ZELI MIRANDA GUTIERREZ GONZALEZ

LINGÜÍSTICA DE CORPUS NA ANÁLISE DO INTERNETÊS

MESTRADO EM
LINGÜÍSTICA APLICADA E ESTUDOS DA LINGUAGEM

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE SÃO PAULO
2007

ZELI MIRANDA GUTIERREZ GONZALEZ
zmgg@uol.com.br - zmgg@ig.com.br

LINGÜÍSTICA DE CORPUS NA ANÁLISE DO INTERNETÊS

Dissertação apresentada à Banca Examinadora da Pontifícia Universidade Católica de São Paulo, como exigência parcial para obtenção do título de MESTRE em Lingüística Aplicada e Estudos da Linguagem, sob orientação do Prof. Dr. Antonio Paulo Berber Sardinha

PUC - SP
2007

BANCA EXAMINADORA

_____

_____

_____

Aos que já foram, mas deixaram muitos ensinamentos:

Maria e Gutierrez

Aos que estão mais perto:

Gerald Schaz pela semente de um sonho
Olavo e Júnior pela compreensão
Guto pela força e dedicação

AGRADECIMENTOS

Gostaria, em primeiro lugar, de agradecer a Deus pela luz, pela força, pelas pessoas e pelos momentos maravilhosos, sem os quais, não seria possível transpor muitos obstáculos encontrados ao longo dessa jornada. Agradecer, também, por me mostrar que quanto mais aprendo, mais percebo que tenho muito a aprender.

Agradeço, em especial, ao meu orientador Prof. Dr. Tony Berber Sardinha, por ter me ensinado a arte de pensar o trabalho acadêmico e pelo constante incentivo, sempre indicando a direção a ser tomada nos momentos de maior dificuldade e pela co-autoria em vários trechos. Agradeço, principalmente, por sua sabedoria, dedicação, cobranças, silêncio, sugestões, sem as quais não conseguiria vencer essa etapa.

Às grandes amigas Tecla, Fran e Silvia pela grande amizade, por todo apoio e conforto nos momentos necessários e cruciais.

A minha mais que irmã Tatiana, quase filha, pelo companheirismo, admiração, respeito e atenção, sempre acreditando que eu seria capaz.

A Profª Estela de Jesus Martins pela amizade e sugestões, além da revisão desse trabalho.

À Solange e Ana Cláudia pela compreensão, sem a qual, eu não poderia concluir esse trabalho a tempo.

A todos os professores, funcionários e alunos do Mestrado em LAEL da PUC, e todos aqueles que, direta ou indiretamente, contribuíram para a realização desta dissertação, dando-me força, incentivo e principalmente, acreditando ser possível trabalhar o tema internetês.

Aos meus amigos Lívia e Júnior, por todo auxílio, com relação às dificuldades nas traduções de palavras e textos.

Aos meus filhos, Olavo e Júnior e ao Guto, meu companheiro, que souberam compreender, como ninguém, a fase pela qual eu estava passando durante a realização deste trabalho. Tentaram, sempre, entender minhas dificuldades e minhas ausências. Agradeço-lhes, carinhosamente, pela aceitação de todas essas falhas.

Ao longo dessa trajetória, muitas pessoas foram importantes e participativas, em maior ou menor proporção. A dificuldade em agradecer não está, particularmente, na citação das pessoas que lembramos, mas na falha daquelas que esquecemos de

mencionar. Então, esta página é dedicada a todos que, de alguma forma, participaram um pouco desses dois anos e, a outros que partilharam muito mais da minha vida contribuindo, aconselhando, motivando, orientando, persistindo, cuidando, ajudando, ouvindo, falando, colaborando e acima de tudo, confiando.

RESUMO

O trabalho que ora se apresenta foi motivado pela necessidade de compreender as modificações na grafia do internetês, bem como identificar a freqüência dessas modificações.

Esse trabalho teve como objetivo principal utilizar uma abordagem baseada em Lingüística de Corpus na identificação das palavras mais freqüentes do internetês, das freqüências de modificações na grafia e os padrões léxico gramaticais. Há vários trabalhos que lidam com a questão do internetês; entretanto, nenhum deles demonstrou empiricamente quão freqüente as modificações ocorrem.

Sendo assim, esse trabalho buscou preencher essa lacuna, sendo, portanto, capaz de demonstrar empiricamente a extensão dessas modificações. Para tanto, encontrou suporte teórico na Lingüística de Corpus, adotando as principais noções apresentadas por Biber (1998), Berber Sardinha (2004, 2006), Sinclair (1991,1996). Por enfocar as freqüências de uso de itens lexicais consideraram-se, mais especificamente, os trabalhos de Berber Sardinha (2000a, 2000b, 2004), Halliday (1991, 1992, 1993). Além da Lingüística de Corpus, o projeto também tocou em questões como: variedades lingüísticas, gênero, registro e grafia internáutica sob a perspectiva de Possenti (2006), Mollica (2007), Thurlow and Brown (2007), Crystal (2001), Othero (2004).

