5
Descrição das regras e aplicação nos dados

Conforme descrito no capítulo anterior, reunimos dados de diferentes ambientes on-line: chats, Orkut e Twitter. Veremos, a seguir, o estudo feito com a intenção de categorizar as palavras que compõem o chamado internetês.

5.1
As categorias do internetês

Como base de nossa análise do internetês, tomamos o estudo de Fusca (2008), que divide as abreviaturas nele contidas em quatro categorias:

1) Abreviaturas que mantêm o primeiro grafema de cada sílaba: *tc* – teclar; *vc* – você; *q* – que; *d* – de; *kd* – cadê.

A autora inclui nesta categoria abreviaturas que tenham sido pluralizadas (*vcs* – vocês), que indiquem marca de nasalidade (*tbm* – também) e todas as que contêm letras que, oralizadas, indicam a sílaba estendida, *kd* – cadê, em que tanto a letra *k* quanto a letra *d* representam exatamente as sílabas completas *ca* e *dê*.

2) Abreviaturas que indicam modo de enunciação oral / falado: *taum* – estão; *long* – longe; *ond* – onde; *bele* – beleza

Segundo Fusca (2008), esta categoria representa as realizações tidas como pertencentes ao modo ao modo de enunciação oral / falado. A nosso ver, algumas formações contidas na categoria anterior deveriam fazer parte desta, como *kd* e *q*, já que é aproveitado o modo de enunciação oral / falado de suas letras na formação da abreviatura.

3) **Abreviaturas em que há simplificação da grafia**: *aki* – aqui; *xau* – tchau; *ki* – que.

Esta categoria contempla as simplificações ortográficas, principalmente de dígrafos, como alerta a autora. Aproveita-se a correspondência do som do grafema único e do dígrafo. O mesmo ocorre no caso do grafema *k*, substituindo o dígrafo

qu. Porém, o uso de *i* caracteriza mais uma vez o modo de enunciação oral / falado da palavra *que*.

Fusca (2008) alerta para o fato de as abreviações serem constituídas por mais de um processo:

> É importante reafirmar que nenhuma abreviatura constitui-se exclusivamente por um único processo. Como vimos anteriormente, as abreviaturas formadas por simplificações gráficas também podem sofrer modificações baseadas na modalidade oral/falada praticada pelo escrevente – o mesmo ocorre com as abreviaturas das outras categorias. Longe de propor categorias estanques e homogêneas, tencionamos demonstrar as regularidades que constituem as abreviaturas, bem como as que contribuem mais fortemente para a formação de cada uma delas. (p. 18)

Veremos, em nossa categorização, que procuramos selecionar as características que mais se destacavam em cada forma de abreviação, classificando as palavras em apenas uma categoria.

4) Abreviaturas formadas por empréstimo: *add* – adicionar.

Aqui estão incluídas as abreviaturas que exploram palavras institucionalizadas em outras línguas, mas que são facilmente entendidas pelos usuários do internetês.

Ao seguir na nossa pesquisa, percebemos a necessidade de modificação no conceito de algumas categorias descritas por Fusca (2008) e de inclusão de novas. Vejamos, então, como procedemos:

1) Abreviaturas que mantêm o primeiro grafema de cada sílaba	Nesta categoria, mantêm-se as abreviaturas elencadas por Fusca (2008), excetuando-se as que exploram o som da letra em sua confecção, como *kd* – cadê, *q* – que, *t* – te etc. Exemplos: *vc* – você; *tc* – teclar; *td* – tudo ou todo; *pq* – porque.
2) Abreviaturas que indicam modo de enunciação oral / falado	Contempla as abreviaturas e palavras que tentam, na forma escrita, representar a oralidade, incluindo as que exploram o som da letra na formação, como as descritas na categoria anterior.

	Exemplos: *q* – que; *d* – de; *amu* – amo; *nunk* – nunca.
3) Abreviaturas em que há redução da grafia	Como no estudo de Fusca (2008), representa as simplificações de grafia, que preferimos chamar de "redução da grafia", por entendermos que a palavra não se torna mais simples, menos complexa do que a anterior. São incluídos dígrafos e outras formas inteligíveis em que se utilizam menos caracteres. Exemplos: *ñ* – não; *aki* – aqui; *bixos* – bichos.
4) Palavras e abreviaturas formadas por empréstimo	Incluímos nesta categoria, além das abreviaturas, toda e qualquer forma de expressão que se utilize de outra língua em sua formação, por acreditarmos ser esse aproveitamento uma característica do internetês. Exemplos: *add* – adicionar; *in* – em; *friend* – amigo.
5) Palavras em que faltam notações lexicais	Esta categoria abrange as palavras, abreviadas ou não, em que o usuário não tenha incluído alguma notação lexical, com o fim de escrever de forma mais fluente. Exemplos: *nao* – não; *ja* – já; *ate* – até; *distancia* – distância.
6) Abreviaturas formadas por siglas e apócopes	Estão inclusas as siglas e as abreviaturas em que houve apócope (queda de um ou mais fonemas no final da palavra). Exemplos: *h* – homem; *cam* – câmera; *comu* – comunidade.
7) Abreviaturas e contrações já consagradas pelo uso	Grupo de abreviaturas e contrações que são facilmente reconhecidas pelas pessoas, independentemente de conhecerem ou não o internetês. Exemplos: *pra* – para; *h* – hora; *av* – avenida.
8) Representação de risadas	Esta categoria contempla as diversas formas de representar as risadas na forma escrita.

	Exemplos: *rs*; *hahaha*; *kkkkk.*
9) Palavras marcadas afetivamente, interjeições e substituições de acento agudo por h	Estão incluídas aqui palavras que representem uma tentativa de inovação na linguagem, sem necessariamente agilizar a escrita, marcando uma aproximação entre os usuários ou uma emoção. Também estão nesta categoria as interjeições e os casos em que os acentos agudos são substituídos por h. Comparando com a categorização de Fusca (2008), estão incluídas aqui palavras em que a nasalidade é representada por letras em vez de til, as quais a autora classificava na categoria "modo de enunciação oral / falado" (*naum* – não) Exemplos: *raivaaaaaaaaa*; *kaos* – caos; *naum* – não; *urrúú.*
10) Palavras que contêm erros ortográficos ou de digitação	Reunião de palavras encontradas no corpus que contenham algum tipo de erro ortográfico, sem qualquer intenção de economia ou de inovação, ou equívoco de digitação. Exemplos: *cazado, min*, *aconpanhar*.
11) Abreviaturas em que houve supressão do meio da sílaba	Esta última categoria abrange as palavras que tiveram quedas de elementos de seu interior, mantendo o as letras iniciais e finais. Exemplos: *qndo* – quando; *qlqr* – qualquer; *cm* – com.

Percebemos que, mais do que uma simplificação da escrita, o internetês não possui apenas siglas e simplificações de escrita, como alerta Eisenkraemer (2006):

> Teóricos definem o *Internetês* como uma simplificação da língua, em que são abundantes as abreviações e siglas, no entanto, verificamos muitos casos contrários, como o da complexificação, em que há o acréscimo de letras à palavra original, mas com um fundo intencional, a fim de expressar, por exemplo, a força do significado de uma palavra. (p.14)

Vejamos, então, os resultados encontrados seguindo a categorização exposta acima. Primeiro, analisemos o corpus coletado nos chats, segundo metodologia exposta no capítulo anterior:

