível em http://vanderwal.net/folksonomy.html.

Wal, T. V. (2007b). A stale state of tagging? Disponível em http://vanderwal.net/random/entrysel.php?blog=1945.

Weeks, J.; Holland, J. & Waites, M. (2003). *Sexualities and Society: A Reader*. Polity Press, Cambridge, UK.

Weinreich, U.; Labov, W. & Herzog, M. (1968). Empirical foundations for a theory of language change.

Weller, S. C. (2007). Cultural consensus theory: Applications and frequently asked questions. *Field Methods*, 19(4):339–368.

Weng, J.; Lim, E.-P.; He, Q. & Leung, C. W.-K. (2010). What do people want in microblogs? measuring interestingness of hashtags in twitter. Em *Proceedings of the IEEE International Conference on Data Mining (ICDM*, pp. 1121–1126.

Zhao, D. & Rosson, M. B. (2009). How and why people twitter: The role that micro-blogging plays in informal communication at work. Em *Proceedings of the ACM International Conference on Supporting Group Work (GROUP)*.

Zipf, G. K. (1935). *The Psycho-Biology of Language*. MIT Press, Cambridge.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Cambridge, MA.

Zollers, A. (2007). Emerging motivations for tagging: Expression, performance, and activism. Em *Proceedings of the International World Wide Web Conference (WWW)*.