O corpus empregado na pesquisa foi coletado em blogs de jovens que utilizam a internet para comunicação. O corpus contém 135.021palavras e 15.552 formas. Para as análises dos itens lexicais consideraram-se as 500 palavras mais freqüentes do corpus de estudo. As freqüências detectadas serviram como base para a descrição das modificações ocorridas na grafia da variante lingüística – o internetês.

Entre os itens mais freqüentes do corpus, selecionou-se o item *'td'* com sentido de tudo, toda, todo, todos, todas, com a finalidade de verificar se os padrões léxico-gramaticais contribuíam para os respectivos sentidos.

Por conseguinte, a pesquisa pretende ter contribuído para o estudo do internetês, uma vez que há poucos trabalhos que demonstrem, de maneira empírica, essas modificações. O trabalho ainda apresenta as limitações da pesquisa e aponta sugestões para futuros estudos.

ABSTRACT

The study presented was motivated by the needs of comprehend the changes in the ortography of the Internet language, such as identify those changes frequency.

The main aim of this study was to focus on the usage of a Corpus Linguistics approach for identification of frequent words most used in the studies corpus, such as frequences of changes in the ortography and the lexican gramathical standards of the internet language. There is a great range of studies on the internet language; however, very few of them has demontrated empirically how frequent changes are.

Therefore, this study has tried to fill this gap by being able to show empirically the changes. The main theoretical underpinning for the research is provided by Corpus Linguistics, assuming the main notions presented by Biber (1998), Berber Sardinha (2004, 2006), Sinclair (1991,1996). For focusing the use frequency of lexican items it was considered, more specificly, the studies of Berber Sardinha (2000a, 2000b, 2004), Halliday (1991, 1992, 1993). Besides the Corpus Linguistics, the project also mentioned in questions such as: linguisctics diversity, genre, registry and internet language ortography along the perspective of Possenti (2006), Mollica (2007), Thurlow and Brown (2007), Crystal (2001), Othero, (2004).

The corpus employed in the study was collected of young people's blogs that use internet for comunication. This corpus contains 135.021 tokes and 15.552 types. For the development of this research and of the analysis of the lexican items it was considered all the 500 most used words in the corpus studies. The frequences were used as base for decription of changes happened in the variant linguistics ortography – the internet language.

Among the most frequent items in the corpus was selected the "td" item with the sense of 'all, every, everything' ('tudo, todo, toda, todas e todos' in portuguese), with the objective of verify the standards lexican-gramathical, contributed for the respective senses.

To sum up, this study hopes it has contributed to the study of the internet language, since there are few studies that have demosntrated empirically how these changes occur. This work also presentes the research limitations and its possible applications in the future.

SUMÁRIO

LISTA DE ANEXOS

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE GRÁFICO

Introdução

Com o advento da Internet, os usuários de computadores passaram a integrar uma rede que revolucionou a comunicação, derrubando fronteiras geográficas e aproximando inúmeros povos e suas diferentes culturas, independente de sua localização física no planeta. Dentre as múltiplas vantagens e novidades advindas da globalização virtual, a linguagem escrita - fonte principal de comunicação – não poderia permanecer incólume: novas palavras e expressões foram sendo inseridas no vocabulário de usuários da rede, atribuídas ao constante uso dos computadores e à busca de uma forma mais ágil de expressão.

O repertório de palavras adotadas em comunicações via Internet apropriou-se de significados tradicionais da língua, deu-lhes novos sentidos, introduziu estrangeirismos e neologismos de forma contínua e dinâmica, de tal forma que muitos deles já constam dos dicionários mais atualizados da língua portuguesa; por exemplo as palavras 'site', 'deletar', 'ciberespaço' e 'link'. Os internautas, cada vez mais ávidos por simplificação e praticidade, foram mais além: palavras vêm sendo abreviadas privilegiando a informação em si mesma, em processos onde apenas uma ou duas letras são suficientes para o entendimento do conteúdo da mensagem; por exemplo '*p/ c*' (para você) ou *'td meu S2'* (todo meu coração). Esse fenômeno, conhecido no Brasil como Internetês "trata-se simplesmente de aspectos da escrita empregada em e-mails, em chats, em blogs (...). Ainda mais especificamente, trata-se da grafia utilizada por certos usuários dos computadores, em geral, jovens adolescentes que passam horas 'teclando', isto é, trocando mensagens por escrito" (Possenti, 2006:30). Portanto, o internetês é visto como uma linguagem surgida no ambiente da Internet, baseada na simplificação informal da escrita. Exemplos seriam as palavras 'não', grafada em internetês como *naum, ñ* ou *num,* ou a palavra 'você', representadas como *vc, ce* ou *c*.