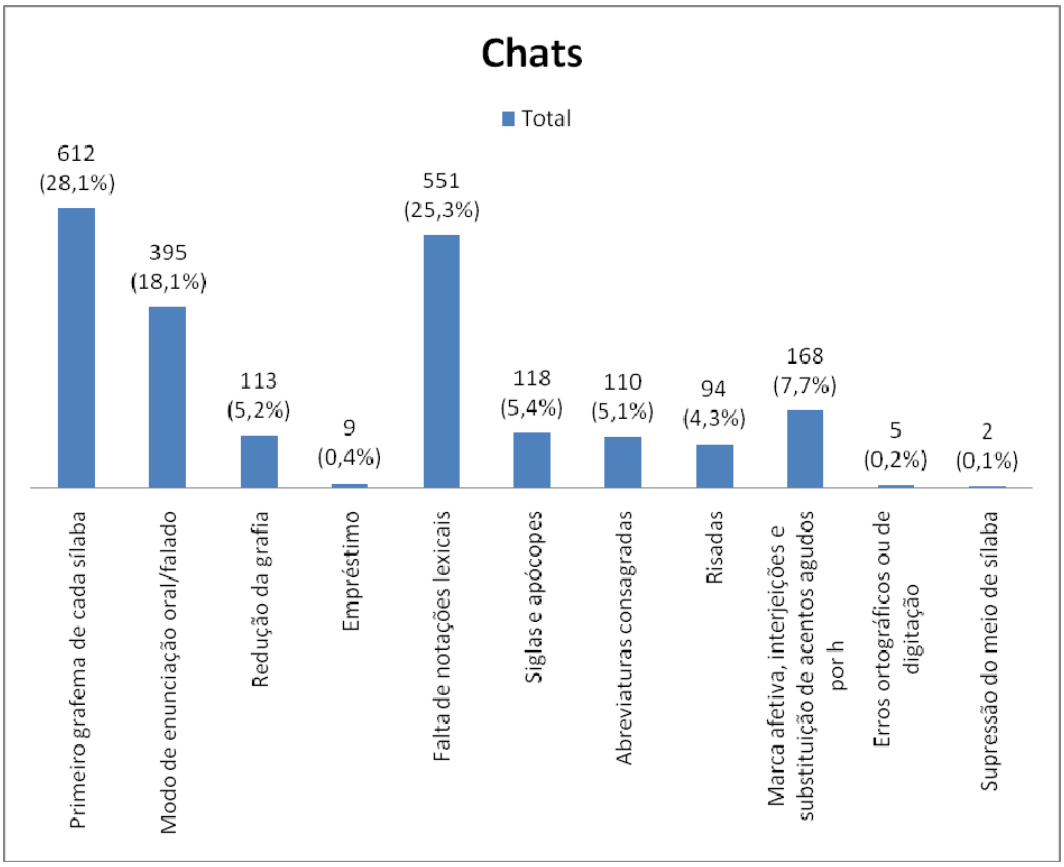


Gráfico 1 – Quantidade de palavras por categorias nos chats

Como primeiro dado, trazemos a porcentagem de palavras de nosso corpus que sofreram qualquer tipo de modificação, e que foram incluídas nas categorias acima: 6,7%. Ou seja, 2177 palavras de um total de 32647 do corpus. Devemos lembrar que contabilizamos apenas as palavras que tiveram pelo menos duas ocorrências no corpus, o que faz com que a percentagem exposta reflita dados aproximados. Porém, podemos ter uma ideia da utilização de termos do internetês quando comparamos aos outros dados, dos outros corpora.

Vemos que a categoria com o maior número de ocorrências é a de abreviaturas que utilizam apenas o primeiro grafema de cada sílaba. As abreviaturas mais utilizadas foram, em ordem decrescente:

(1) *vc* – **você**, com 309 ocorrências;
(2) *tc* – **teclar**, com 123 ocorrências;
(3) *td* – **tudo** ou **todo**, com 33 ocorrências;
(4) *pq* – **porque**, com 25 ocorrências;

Percebemos que as abreviaturas com maior número de ocorrências nesta categoria são palavras típicas de uma conversação em chats. A nosso ver, como o contexto e a recorrência dão pistas de seus significados, os usuários não veem problemas em ganhar tempo abreviando desta forma, já que a velocidade da conversação síncrona os obriga a terem tal habilidade.

A segunda categoria com maior número de ocorrências é de palavras escritas sem alguma notação lexical, que traz como exemplos de palavras mais utilizadas:

(1) *nao* – **não**, com 117 ocorrências;
(2) *alguem* – **alguém**, com 58 ocorrências;
(3) *ola* – **olá**, com 48 ocorrências;
(4) *so* – **só**, com 43 ocorrências;

Analisamos que a falta de notações lexicais se explica também pela necessidade de velocidade de uma conversa escrita síncrona, como nos chats. Além disso, em seu início, tais programas de bate-papo não permitiam o uso de caracteres especiais, o que acabou condicionando alguns usuários a procederem de tal maneira. Contudo, a principal razão para esta ser a segunda categoria com maior número de ocorrências deve ser a necessidade de rapidez. Também aqui, não há qualquer problema de compreensão do significado das palavras, já que a recorrência e o contexto determinam a correta decodificação.

Outra análise que podemos fazer dos resultados encontrados nesta categorização é referente à terceira categoria com maior número de ocorrências, a de abreviaturas baseadas no modo de enunciação oral/falado. Vejamos as abreviaturas mais utilizadas:

(1) *q* – **que**, com 94 ocorrências;
(2) *ta* – **está**, com 72 ocorrências;
(3) *to* – **estou**, com 46 ocorrências;
(4) *tô* – **estou**, com 21 ocorrências;

O chat, como sabemos, tenta representar uma "conversação oral escrita", ou seja, traz a sincronia e a informalidade da conversação oral por meio da escrita. Com isso, os usuários acabam tentando dar à escrita características orais, explorando o som de algumas letras ou reproduzindo palavras como elas são ditas. Vemos os exemplos acima, em que, com 94 ocorrências, está a abreviatura *q*, em que temos o mesmo som na leitura da letra e da palavra inteira (*que*). Já em (2), (3) e (4), temos palavras tipicamente utilizadas na conversação informal oral, em que são suprimidas as sílabas iniciais do verbo *estar*.

É importante salientarmos que a categoria que demonstra palavras com erros ortográficos traz consigo 5 ocorrências (0,2% do total de palavras do internetês), de apenas duas palavras (*cazado* – casado e *sinomose* – cinomose).

Seguimos, agora, com a análise dos dados coletados no Orkut, que tem, como característica principal em comparação com os chats, a comunicação assíncrona:

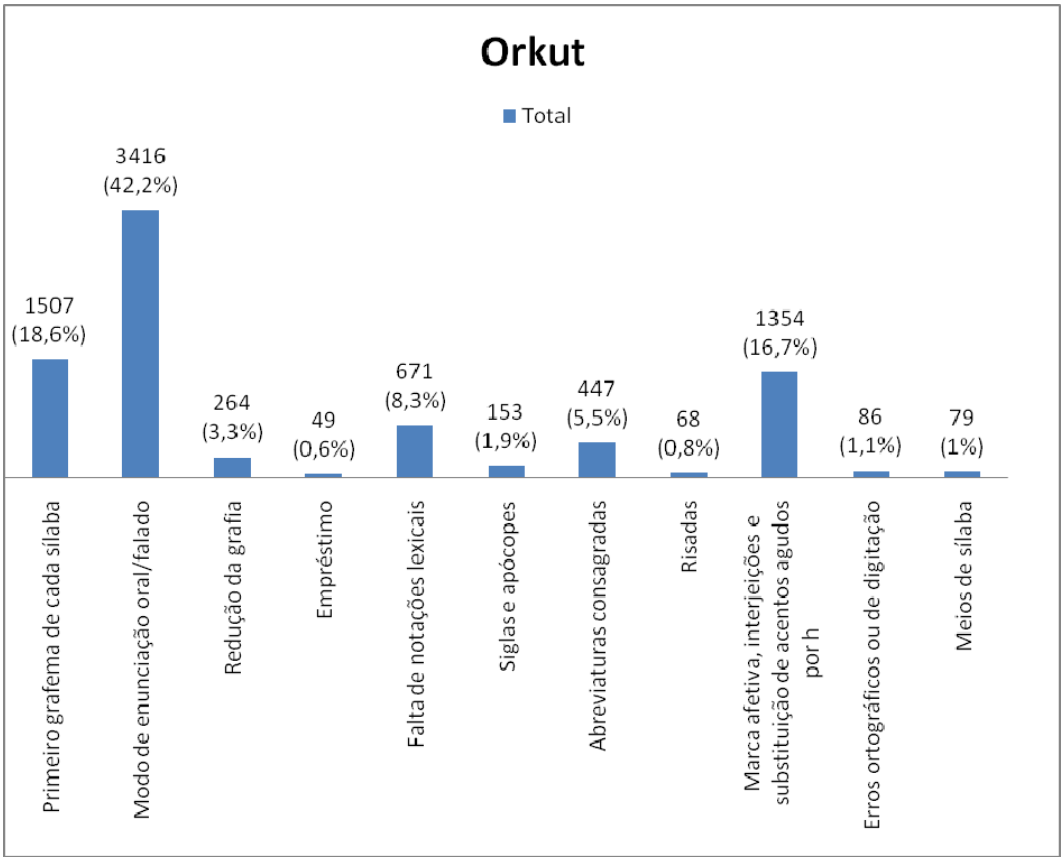


Gráfico 2 – Quantidade de palavras por categorias no Orkut

Com os dados do Orkut, já percebemos algumas diferenças significativas em relação ao corpus dos chats. Primeiramente, temos uma percentagem maior de

expressões típicas do internetês, contabilizando o total de palavras reunidas. De 40802 palavras, 8095 estão dentre as categorias descritas, ou seja, 20%, número bem superior ao do chat.

Os dados nos mostram que, diferentemente do que poderíamos imaginar, a assincronia da comunicação via Orkut faz com que seus usuários utilizem mais palavras do internetês do que os dos chats. Vejamos mais detalhadamente as características dessas palavras, para tentarmos encontrar os motivos que causam esse número maior de palavras do internetês.

Primeiramente, percebemos que a categoria de abreviaturas mais utilizadas não é a mesma do corpus de chats. Aqui, temos a que utiliza palavras que se aproximam do modo de enunciação oral / falado em primeiro lugar, com 42,2%. Vejamos as palavras mais utilizadas:

(1) *q* – **que**, com 861 ocorrências;
(2) *d* – **de**, com 206 ocorrências;
(3) *amu* – **amo**, com 202 ocorrências;
(4) *t* – te, **com** 162 ocorrências;

Com as palavras (1), (2) e (4), temos o aproveitamento do som da letra; já em (3), vemos a oralização da palavra *amo*. Em uma análise mais detalhada das palavras desta categoria, cogitamos que, por terem mais tempo para escreverem seus recados, devido à assincronia do Orkut, os usuários procuram inovar mais, utilizando palavras que caracterizem seu discurso como próprio da Internet. Ou seja, enquanto no corpus do chat percebemos que a preocupação maior é com a rapidez, no Orkut o interesse maior é mostrar que se tem habilidade e criatividade ao se criar textos no ambiente on-line.