O internetês ocorre em vários idiomas, não somente no português. No inglês, por exemplo, é chamado de 'netspeak' e seus princípios parecem ser semelhantes aos do internetês, tais como as contrações (contractions), diminuição de palavras

(shortenings), acrônimos (acronyms), homofonia letras/números (letter/number homophones) entre outros. Tais fenômenos respondem pela grafia de palavras como em: *'Gd'* (*good*), *'msg'* (*message*), *'mon'* (*Monday*), *'aft'* (*after*), '*BFPO*' (*British Forces Posted Overseas*), '*DI*' (*Detective Inspector*), *'2getha'* (*together*), *'1s'* (*ones*) que significam respectivamente: bom, mensagem, segunda-feira, depois, Forças da Marinha Britânica, Inspetor Detetive, juntos, uns.

O internetês, com sua nova forma de grafar palavras, é mais comumente utilizado em meios eletrônicos de comunicação informal, tais como chats, blogs e mensagens de texto via celular[1]. O resultado desta comunicação rápida e instantânea traduz-se tanto em uma economia de caracteres digitados quanto em uma despreocupação com as normas ortográficas e gramaticais da língua portuguesa. Esta busca por economia é aparentemente movida por uma urgência na digitação, visto que muitas vezes o interlocutor está online, à espera de resposta, ou por limitações dos métodos de input de texto, como o teclado (keypad) restrito de muitos celulares que dificultam a digitação.

Dentre as características mais comuns nas mensagens em internetês estão abreviações, símbolos próprios e uma diversidade de pontuações utilizadas como recursos de comunicação. Como exemplo de abreviações estão os itens *'c'* (você), *'p'* (para) e *'qdo'* (quando); já entre os símbolos próprios utilizados com a finalidade de expressar sentimentos estão os chamados emoticons ou smile; **:-),** ou **;-)**, e a diversidade de pontuações dá-se principalmente com a finalidade de expressar a entonação ou surpresa sobre algum assunto, como em: '*Nossa!!!'* e '*Viu???'*.

A prática do internetês e as conseqüentes modificações gráficas utilizadas pelos seus usuários podem advir de fatores como: (1) ambientes síncronos com vários interlocutores, levando-os a expressar-se no mais curto espaço de tempo possível; (2) a influência da oralidade na escrita on-line; (3) o desejo de, por meio de símbolos, emoticons e sinais gráficos, facilitar a interação e criar vínculos afetivos entre os participantes (Othero, 2004; Santella, 2006). Esses recursos são, portanto, formas variantes convencionadas pela comunidade internáutica, com o intuito de aproximar a comunicação virtual das interações face-a-face, onde as palavras ganham inúmeras

[1] As mensagens de texto via celular são internacionalmente conhecidas como SMS (Short Message Service); no Brasil, correspondem aos Torpedos.

formas e sentidos com o auxílio de elementos paralingüìsticos, conforme a intenção do locutor.

Alguns especialistas em estudos da linguagem encaram o internetês como uma variedade de língua, como tantas outras. Segundo Possenti (2006:24) “uma coisa é a grafia, outra, a língua. Não há linguagem nova, só técnicas de abreviação. As soluções gráficas são até interessantes, pois a grafia cortada é a vogal”. Desta forma, as técnicas de abreviação com eliminação de vogais e consoantes não comprometeriam a língua, que é formada por regras e leis combinatórias (sintaxe e gramática). Ao contrário, as abreviações trariam soluções práticas, com intuito de agilizar a comunicação. Por exemplo, a palavra ‘gente’ nos meios digitais é grafada como *gnt,* assim o nome da letra ‘g’ sonoriza ‘ge’ e o nome da letra ‘t’ sonoriza ‘te’. Desta forma, ao suprimir principalmente as vogais, o nome das consoantes substitui o ‘som’ das vogais que não são escritas (Possenti, 2006).

Por outro lado, o internetês vem sendo fortemente criticado por alguns gramáticos puristas que acreditam na descaracterização da língua portuguesa por usuários que dele se utilizam para sua comunicação virtual. Os críticos afirmam que esse tipo de grafia transgride as normas estabelecidas pela gramática tradicional e seus autores – ainda segundo Possenti (2006:30) – “acusam os internautas de fazerem a espécie regredir e de destruírem a nossa amada língua portuguesa”.