A hipótese da criatividade é reforçada se investigarmos, em vez do número de ocorrências total de palavras, o número de variação tipológica, ou seja, quantas palavras diferentes foram utilizadas em cada categoria. Vejamos o gráfico:

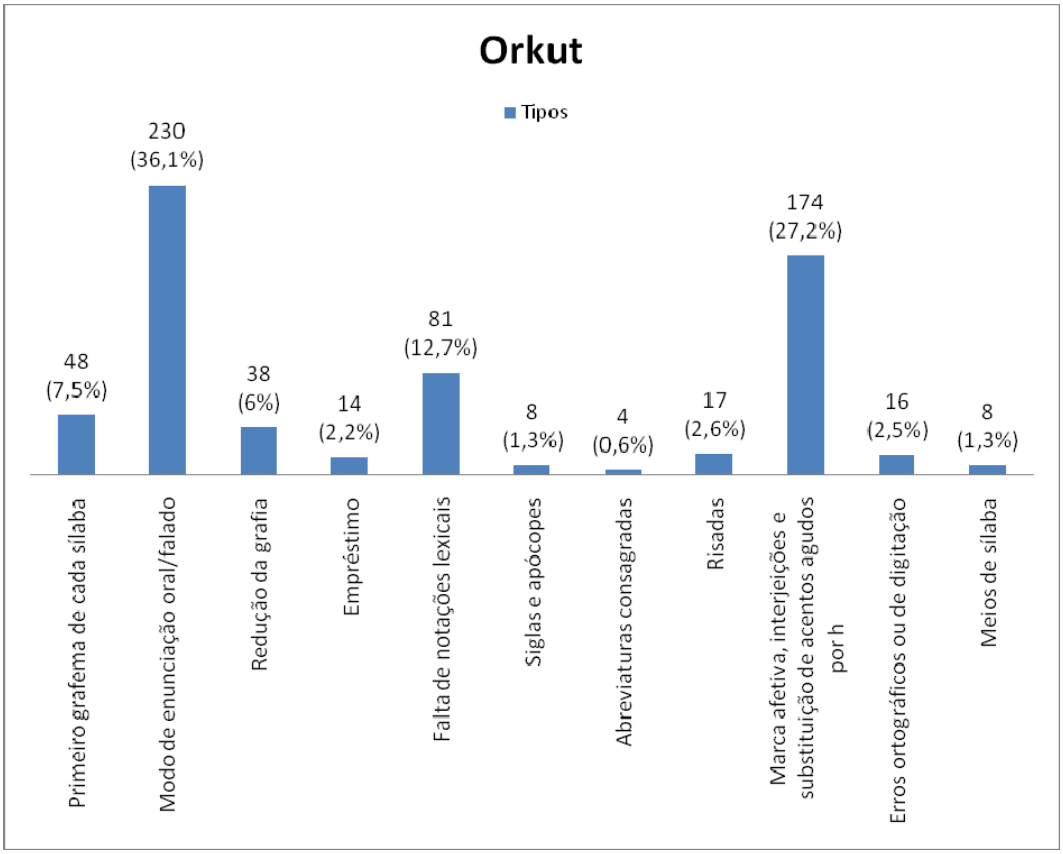


Gráfico 3 – Variação de palavras por categorias no Orkut

Enquanto há apenas 48 palavras diferentes na categoria que utiliza o primeiro grafema de cada sílaba, percebemos o número de variação de palavras que utilizam o modo de enunciação oral / falado: 230. Também vemos que há bastante variação na categoria que representa marcas afetivas, interjeições e substituições de acentos agudos por h: 174. Esses dados também sugerem que se procura muito mais no Orkut, e há tempo para isso devido à assincronia da comunicação, a inovação na linguagem, vistas as categorias que contêm maior variedade de palavras diferentes.

Também no corpus do Orkut encontramos pouquíssimas ocorrências de palavras com simples erros ortográficos, sem qualquer intenção de inovação na linguagem: 86, o que representa 1,1% do total de palavras do internetês. Além disso, os erros encontrados são simples, como trocas de letras (*min* – mim, *en* – em, *concelhos* – conselhos).

Vamos, então, à categorização das palavras encontradas no corpus do Twitter:

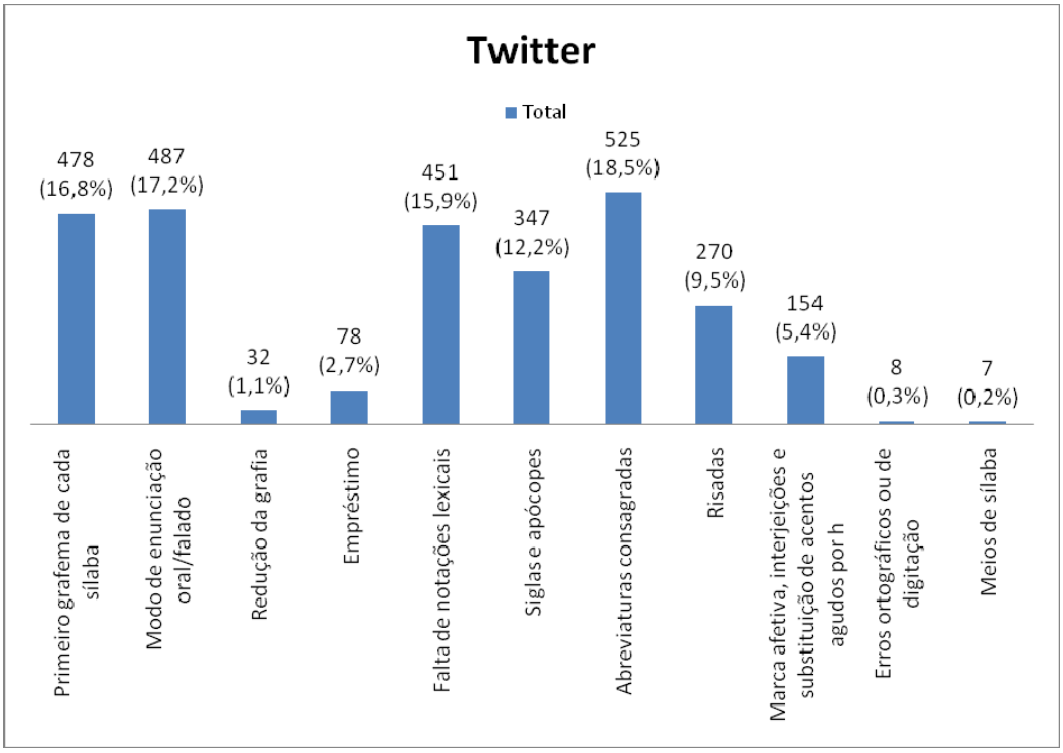


Gráfico 4 – Quantidade de palavras por categorias no Twitter

Das 59.805 palavras que compõem o corpus do Twitter, 2.837 estão entre as categorias acima expostas, ou seja, 4,7%, o menor índice entre os três corpora coletados. É interessante ressaltar esse dado, já que o Twitter restringe os usuários a utilizarem apenas 140 caracteres em cada mensagem, o que deveria obrigá-los a abreviar mais do que nos outros ambientes on-line. Porém, o recurso hipertextual dos links acaba expandindo o espaço dos tweets, o que faz com que os usuários redirecionem textos maiores para outras páginas que não imponham restrições de número de caracteres.

Vejamos, então, a categoria em que tivemos um maior número de ocorrências nesse corpus: a das abreviaturas consagradas. As palavras mais utilizadas desta categoria foram, respectivamente:

(1) *pra* – **para**, com 338 ocorrências;
(2) *h* – **horas**, com 77 ocorrências;
(3) *pro* – **para o**, com 55 ocorrências;
(4) *Av* – **avenida**, com 7 ocorrências;

Com esse resultado, diferente dos encontrados nos outros corpora, percebemos um dado importante sobre as mensagens postadas no Twitter. Como não há muito espaço para a contextualização do que se quer transmitir (como no Orkut) nem a interação direta síncrona (como nos chats), supõe-se que seja proposital o uso de abreviaturas e contrações já conhecidas, que não causem qualquer tipo de estranhamento por parte de quem as leia. As inovações linguísticas acontecem, porém, em menor número, já que a restrição do número de caracteres impõe que o tweet seja rápido, sucinto e inteligível.

Os números sugerem que os usuários têm como principal intenção garantir a inteligibilidade de suas mensagens é o grande número de ocorrências de palavras em que faltam apenas notações lexicais (15,9%). Devemos nos lembrar de que muitas mensagens são enviadas para o Twitter via SMS ou via Internet para celular, o que explicaria também essa percentagem, já que, ao digitar mensagens no teclado do telefone, as pessoas procuram agilidade, na maioria das vezes, não utilizando as referidas notações.

Em uma análise qualitativa dos dados, pudemos perceber também que, no corpus do Twitter, as abreviaturas das categorias que mantêm o primeiro grafema de cada sílaba e que indicam o modo de enunciação oral / falado são mais comuns, triviais, do que as que figuram essas mesmas categorias no corpus do Orkut. Enquanto no Twitter temos palavras como *blz* – beleza, *td* – tudo, *qd* – quando, *tou* – estou, *durmir* – dormir; no Orkut temos: *smp* – sempre, *nd* – nada, *sabs* – sabes, *pçoa* – pessoa, *vlhu* – velho.