Apesar de haver muitos trabalhos sobre internetês (Crystal, 2001, Thurlow and Brown, 2006, Jaffe, 2000; Androutsopoulos, 2000; Othero, 2004), há ainda muitas questões em aberto que precisam de resposta. Por exemplo, não sabemos ainda qual a extensão do uso de internetês na prática. Ou seja, apesar de estudos anteriores, terem identificado características do Internetês, como a perda de vogais, consoantes e substituições, não há dados que informem quantas e quais palavras são mais propensas a sofrer essa transformação. Além disso, também não sabemos qual a incidência de uso de internetês, ou seja, em um texto online onde ocorram palavras escritas em internetês, quantas dessas palavras sofrem mudança? Por fim, também não sabemos ainda quais processos de mudança de ortografia ocorrem; por exemplo, a mudança de ‘qu’ para ‘*k*’, ‘ão’ para *‘aum’*, entre outros e, quais os mais comuns. Para todas essas questões, precisamos de pesquisa empírica, que olhe

diretamente os textos em que aparece o internetês para entendermos melhor como ocorre esse fenômeno na prática.

Desta forma, neste trabalho, não pretendemos tomar partido em relação à validade ou não da grafia internáutica, propomo-nos simplesmente a compreender as modificações ocorridas na passagem da norma padrão para o internetês e a investigar essas modificações mais sistematicamente, focando o nível léxico-gramatical da língua através da exploração de um corpus eletrônico composto por blogs. Segundo Marcuschi (2005:61), "a princípio os blogs eram listas de links e sites interessantes que poderiam ser consultados, bem como notas de atalhos para navegação". Atualmente, os blogs são utilizados para anotações diversas, como poemas, críticas literárias, críticas de cinema, letras de música, exposição de idéias, opiniões políticas; enfim, qualquer texto que possa ser considerado dialógico no ambiente virtual. Como exemplo, destacamos o "Blog Especialista em Assuntos Diversos[2]", contendo várias alternativas de interação como links para filmes, poemas, piadas, letras de música, entre outras opções.

Dentre as várias definições de 'blog' obtidas no decorrer deste trabalho, a que consideramos como a mais adequada para fins de uma pesquisa sobre o internetês é a de Schittine (2004:186): "o blog, ou seja, o diário íntimo na internet, é um híbrido de vários tipos de escrita. Dividido entre vários estilos, ele se aproxima de uns, se afasta de outros, mas acaba tendo um pouco de cada um deles".

O gênero blog focado neste trabalho é um diário público, acessado por jovens que se comunicam, expressam seus sentimentos e pensamentos, bem como opinam sobre assuntos diversos. O gênero blog é movido por forças sociais e tecnológicas, operando em contexto definido e utilizando o internetês como variedade de língua; constitui-se, portanto, em um conjunto de dados lingüísticos autênticos em linguagem natural (Marcushi, 2002).

A pesquisa pretende levantar e apontar as probabilidades de uso de itens lexicais típicos do internetês encontrados em registros de blogs, recorrendo para tanto aos preceitos teóricos da Lingüística de Corpus (Biber et al., 1998; Berber

[2] http://empapucado.blogspot.com/.

Sardinha, 2004). Optamos por utilizar essa área de estudo porquanto, segundo definição de Berber Sardinha (2004:3):

> "A Lingüística de Corpus ocupa-se da coleta e exploração de corpora, ou conjuntos de dados lingüísticos textuais coletados criteriosamente, com o propósito de servirem para a pesquisa de uma língua ou variedade lingüística. Como tal, dedica-se à exploração da linguagem por meio de evidências empíricas, extraídas por computador."

A utilização do computador, ferramenta de suma importância para a Lingüística de Corpus, possibilita a execução de tarefas mais complexas, envolvendo extensos bancos de dados com grande potencial de pesquisa e análise, viáveis somente através de programas computacionais.

Esse contexto tecnológico favorece o foco da presente pesquisa, a qual tem a pretensão de investigar a grafia dos meios digitais por intermédio de corpora coletados de blogs, valendo-se das teorias da Lingüística de Corpus e contando com o auxílio da ferramenta computacional WordSmith (Scott, 1997).

1 - Objetivos da Pesquisa

O primeiro objetivo do presente trabalho consiste em utilizar os procedimentos e as noções oferecidas pela Lingüística de Corpus com o propósito de detectar as modificações mais freqüentemente utilizadas na grafia do internetês. O segundo objetivo de pesquisa consiste em extrair e mensurar as freqüências das formas dos itens lexicais em um dado corpus, com vistas à determinação de uma eventual padronização da linguagem da Internet.