Percebemos também que tanto as palavras dos chats quanto as do Orkut procuram reproduzir mais um discurso oralizado, já que suas mensagens geralmente têm caráter pessoal, afetivo, entre os usuários, diferentemente do Twitter, em que são postadas mensagens mais próximas a pequenos avisos, afastando a intenção de oralizar a linguagem neles utilizada.

São diferenças sutis, mas que existem, já que o ambiente assíncrono do Orkut, como já dissemos anteriormente, permite uma maior inovação na linguagem, além de o contexto, sem restrições como a do Twitter, de apenas 140 caracteres, facilitar o entendimento das novas palavras.

Quando analisamos a variação de palavras utilizadas no corpus do Twitter, temos o seguinte resultado:

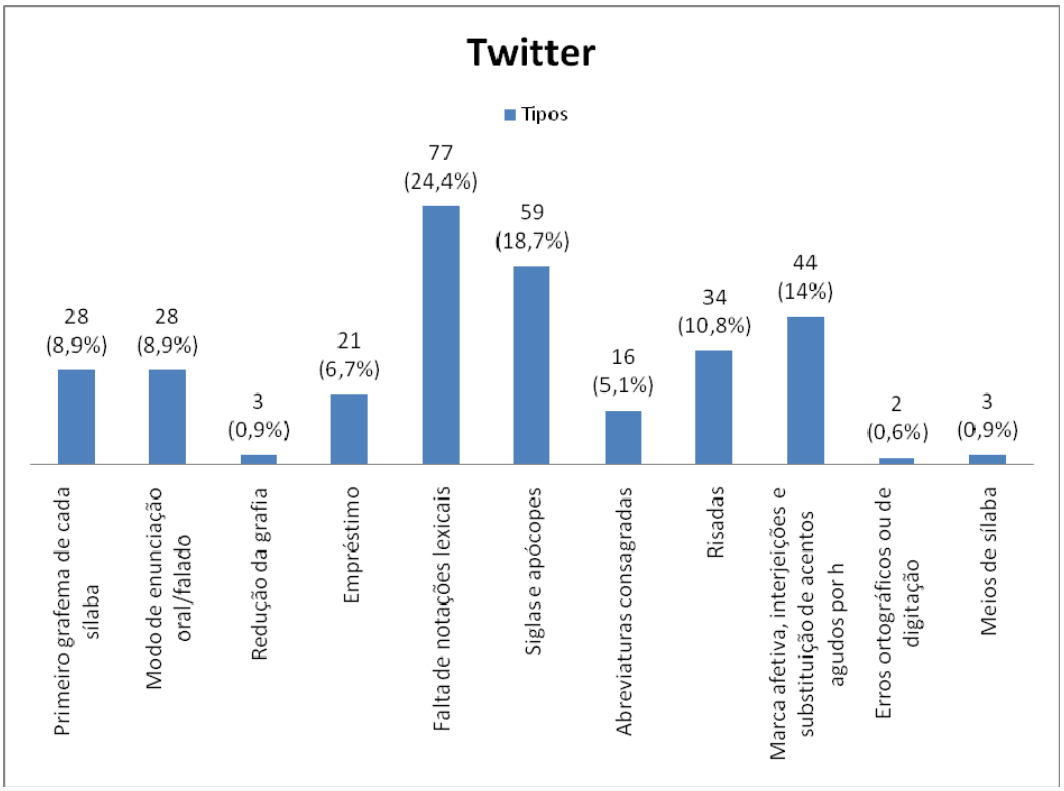


Gráfico 5 – Variação de palavras por categorias no Twitter

A maior variação ocorre na categoria em que são utilizadas palavras sem notações lexicais, o que demonstra uma maior funcionalidade desta categoria para o Twitter. Há também bastante variação na categoria que abrange siglas e apócopes, classificando-a também como bastante funcional no Twitter. Em uma análise qualitativa, percebemos que tais siglas e apócopes, no caso do Twitter, também são bastante conhecidas, como siglas de estados (SP, RJ etc.), e outras do próprio Twitter (DM, TT etc.), o que reforça nossa afirmação sobre a preocupação dos usuários com a clareza do texto, devido à pouca contextualização causada pelo limite de 140 caracteres.

Vemos também que a categoria com o maior número de ocorrências (abreviaturas consagradas) tem pouca variação, o que nos mostra que tais palavras são bastante recorrentes, trazendo uma segurança para os usuários, mas não são tão funcionais como as das outras categorias.

No corpus do Twitter, também encontramos um número irrisório de palavras que ferem a ortografia da língua portuguesa: 8, que representa 0,3% das palavras do internetês.

Ou seja, após analisarmos os três corpora, de um total de 13108 palavras categorizadas como típicas do internetês, encontramos apenas 99 ocorrências de erros ortográficos ou de digitação, o que representa menos de 1%. Os dados mostram que não se justifica o medo de diversos professores tradicionais, os quais afirmam ser o internetês um grande vilão, ameaçando a pureza e integridade de nossa língua portuguesa.

Como resumo dessa análise comparativa de categorias de palavras do internetês, podemos desenhar o seguinte quadro:

Chats	**Percentagem aproximada de palavras do internetês**: 6,7%; **Categorias com maior ocorrência de palavras**: Abreviaturas que utilizam o primeiro grafema de cada sílaba, seguida das palavras que não utilizam notações lexicais; **Conclusões**: Devido à comunicação síncrona, possivelmente a principal intenção dos usuários é agilizar a digitação da sua linguagem. Há também a tentativa de trazer um caráter oral à conversação escrita, já que temos bastantes palavras típicas de uma conversação oral informal, como *tá* – está, *tô* – estou, *num* – não.
Orkut	**Percentagem aproximada de palavras do internetês**: 20%; **Categorias com maior ocorrência de palavras**: Palavras que se aproximam do modo de enunciação oral / falado, seguida das que utilizam o primeiro grafema de cada sílaba; **Conclusões**: O fato de ser um ambiente de comunicação assíncrona incentiva os usuários a criarem novos termos, inovando na exploração de sons de letras e sinais. Possibilidade reforçada quando analisamos a variedade de palavras utilizadas no corpus.
Twitter	**Percentagem aproximada de palavras do internetês**: 4,7%; **Categorias com maior ocorrência de palavras**: Abreviaturas e contrações consagradas pelo uso, seguida das palavras que não utilizam notações lexicais. **Conclusões**: A limitação de 140 caracteres faz com que os usuários utilizem abreviações e outros tipos de palavras bastante conhecidas, para que não se corra o risco de não serem entendidos.

5.2
A classificação das palavras quanto ao número de sílabas

Achamos importante, antes de analisarmos os padrões silábicos das palavras do internetês, mostrarmos a divisão das palavras que compõem as categorias do internetês levando em conta seu número de sílabas. Para fazermos tal classificação, deixamos de lado as palavras que não estavam na língua portuguesa (da categoria palavras e abreviaturas formadas por empréstimo) e as representações de risadas. Além disso, consideramos a variação de palavras, não seu número de ocorrências.

Primeiramente, vejamos os resultados do corpus dos chats:

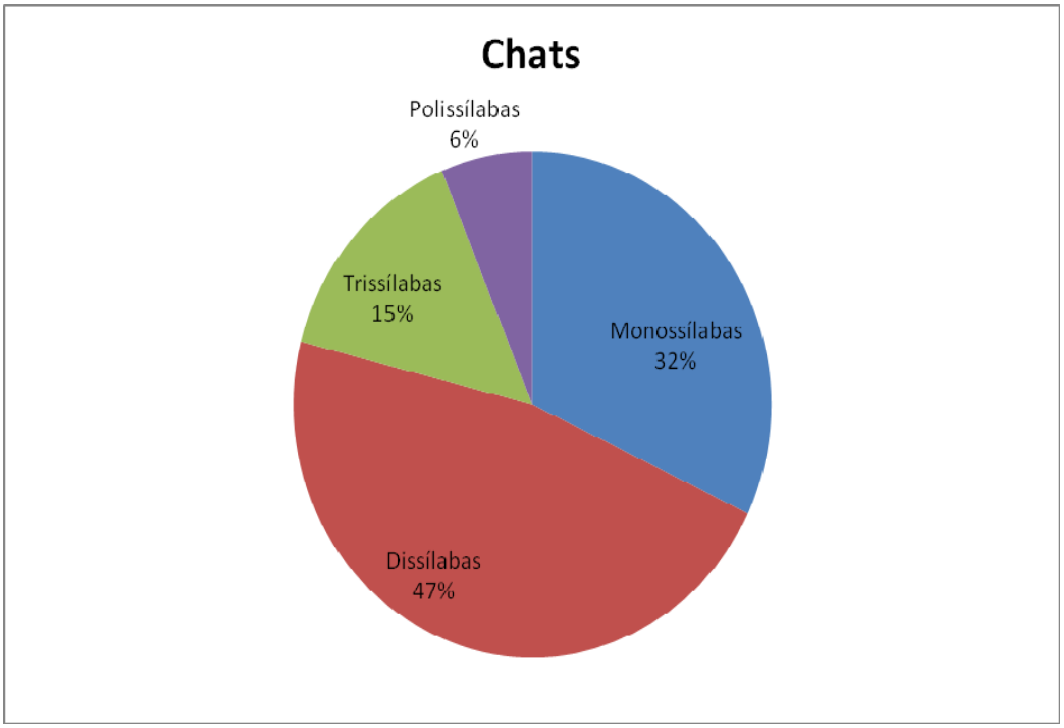


Gráfico 6 – Quantidade de palavras classificadas quanto ao número de sílabas nos chats

Devemos levar em conta que a linguagem utilizada em chats, Orkut e Twitter, em geral, é informal, o que nos leva a acreditar que a maioria das palavras utilizadas, abreviadas ou não, sejam mais curtas, em sua maioria mono e dissílabas. Porém, vale a pena ressaltar que o índice de palavras polissílabas no internetês dos chats é baixo, o que mostra a dificuldade de se abreviar ou modificar palavras maiores. Analisando palavra por palavra, pudemos ver que a única estratégia utilizada nas palavras polissílabas foi a retirada de notações lexicais, ou seja, modificação que não dificulta o entendimento das palavras.