Com esse objetivo em mente, a pesquisa pretende responder às questões:

1. Quais palavras são encontradas com maior freqüência no corpus de estudo?

2. Quais palavras caracterizam a utilização do internetês?

3. Quais as modificações que ocorrem com maior freqüência na formação das palavras do internetês?

4. Quais padrões léxico-gramaticais são mais comumente verificados no internetês?

2 - Organização da dissertação

A dissertação está organizada da seguinte forma:

O Capítulo 1 apresenta e discute a fundamentação teórica da pesquisa, retomando – além do conceito de Lingüística de Corpus e de suas principais características – a definição da terminologia adotada na área. Em seguida, faz um breve histórico das pesquisas realizadas com o auxílio da Lingüística de Corpus, destacando as noções mais relevantes para o estudo, tais como freqüência, concordância e padrões. Por fim, aborda a teoria da visão probabilística e sua aplicação em pesquisas, complementada pela discussão dos conceitos de pesquisa quantitativa ou qualitativa.

O Capítulo 2 discorre mais detalhadamente sobre a metodologia adotada, incluindo os objetivos e questões da pesquisa. Inicialmente, lista os procedimentos de coleta do corpus e apresenta o programa computacional WordSmith, fundamental para as análises desenvolvidas. A seguir, define os critérios para definição das palavras em internetês, seguidos da apresentação da lista de palavras, concordâncias e padrões. Finalmente, descreve os procedimentos seguidos na análise para a obtenção dos resultados que serão detalhados no capítulo 3.

O capítulo 3 apresenta e discute os resultados da pesquisa, compostos primeiramente de duas listas: as palavras mais freqüentes do corpus de estudo, e as

palavras mais freqüentes em internetês. As seções seguintes mostram os resultados das análises dos itens lexicais, bem como das classes gramaticais, modificadas ou não e suas respectivas percentagens. O capítulo encerra-se com uma listagem dos padrões léxico-gramaticais do internetês.

Seguem-se, finalmente, as considerações finais acerca das conclusões obtidas no decorrer da pesquisa, seguidas das referências bibliográficas e dos anexos mencionados ao longo da dissertação.

Capítulo 1 - Fundamentação Teórica

Neste capítulo, apresentaremos a fundamentação teórica que embasa o presente estudo. Inicialmente, buscaremos conceituar e caracterizar a Lingüística de Corpus como área de conhecimento, bem como definir o sentido do termo *corpus* neste contexto específico. Em seguida, após um breve histórico sobre a Lingüística de Corpus, seus campos de interesse e potencialidades transdisciplinares de aplicação, discutiremos a relevância das pesquisas desenvolvidas na área para os estudos lingüísticos em geral. Por fim, enfocaremos especialmente o estudo da probabilidade de ocorrências de traços léxico-gramaticais e sua relevância para o presente trabalho.

1.1 Lingüística de Corpus e Corpus: Conceituação e Características

A Lingüística de Corpus é uma área do conhecimento que estuda a linguagem por meio da utilização de grandes quantidades de dados empíricos relativos ao efetivo uso da linguagem, com o auxílio do computador.

A principal característica da Lingüística de Corpus é a observação de dados empíricos de uma ou mais línguas – ou variedades de língua - armazenados em bancos de dados que compõem um corpus, com a utilização de ferramentas eletrônicas especialmente desenvolvidas para auxiliar o pesquisador na análise dos dados, facilitando assim o seu trabalho quanto à verificação dos fenômenos da língua em uso.

As pesquisas desenvolvidas pela Lingüística de Corpus apresentam, também, outras características que Biber (1998: 4)[3] aponta como fundamentais:

São empíricas, analisando os padrões reais de uso da língua em textos naturais;

[3] - It is empirical, analyzing the actual patterns of use in natural texts;
- It utilizes a large and principled collection of natural texts, known as a 'corpus', as the basis for analysis;
- It makes extensive use of computers for analysis, using both automatic and interactive technique;
- It depends on both quantitative and qualitative analytical techniques.

Utilizam extensa e criteriosa coleção de textos naturais como base de análise – o *corpus*;

Fazem uso extenso de computadores em suas análises, recorrendo a técnicas tanto automáticas quanto interativas;

Apóiam-se em técnicas de análise tanto qualitativas quanto quantitativas. (tradução da autora);

Incorporando as características acima listadas, Berber Sardinha (2004:18) considera a definição de *corpus* proposta por Sanchez (1995: pp. 8-9) como a mais completa:

> "Um conjunto de dados lingüísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos em amplitude e profundidade, de maneira que sejam representativos da totalidade do uso lingüístico ou de algum de seus âmbitos, dispostos de tal modo que possam ser processados por computador, com a finalidade de propiciar resultados vários e úteis para descrição e análise".

Ao justificar sua preferência pela definição de Sanchez, Berber Sardinha (2004:18-9) aponta aspectos nela incluídos de capital relevância para a análise acurada de um corpus, tais como:

- A origem: os dados devem ser autênticos;
- O propósito: sua finalidade é servir de objeto ao estudo lingüístico;
- A composição: o conteúdo deve ser criteriosamente escolhido;
- A formatação: os dados devem ser legíveis por computador;
- A representatividade: o corpus deve ser representativo de alguma língua, ou variedade de língua";
- A extensão: para fins representativos, o corpus deve ser vasto.