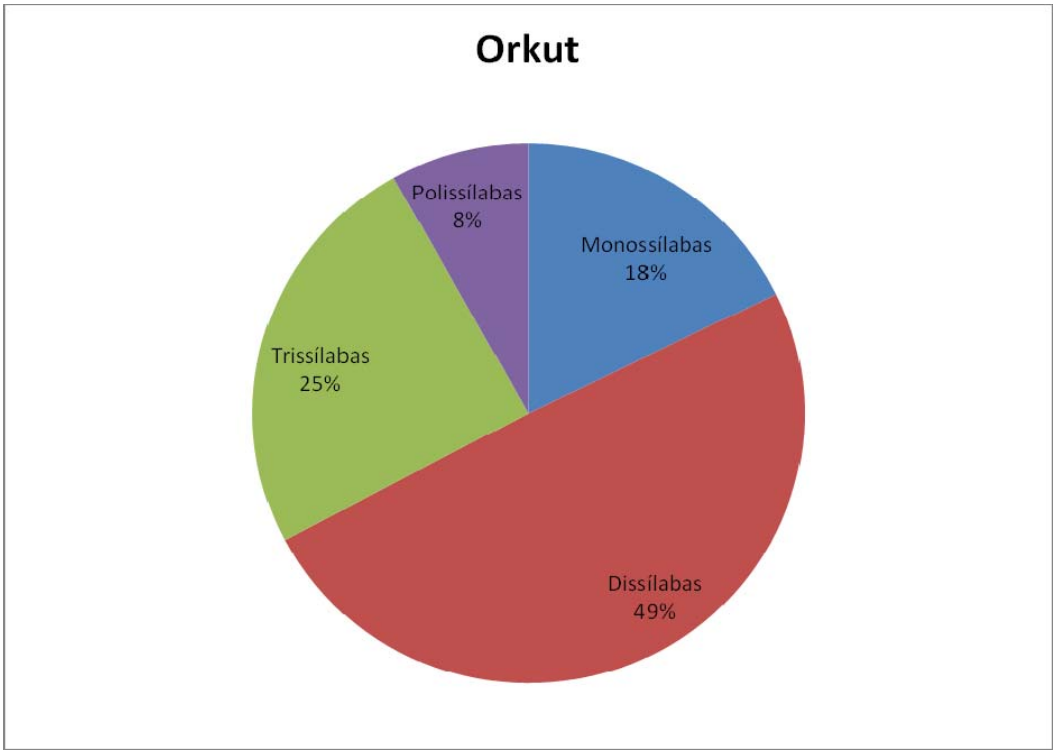


Gráfico 7 – Quantidade de palavras classificadas quanto ao número de sílabas no Orkut

Característica semelhante vemos nos dados do corpus do Orkut. O número de polissílabas aqui é um pouco maior. Percebemos também que tais palavras sofrem poucas modificações. Além de palavras sem notações lexicais, encontramos algumas em que é aproveitado o som de uma letra apenas (important, acontec), e outras que se enquadram na categoria modo de enunciação oral / falado, mas que também não dificultam o entendimento (encontradu, esquecendu).

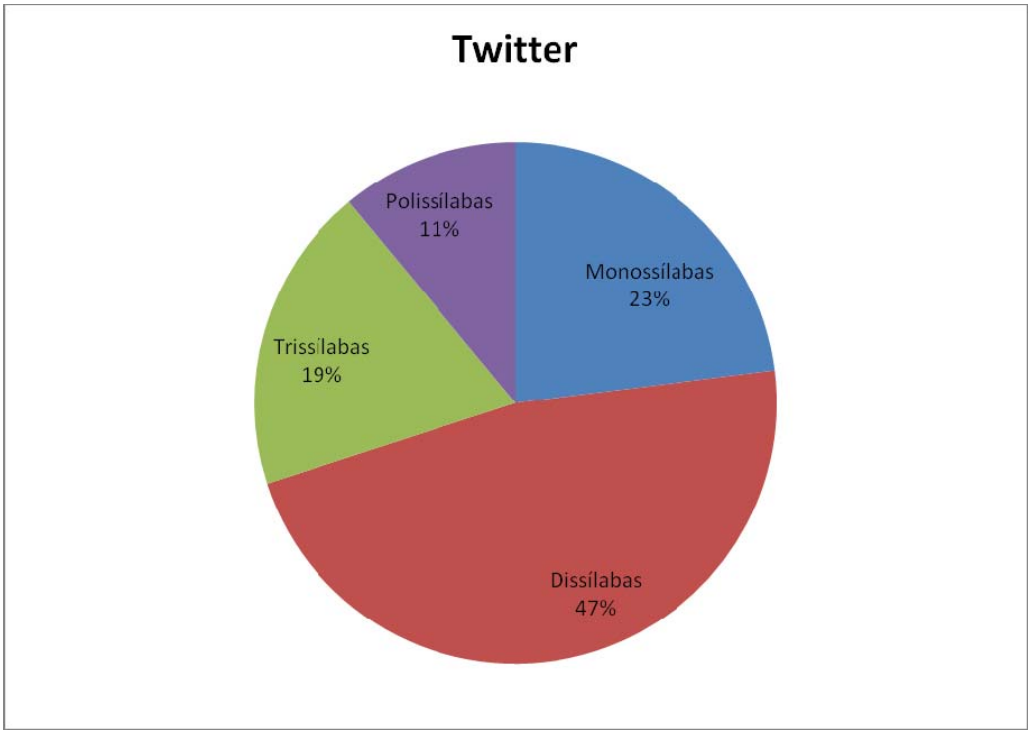


Gráfico 8 – Quantidade de palavras classificadas quanto ao número de sílabas no Twitter

Nos dados do Twitter, apesar de o índice de polissílabas ser um pouco maior do que nos outros corpora (o que nos faz acreditar que a linguagem seja um pouco menos informal), também percebemos, em uma análise qualitativa, que tais palavras sofrem modificações que não dificultam seu entendimento, principalmente abreviaturas conhecidas (*av* – avenida, *bijus* – bijuteria, *TV* – televisão, *tel* – telefone).

Chegamos a mais uma conclusão interessante, analisando o total de palavras dos três corpora quanto ao número de sílabas das palavras do internetês. Raros são os casos em que são abreviadas palavras com mais de três sílabas, apenas 8,4% do total. O fato de os usuários não inovarem muito na transposição dessas palavras para o internetês sugere a dificuldade de se abreviar palavras mais complexas, já que a grande maioria das palavras que compõem essa linguagem são monossílabas ou dissílabas (69,8%).

Passemos, então, a uma outra descrição, no nível silábico.

5.3
A análise dos padrões silábicos

A análise dos padrões silábicos das abreviaturas teve como base o estudo de Camara Jr. (1976), em que foram feitas análises da estrutura da sílaba em língua portuguesa. O autor nos mostra que as possibilidades de sílaba em língua portuguesa são: V, CV, VC e CVC, sendo V o elemento central da sílaba, ou seja, a vogal e C o elemento marginal, ou seja, as consoantes e semivogais. Vejamos as próprias palavras do autor no que diz respeito às sílabas livres e às travadas:

> Se chamarmos simbolicamente V o centro da sílaba e C um elemento marginal, teremos os tipos silábicos: V (sílaba simples), CV (sílaba complexa crescente), VC (sílaba complexa crescente-decrescente). Conforme a ausência ou presença (isto é, V e CV de um lado, e, de outro lado, VC e CVC), temos a sílaba aberta, ou melhor, livre, e a sílaba fechada, ou melhor, travada.

Continuando sua explanação, Camara Jr. ainda alerta que algumas sílabas CVC, que seriam consideradas sílabas travadas, podem ser consideradas livres:

> Já sabemos, por outro lado, que há em português, como alofones assilábicos, as vogais altas /i/ e /u/ (*pei-to*, *pau-ta*). Se eles funcionam como C, são não obstante de natureza V e surge o problema de representar tais sílabas como CVC ou CVV. [...] Em (C)VC está contido o conceito de sílaba travada, enquanto que em (C)VV está contido o conceito de sílaba livre.

Consideramos a classificação que contempla apenas uma vogal na sílaba (CVC). Ou seja, não adotamos o padrão CVV para sílabas como as mencionadas na citação acima.

Juntamente ao estudo de Camara Jr. (1976), adotaremos o trabalho de Veloso (2003), o qual acrescenta à lista de possibilidades de sílabas em língua portuguesa os padrões CCV e CCVC, os quais também são admitidos por Mattoso Jr., porém, mais explicitamente comentados em Veloso (2003).

Outro padrão encontrado em nossa pesquisa e admitido por Camara Jr. é o CVCC, o qual completa nossa lista de possibilidades de sílabas em língua portuguesa:

Tipos de sílabas	**Exemplos**
V	**á**-gua, sa-**í**-da, en-te-**a**-do
CV	**vo**-**cê**, **te**-clar, **tu**-**do**

CCV	**pro**-fes-sor, **pra**, **cri**-ar
VC	**an**-te-ri-or, **ex**-pres-so, **in**-tri-ga
CVC	**por**-que, **tam**-**bém**, **com**
CCVC	te-**clar**, **quan**-to, con-tem-**plar**
CVCC	de-**pois**, **mãos**, ar-ma-**zéns**

Trabalhamos com os mesmos dados já categorizados acima para a busca dos padrões silábicos mais abreviados. Entendemos, então, que nossa análise é válida em duas dessas categorias: as abreviaturas que utilizam o primeiro grafema de cada sílaba e as abreviaturas em que houve supressão do meio da sílaba. As outras categorias apresentam outras motivações em sua formação, como exploração dos sons das letras, redução da grafia, empréstimos de outras línguas etc., por isso não as analisamos nesse nível silábico.

Para fazer tal categorização silábica, consideramos as semivogais como C, tratando como V apenas as vogais, os elementos centrais da sílaba, segundo Camara Jr. (1976). Ou seja, em sílabas como **qua** de *quase*, adotamos o padrão CCV, já que *u* é apenas um elemento marginal que acompanha a vogal *a*.

Outra observação pertinente é o fato de considerarmos sílabas como **quer** de *qualquer* como CVC, já que temos um dígrafo *qu*.

Como não há razão para separar os corpora nesse nível de análise, unimos os dados dos chats, Orkut e Twitter. Fizemos, então, uma observação mais detalhada das palavras das categorias descritas anteriormente, separando cada uma de suas sílabas, de modo a encontrarmos os seguintes resultados:

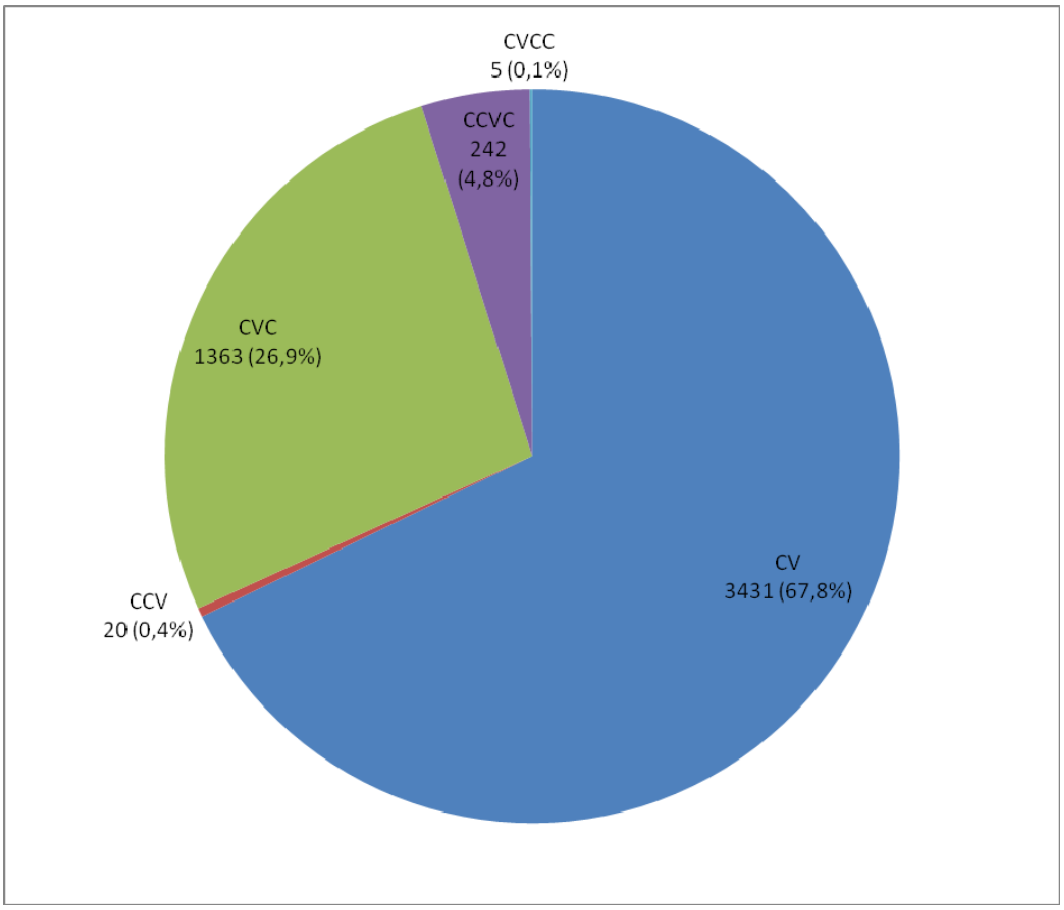


Gráfico 9 – Padrões silábicos mais abreviados em todos os corpora

A chamada sílaba canônica CV aparece como a mais abreviada em nossos dados, com ampla diferença para a segunda mais abreviada, CVC. Percebemos que os padrões V e VC não tiveram qualquer ocorrência de abreviação, o que leva à conclusão de que a abreviação desses tipos silábicos dificultaria decisivamente o entendimento da palavra que se quisesse escrever.

Os dados nos mostram, então, que as sílabas do padrão CV são as que apresentam maior facilidade de entendimento se abreviadas, configurando-se como mais produtivas.

Pudemos perceber que, em alguns casos, a nasalidade das sílabas do padrão CVC é marcada, como nas abreviaturas:

(1) *tbm* – também;

(2) *cntg* – contigo;

(3) *smp* – sempre;

O exemplo (1) é curioso, já que, apesar de ambas as sílabas serem do mesmo padrão e conterem nasalidade, somente a segunda é marcada. Uma possível explicação seria o fato de ser a segunda sílaba a tônica. Porém, no exemplo (2), a nasalidade marcada não está na sílaba tônica. Já, no exemplo (3), é possível que a letra *m* apareça na abreviatura para diferenciá-la da sigla *sp*, de *São Paulo*.

Devemos deixar claro que, segundo nossos dados, há equilíbrio entre a marcação e a não marcação da nasalidade em uma mesma palavra, posto que temos 72 ocorrências de *tbm*, e 50 de *tb*. No caso do pronome *contigo*, temos 16 ocorrências de *ctg* e 8 de *cntg*. No caso do advérbio *sempre*, todas as ocorrências foram da abreviatura *smp*, já que, segundo nossa percepção, a intenção não seria marcar a nasalidade existente na primeira sílaba, mas diferenciar a sigla da referente a São Paulo.

Também no padrão silábico CVC, encontramos marcação da presença da letra *s* em fim de sílaba, ora representando plural, ora não, como nos exemplos a seguir:

(1) *bjs* – beijos;
(2) *vcs* – vocês;
(3) *tds* – todos;
(4) *msm* – mesmo;

Nos exemplos de (1) a (3), temos a presença do *s* marcando o plural das palavras, estratégia bastante recorrente. Já em (4), vemos o uso do *s* não para marcar o plural, mas para deixar clara a intenção de abreviar a palavra *mesmo*, já que o uso de *mm*, utilizando somente as letras iniciais de cada sílaba, poderia trazer confusão na associação da abreviatura à palavra abreviada.

Quanto às sílabas de padrão CCVC, quase todas as ocorrências tratam das sílabas **quan**, de *quando*, e **qual**, de *qualquer*. A única sílaba além dessas que teve sua abreviação nesse padrão foi **clar**, de *teclar*, devido à grande recorrência, principalmente no corpus de chats, em que a chamada para a conversa caracteriza-se pela frase: *Alguém q tc?* – Alguém quer teclar?.

A abreviatura da sílaba do padrão CCV teve poucas ocorrências, todas com as sílabas **pra**, contração de *para*, **pre**, de *sempre*, e **quer** de *qualquer*.

Tivemos apenas 5 ocorrências de abreviação de uma sílaba do padrão CVCC, no corpus do Orkut: **pois**, de *depois*.

Dadas as observações, percebemos que, salvo nos casos de sílabas dos padrões CV e CVC, os casos de abreviações dos outros padrões silábicos são bastante específicos. Ou seja, 97,4% das sílabas abreviadas são dos padrões CV ou CVC, o que mostra uma grande diferença para os demais padrões silábicos em relação à possibilidade de entendimento da palavra abreviada.

Este estudo no nível silábico se mostra importante para uma possível formalização posterior das características do internetês para a criação de programas que entendam automaticamente as palavras do internetês. Veremos exemplos dessa formalização mais à frente.

5.4
Uma análise mais detalhada da estrutura silábica

Para que consigamos uma análise ainda mais específica sobre como se dá a abreviação das palavras do internetês no nível silábico, tomamos como base o trabalho de Silva (2009), o qual trata da investigação sobre quais constituintes silábicos são os mais suprimidos nas abreviaturas de chats.