A utilização adequada de um corpus, por meio de ferramentas computacionais, possibilita aos estudiosos da linguagem encontrar provas da ocorrência de um dado fenômeno da língua, bem como determinar a freqüência de cada ocorrência sob estudo.

Dentre os tipos de corpora utilizados pela Lingüística de Corpus, podemos destacar o *corpus geral* e o *corpus especializado*. O *corpus geral,* também conhecido como *corpus de língua geral*, compõe-se de uma coletânea de textos com a finalidade de permitir pesquisa de uma determinada língua, sem levar em conta distinções genéricas, varietais, dialetais, lexicais, etc. Sinclair (1991) refere-se ao corpus geral como "um tipo de corpus que armazena uma enorme quantidade de dados e está em constante atualização da língua-alvo". Entretanto, no caso de interesse de análise de determinada variedade, como o internetês, ou domínio específico, como a linguagem jurídica, esse tipo de *corpus* não é apropriado e nesse caso opta-se pela compilação de um *corpus especializado.*

O *corpus especializado*, por sua vez, é desenvolvido para atender às necessidades específicas de um trabalho de pesquisa em particular, de acordo com os objetivos propostos, como, por exemplo, uma coletânea de textos de e-mails reunidos com o propósito de verificar a variedade lexical e gramatical no gênero. Vários corpora especializados têm sido coletados com propósitos e conteúdos distintos, segundo os objetivos da pesquisa, e portanto diferenciados pelo modo, tempo, seleção, conteúdo, autoria, disposição interna e finalidade (Berber Sardinha, 2004:20-21).

1.1.1 Lingüística de Corpus e Corpora: histórico e pesquisas

O desenvolvimento das pesquisas lingüísticas que se utilizam da Lingüística de Corpus está intimamente ligado à evolução tecnológica, mais especificamente à existência do computador eletrônico. Entretanto, desde a Grécia antiga, passando pela Idade Média até o início do século XX, já se produziam pesquisas baseadas em corpus. Essas pesquisas eram realizadas em documentos 'manuscritos', adotando o sentido primário da palavra *corpus*, ou seja, uma grande quantidade de documentos. Obviamente, em vez de eletrônicos, os corpora eram coletados, armazenados e analisados manualmente, recorrendo muitas vezes a um grande contingente de

especialistas para a análise e verificação quanto à ocorrência ou não dos fenômenos estudados em uma dada língua (Berber Sardinha, 2004).

Em 1959, Randolf Quirk e sua equipe compilaram o corpus SEU (*Survey of Englih Usage*) contendo 1 milhão de palavras, as quais foram separadas em fichas individuais e analisadas gramaticalmente. Somente em 1989 – ou seja, 30 anos depois - o SEU foi totalmente transformado em corpus eletrônico. Posteriormente, sob a coordenação de Jan Svartvik, da Lund University, Suécia, a parte falada do SEU converteu-se no SSE (*Survey of Spoken English)*. O corpus ficou então conhecido como *London-Lund Corpus (LLC)*, sendo considerado, nos anos 90, como o maior corpus falado em língua inglesa, totalizando cerca de 500 mil palavras.

O trabalho de Quirk e sua equipe contribuiu, posteriormente, para o desenvolvimento de etiquetadores computadorizados, além de servir de embasamento para a famosa obra *Compreehensive Grammar of the Englih Language de autoria de* Quirk, R.; Greenbaum, S.; Leech, G.; Svartvik, J. (1985) bem como para a formação de vários corpora, como o Brown, considerado um marco para o desenvolvimento da Lingüística de Corpus.

Em 1957, Noam Chomsky publicou a obra *Syntactic Structures,* divulgando a lingüística gerativista com sua visão racionalista da linguagem contrapondo-se, portanto, à visão empirista até então defendida por estudiosos que utilizavam corpora em suas pesquisas.

Chomsky (1957) e, por conseguinte, seus seguidores durante as décadas de 50 e 60, sugeriam que a utilização de corpora em pesquisas, devido à grande quantidade de textos cuja análise, por meios manuais, demandaria um contingente extraordinário de analistas, não seria confiável. Declarava, ainda, que um corpus constituído por dados empíricos em nada contribuiria para a análise lingüística, posto que o falante nativo poderia ter acesso às propriedades intrínsecas da língua por meio da intuição. Em outras palavras, a lingüística gerativista chomskyana recorria a julgamentos introspectivos dando enfoque aos agrupamentos sintáticos possíveis segundo o conhecimento que um falante nativo possui de sua língua.