Como referência de estudo, a autora utiliza o trabalho de Blevins (1995), que divide a sílaba em ataque e rima, e a rima em núcleo e coda. Segundo Silva (2009), ainda se referindo ao trabalho de Blevins (1995):

> Os constituintes silábicos podem ser preenchidos ou vazios, conforme estejam ou não associados a segmentos no nível terminal da representação, respectivamente. Podem ou não ser ramificados, conforme dominem um ou mais segmentos. A rima é o único constituinte obrigatoriamente preenchido (a não ser em casos excepcionais de existência de núcleos vazios). (p.25)

Tomemos como exemplo a palavra *mesmo*. Na sílaba *mes*, temos **m** como ataque, **es** como rima. Subdivindo-se a rima, temos **e** como núcleo e **s** como coda. Na sílaba *mo*, da mesma palavra tomada como exemplo, temos **m** como ataque e **o** como rima, a qual não apresenta subdivisão.

Em seu trabalho, Silva (2009) apresenta algumas categorias de sílabas abreviadas nos chats. A autora traz alguns exemplos de palavras em que as sílabas sofreram:

1) **Queda de ataque**, como em *kidu* – querido, em que a letra *r* da sílaba *ri* foi totalmente retirada;

2) **Queda de parte de ataque ramificado**, como em *tc* – teclas, em que percebemos a ausência da letra *l* na sílaba *clas*;

3) **Queda de rima completa**, como em *ctg* – contigo, em que, no caso das três sílabas da palavra, são excluídas toda a rima;

4) **Queda de núcleo apenas**, como em *pds* – podes, em que há ausência da letra *e* na sílaba *des*;

5) **Queda de parte do núcleo**, como em *na* – não, em que a letra *o* é retirada;

6) **Queda de coda apenas**, como em *memo* – mesmo, em que há a retirada da letra *s* na sílaba *mes*;

7) **Queda de sílaba**, como em *nina* – menina, em que é retirada toda a primeira sílaba;

8) **Falta de palavra**, como em *embarcação pesca* – embarcação de pesca, em que há ausência de uma palavra em uma expressão.

Aproveitamos os dados em que analisamos os padrões silábicos e, dessas categorias elencadas por Silva (2009), registramos em nossos dados a ocorrência de abreviações que se encaixam em apenas três delas: **queda de rima completa**, **queda de parte de ataque ramificado** e **queda de núcleo apenas**. Vejamos as estatísticas:

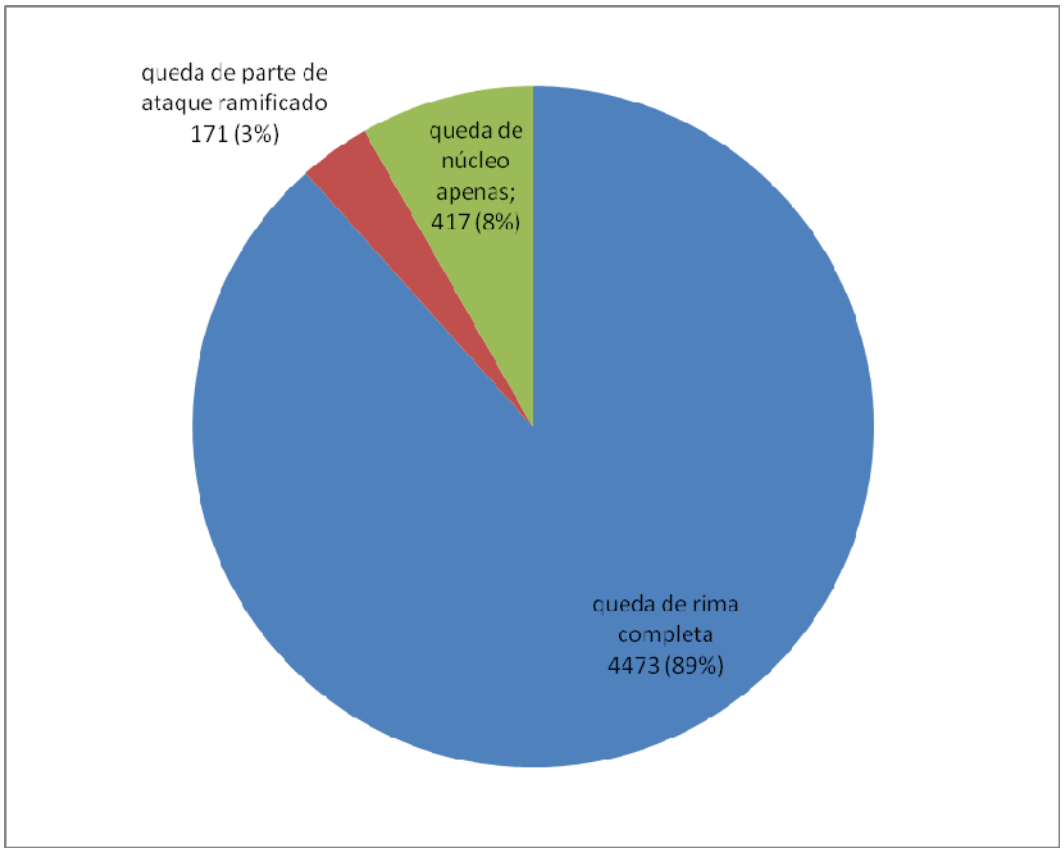


Gráfico 10 – Quedas silábicas em todos os corpora

Percebemos que a parte da sílaba mais suprimida nas abreviaturas do internetês é a rima, em 89% dos casos. Ou seja, as abreviaturas que mais ocorrem são as que mantêm a primeira letra de cada sílaba, suprimindo a rima completa, como *vc* – você, *pq* – porque. São poucos os casos (8%) em que apenas o núcleo é retirado, sendo preservada a coda, como na primeira sílaba de *msm* – mesmo, nas segundas de *tbm* – também, e *tds* – todos. Mais raras ainda são as quedas de ataques ramificados, como nas segundas sílabas de *smp* – sempre e *tc* – teclar.

Reunidos esses dados, vejamos como poderíamos tratá-los futuramente, com a criação de autômatos que expandissem as abreviaturas.

5.5
As diversas contribuições no estudo do reconhecimento automático

Para ilustrarmos como nossos dados podem ser úteis para a criação de um programa específico de reconhecimento automático de palavras do internetês em

língua portuguesa, mostraremos como cada estudo citado neste trabalho poderia contribuir com sua peculiaridade. Vejamos:

1) Primeiramente, como uma forma de pré-processamento de nossos dados, poderíamos criar um banco de dados com todas as palavras da língua portuguesa associadas às suas formas no *compression model*, conforme o trabalho de Shieber e Baker (2003), em que se retiram todas as vogais das palavras, exceto as iniciais. Assim, abreviaturas em que são retiradas apenas as vogais, como *vc* – você, já seriam reconhecidas pelo programa.

Também teríamos uma lista com abreviaturas e siglas já conhecidas, como as que foram incluídas nessa categoria em nosso trabalho, associadas às palavras correspondentes.

Outra lista que facilitaria o conhecimento prévio da palavra não dicionarizada seria a correspondente às palavras sem notação lexical.

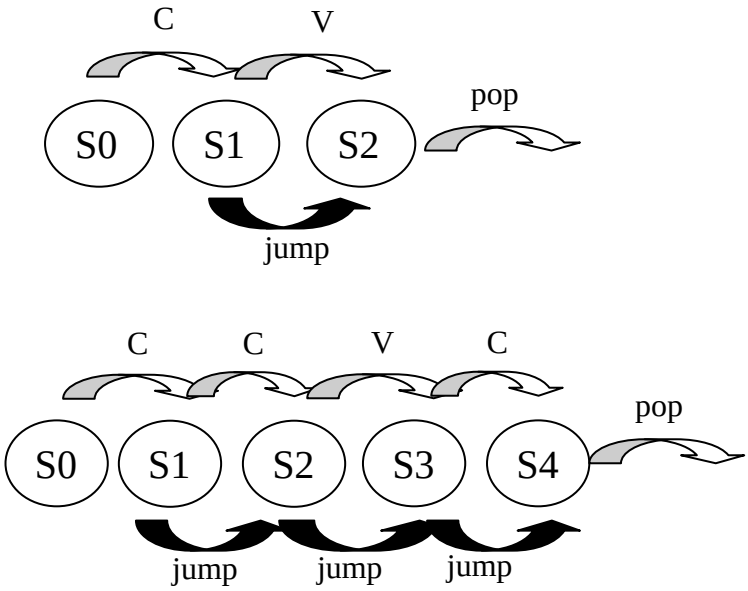
Porém, obviamente, muitas outras permaneceriam desconhecidas, e seriam submetidas a outras regras:

2) Seria feita uma comparação da sequência de letras da palavra de entrada com a lista de palavras dicionarizadas, de modo que fossem encontrados os casos de apócope, como *aniv*. Ou seja, se a sequência de letras da abreviatura corresponder exatamente à da palavra dicionarizada, como em *aniv* – aniversário, o programa daria a resposta. Obviamente, haveria várias opções de resposta, como, neste caso, aniversário e aniversariante. Porém, o usuário poderia selecionar a forma correta manualmente, tendo acesso ao contexto em que a palavra ocorreu.

Não obtendo sucesso na associação, o programa submeteria a palavra desconhecida aos passos seguintes.

3) Reuniríamos um banco de dados com os diagramas de autômatos de estados finitos (Allen, 1994), em que houvesse todas as sílabas possíveis em língua portuguesa e os possíveis caminhos para obtermos uma abreviatura, inclusive com as letras que compõem sílabas e as que não as compõem. Se temos, por exemplo, a abreviatura *nunk* (nunca), o programa não reconheceria a possibilidade da sílaba CVCC, já que não há sílabas em língua portuguesa com a consoante *k* depois de *n*. Seria reconhecida, então a sílaba possível *nun* (CVC) e *k*, como não forma sílaba inteira, teria seu significado associado ao modo de enunciação oral / falado *ca*. Chegaríamos, então, a uma sílaba CV, possível em

língua portuguesa. Juntando-se as sílabas, chegaríamos à palavra *nunca*. Vejamos dois exemplos de padrões sílabicos descritos em nosso banco de dados:



De acordo com nossos dados, só há uma possibilidade de abreviarmos sílabas do tipo CV, como **te**, de *teclar*: S0-S1-jumpS2-pop. Ou seja, mantém-se a primeira letra, a consoante.

Quanto às possibilidades de abreviação de sílabas do tipo CCVC, teríamos as seguintes, segundo nossos resultados: S0-S1-jumpS2-jumpS3-S4-pop (*qnd* – **quan**, de *quando*); S0-S1-jumpS2-jumpS3-jumpS4-pop (**qd** – *quan*, de *quando*).

Também utilizaríamos as informações sobre as quedas mais frequentes (rima, núcleo, ataque).

4) Do trabalho de Park e Byrd (2001), poderíamos aproveitar a ideia de substituir cada letra da abreviatura por *c*, representando caracteres alfabéticos, e *n* representando caracteres numéricos e símbolos. Assim, quando a abreviatura contiver um número ou um símbolo, será associado a esse número ou símbolo o seu significado. Por exemplo, ao processar a palavra do internetês *T+*, consideraríamos a sequência *cn* (caractere alfabético, seguido de caractere numérico), sendo que o *n* seria tratado como um símbolo e os possíveis significados atribuídos a ele seriam *mais* e *mas*. O processo para descobrir o significado de *T* (até), seria outro, conforme os passos a seguir.

Para chegarmos ao nível silábico, antes teríamos de reconhecer onde começa e onde termina cada sílaba de cada abreviatura. Para isso, vejamos o passo seguinte:

5) Ao ser computada uma palavra do internetês, não reconhecida, um programa, como o proposto no trabalho de Gouveia, Teixeira e Freitas (2000), tentaria dividi-la em sílabas. Aqui, poderíamos aproveitar a estratégia proposta por Cook e Stevenson (2007), nos seus *lexical blendings*. Em vez de procurarmos por palavras aglutinadas possíveis, utilizaríamos a ideia para procurarmos sílabas possíveis. Se, por exemplo, tivéssemos a abreviatura *tc*, primeiramente, o programa reconheceria que não seria possível uma palavra com a sequência de duas consoantes CC. Com a informação de que os dados são provenientes de chats, o primeiro caminho a ser seguido pelo programa seria o de reconhecer cada consoante como o início de sílaba, já que, segundo nossos dados, a abreviatura mais recorrente em tal ambiente é a que utiliza o **primeiro grafema de cada sílaba**. Então, teríamos uma palavra com duas sílabas, as quais seriam submetidas aos diagramas criados.

Seguindo o modelo de Cook e Stevenson (2007), se tivéssemos, por exemplo, a abreviatura *qro*, teríamos o modelo CCV, possível para sílabas em língua portuguesa. Porém, a informação de que não há sílabas válidas com a sequência das consoantes *qr* faria o programa reconhecer outras possibilidades de sílabas da abreviatura *qro*. Como também não temos a sequência CC, para considerarmos *qr* uma sílaba válida, o programa chegaria à conclusão de que, na verdade, teríamos na palavra a sílaba válida *ro* (CV) e *q* representaria uma outra sílaba, a qual seria submetida ao diagrama.

Haveria, nesta etapa, a informação de que as letras *s*, *m* e *n* poderiam fazer parte da sílaba correspondente à consoante anterior, pois esse é um dado encontrado em nosso trabalho. Ou seja, nos casos de abreviaturas do tipo *cntg* (contigo), a consoante *n* poderia fazer parte da mesma sílaba da consoante *c*, em vez de ser a inicial de uma nova sílaba.

Um maior detalhamento sobre as abreviaturas poderia contribuir ainda mais. Por exemplo, em nossa investigação, pudemos perceber que as palavras do internetês raramente têm mais de três sílabas. Ou seja, no caso de *cntg*, com essa informação, o programa chegaria à conclusão de que a consoante *n* faz parte da sílaba anterior, iniciada por *c*;

6) Divididas as sílabas da abreviatura, cada uma seria analisada segundo o diagrama proposto no passo 3). Reconhecida, elas seriam unidas, de modo que

formassem a palavra. Sendo essa palavra reconhecida, estaria completo o processo de reconhecimento da palavra do internetês.

Vejamos, agora, um exemplo concreto de como uma palavra seguiria tais passos até a ser descoberta:

Pensemos que o programa recebe como entrada a palavra *msmo* e a informação de que ela foi coletada no Orkut.

– o primeiro procedimento seria compará-lo ao banco de dados que contém palavras sem vogais. Não seria encontrada nenhuma correspondência;

– depois, ela seria submetida ao banco de abreviaturas já conhecidas e às palavras sem notações lexicais. Também não seria encontrada correspondência;

– o programa tentaria associar a sua sequência de letras às sequências das palavras dicionarizadas, não encontrando ainda um resultado;

– associaríamos as etiquetas cccc para a palavra, já que ela é composta apenas por caracteres alfabéticos, não havendo a necessidade, então, de se procurar significados alternativos para qualquer caractere numérico;

– o programa procuraria as sequências silábicas válidas. Primeiramente, analisaria que não há a possibilidade da sílaba CCCV (*msmo*). Então, procuraria outras possibilidades. Com as informações detalhadas sobre as possíveis sílabas em língua portuguesa, o programa chegaria à conclusão de que há duas sílabas na palavra, já que *mo* compõe a sílaba CV, a mais recorrente, e *ms* são consoantes que fazem parte da mesma sílaba, já que também há a informação de que a letra *s* pode fazer parte da mesma sílaba da consoante anterior;

– cada sílaba seria submetida, então, aos diagramas de nosso banco de dados. Nesse momento, haveria também a informação de que, tratando-se de dados provenientes do Orkut, o primeiro caminho a ser seguido é aquele em que se aproveita o modo de enunciação oral / falado da letra. Além disso, outra informação que direcionaria a sílaba ao diagrama correto seria a que leva em conta os padrões silábicos em que mais ocorre queda de núcleo, de rima completa, de parte do ataque ramificado etc. No caso de queda de núcleo apenas, o padrão mais abreviado é o CVC. Ou seja, a sílaba abreviada *ms* seria submetida ao diagrama de sílabas CVC. Com essas informações, o diagrama encontraria facilmente o caminho em que *ms* se apresenta como abreviatura de *mes*, já que é aproveitado o som do fonema /m/;

– seriam, então, unidas as sílabas em uma só palavra, *mesmo*, considerada válida, encerrando o processo.

Obviamente, há muitos detalhes que compõem as palavras do internetês, os quais devem ainda ser pesquisados. Nossa proposta não pretende contemplar todos os casos de palavras não dicionarizadas, mas mostrar que é possível criar regras que tornem programas capazes de encontrar o significado correto para elas. Como alertamos, é necessário aproveitar a características de diversos estudos, de diversos autores, para que consigamos chegar a uma proposta consistente para a língua portuguesa.

Vimos que nossa proposta não é se basear apenas em listas prontas, as quais se mostrariam desatualizadas rapidamente, mas dar ao usuário uma resposta que se mantenha eficaz por um bom período de tempo.

O reconhecimento automático de palavras do internetês se mostra indispensável nos dias de hoje, em que, cada vez mais, os conteúdos de redes sociais, blogs e programas de bate-papo são objetos de estudo.

O trabalho de Freitas (no prelo 2011) fala sobre a importância de se formar corpora de dados digitais, além dos textos formais:

> Acreditamos que essa breve discussão seja suficiente para ilustrar a relevância de um corpus de blogs. Embora seja crescente o interesse na compilação de novos corpora para a língua portuguesa, notamos que, em geral, predominam corpora compostos por textos de jornal. Com relação a textos criados especificamente para ambientes digitais – como os blogs – , e no que pese a relativa facilidade na obtenção de dados de blogs – ou de outros textos veiculados na internet – e o interesse pela linguagem utilizada nos textos da CMC, temos conhecimento apenas do Corpus ANCIB, com cerca de 81.000 frases e que corresponde a mensagens de correio eletrônico da lista ANCIB, cujo conteúdo é bastante previsível e institucional (anúncios de conferências, majoritariamente), e da Coleção Dourada do Segundo HAREM (Carvalho et al, 2008) que inclui, dentre outros tipos de textos, blogs e textos da Wikipédia. (p. 3)

Desejamos também deixar claro que nossa pretensão não é modificar o corpus coletado, convertendo todas as palavras do internetês para palavras dicionarizadas, o que descaracterizaria os dados. Queremos fazer com que tais dados ganhem credibilidade, esclarecendo como funcionam os processos no caso da escrita na Internet, e facilitando o entendimento de diversas palavras por muitos não conhecidas.

A descrição de procedimentos a serem utilizados para o reconhecimento de abreviações por sistemas computacionais pode também, talvez, contribuir para

esclarecer as estratégias que nós, humanos, falantes de uma língua que dominam a escrita, utilizamos para o reconhecimento de abreviações desconhecidas.