Sinclair (1996 apud Berber Sardinha, 2004:32), rebatendo as considerações de Chomsky, sustenta que:

> "...o que o falante nativo pode informar é somente se o traço ou estrutura em questão é intuitivamente possível ou não, pois o ser humano, ao contrário do que em geral se pensa, não é bem organizado para isolar conscientemente o que é central e típico da linguagem; aquilo que é incomum é percebido imediatamente, mas os eventos costumeiros são apreciados subliminarmente".

Por outro lado, a posição empirista defendida por Halliday (1991) descreve a linguagem como um sistema probabilístico, diferenciado por contextos nos quais os falantes empregam a linguagem (Berber Sardinha, 2000a, 2004). De modo complementar, a abordagem empirista fundamenta-se na observação de ocorrências de traços ou estruturas por intermédio da freqüência de uso em contextos definidos.

Nessa perspectiva, Berber Sardinha (2004:15) sustenta que "a Lingüística de Corpus trabalha dentro de um quadro conceitual formado por uma abordagem empirista e uma visão da linguagem como sistema probabilístico". A partir dessa afirmação, é importante destacar aqui duas questões que consideramos altamente cruciais para o entendimento das pesquisas na área: em primeiro lugar, a importância do corpus como fonte de informações, visto que corresponde ao registro da linguagem natural efetivamente utilizada pelos usuários da língua em situações reais; em segundo lugar, a importância da investigação das freqüências de traços lingüísticos (léxico, sintáticos, semânticos e discursivos), pois a comprovação da freqüência atestada é que levará o pesquisador à probabilidade teórica (Berber Sardinha, 2004).

Tendo em vista as considerações aqui apresentadas, recorremos novamente a Biber (1998:11) quando sustenta que "a Lingüística de Corpus pode ser aplicada em investigações empíricas em qualquer área lingüística[4]". Corroborando tal assertiva, inúmeras pesquisas têm sido desenvolvidas na área, confirmando a potencialidade da utilização da Lingüística de Corpus em diversas áreas, tais como a lexicografia, a sociolingüística, a semântica, entre outras.

Além dos estudos acima mencionados, vale ressaltar que a disponibilidade de corpora eletrônicos vem contribuindo de forma significativa para o desenvolvimento de centros de pesquisas baseadas em corpus e no processamento de linguagem natural.

[4] The corpus-based approach is that it can be applied to empirical investigations in almost any area of linguistic.

Alguns centros de pesquisa são mantidos em empresas, com o propósito de pesquisar o processamento automático de textos visando fornecer subsídios para a informatização de grandes bases de dados, bem como para a montagem de sistemas inteligentes de reconhecimento de voz e de gerenciamento de informação. Dentre as empresas que aplicam em pesquisas na área estão a Google e Microsoft.

1.1.2 Representatividade e Tamanho do Corpus

A representatividade de um corpus é assunto polêmico na Lingüística de Corpus. Ao citar a representatividade, remetemos ao sentido de extensão do corpus, isto é, quanto maior o corpus, mais representativo ele se torna. O ideal seria atingir uma representatividade total, inserindo nele a totalidade da língua.

Por outro lado, a compilação de um corpus representativo que abrangesse uma língua integralmente afigura-se tarefa praticamente impossível, devido à sua constante mutação. Outro ponto importante prende-se ao desconhecimento do tamanho da população[5]; torna-se, portanto, bastante remota a possibilidade de determinar o tamanho de um corpus ideal que represente essa população.

A melhor forma de compilar um corpus representativo é através do acúmulo de uma grande quantidade de palavras, textos, registros e gêneros, procurando tornar a amostra a maior possível, aproximando-a assim do número máximo da população da qual deriva (Berber Sardinha 2004:23). Em outras palavras, o corpus deve ser o maior possível a fim de que possa representar um determinado tipo de variedade de linguagem, funcionando como amostra desta (Berber Sardinha 2004:23). Nesse contexto, Sinclair (1997, apud Berber Sardinha, 2004) sustenta que “um corpus deve ser tão grande quanto à tecnologia permitir, daí a grande variedade no tamanho dos corpora existentes”.

Um aspecto relevante para constituição de um corpus representativo é, por conseguinte, o seu tamanho. Apesar disso, a definição de critérios para a determinação do tamanho ideal de um corpus ainda é aspecto pouco pesquisado. Berber Sardinha (2000a), sob uma perspectiva histórica, sugere uma classificação

[5] População, neste contexto, é entendida como a linguagem (Berber Sardinha, 2004:23).

baseada na observação dos corpora utilizados em quatro anos de conferências de Lingüística do Corpus. Essa classificação foi realizada com base na monitoração dos corpora efetivamente empregados pela comunidade de lingüistas, segundo estudos apresentados nas principais conferências de Lingüística de Corpus (ICAME 1995; ICAME 1996; ICAME 1997; PALC 1997; TALC 1998):

Tamanho em palavras	Classificação
Menos de 80 mil	Pequeno
80 a 250 mil	Pequeno-médio
250 mil a 1 milhão	Médio
1 milhão a 10 milhões	Médio-grande
10 milhões ou mais	Grande

Quadro 1: classificação do tamanho do corpus, segundo Berber Sardinha (2000a).

De acordo com o quadro 1, um corpus considerado pequeno é formado por até 80 mil palavras, um médio até 1 milhão, até chegar a um corpus grande – composto de 10 milhões de palavras ou mais.

Outro aspecto importante, também associado à representatividade, é a questão da probabilidade, relacionada à maior ou menor freqüência de uso de determinadas palavras. Segundo pesquisas na área (Berber Sardinha, 2004, 2000a), a maior parte das palavras que compõem uma dada língua ocorre em baixíssima freqüência. Para que haja a probabilidade de essas palavras ocorrerem no corpus, é necessário construir um corpus composto de grande quantidade de palavras; nas palavras de Berber Sardinha (2004:23), “quanto maior a quantidade de palavras, maior a probabilidade de aparecerem palavras com baixa freqüência”.

O sentido das palavras também está diretamente relacionado à representatividade. As formas com freqüência alta muitas vezes ocultam vários sentidos. Entre os vários sentidos podemos distinguir quais os sentidos mais freqüentes são atribuídos às formas e quais os menos freqüentes. Assim, mesmo as formas que apresentam alta freqüência têm sentidos raros (por exemplo, 'serviço'

entendido como 'saque' no jogo de tênis) e, portanto, esses sentidos terão maior probabilidade de ocorrer quanto maior for o corpus (Berber Sardinha, 2004). Dessa forma, “para que seja representativo, um corpus deve conter o maior número possível de sentidos de cada forma” (Berber Sardinha, 2004:24).

1.1.3 Freqüência

Para a investigação da linguagem do ponto de vista probabilístico, é preciso ter em mãos um instrumento que nos permita aferir essa probabilidade. O instrumento mais básico, nesse sentido, é a lista de freqüência de palavras, que registra cada palavra e sua ocorrência no corpus. A lista de freqüência de palavras é valiosa para o estudo da Lingüística de Corpus, pois revela a ocorrência de cada palavra naquele corpus específico, além de listar todas as palavras que o compõem. Partindo destes dados, pode-se então determinar quais palavras são mais ou menos freqüentes no corpus sob estudo.

O levantamento de listas de freqüência foram feitos por pesquisadores em várias línguas. Em língua portuguesa, destacamos o Frequency Dictionary of Portuguese Words, elaborado por Ducan (1972) como parte de sua tese de doutorado em Standford, “a partir de um corpus de português europeu com 500 mil palavras” (Ducan 1972, apud Berber Sardinha, 2004:161).

No Brasil, as primeiras listas de freqüência foram realizadas por Maria Tereza Camargo Biderman, uma das responsáveis pela compilação do Corpus do Português Brasileiro Contemporâneo da Universidade Estadual Paulista (Unesp), considerado em 2001 como o maior banco textual do português brasileiro (Época, 2001). Este trabalho serviu de subsídio para a preparação do Dicionário Didático de Português (1998), com especial ênfase à explicitação dos significados a partir dos usos lingüísticos contextualizados.

1.1.4 Concordâncias

Uma prática freqüente dos Lingüistas de Corpus é a observação das palavras em seu contexto original. Esse procedimento possibilita evidenciar claramente o significado e/ou o uso contextualizado das palavras.

Para tanto, é necessário recorrer a um instrumento essencial, ou seja, à lista de concordâncias. Para melhor entendimento e verificação de exemplos, vide seção 2.6.1, capítulo 3.

Berber Sardinha (2004:187) define concordância como "uma listagem das ocorrências de um item específico, dispostas de tal modo que a palavra de busca (aquela que se tem interesse em investigar) aparece centralizada na página (ou na tela do computador). A palavra de busca é acompanhada pelo seu cotexto original, isto é, pelas palavras que ocorrem junto com ela no corpus". A palavra de busca, ou a palavra que se deseja investigar, também pode ser chamada de nódulo. O conceito de concordância não deve ser confundido com o sentido do termo em relação a concordâncias gramaticais (verbal e nominal).

Outras definições são igualmente essenciais para o entendimento das análises que utilizam concordâncias, ou seja, os colocados. Assim, de acordo com Berber Sardinha (2004:188), "colocados são palavras que ocorrem ao redor do nódulo ou da palavra de busca, em posições relativas (primeira à esquerda, segunda à esquerda); diferem, portanto, de 'palavra de contexto' pois esta é opcional, definida pelo usuário no momento da busca". Os colocados, contudo, são todas as palavras que ocorrem perto do nódulo, dentro de um horizonte especificado incluindo as palavras de busca que existirem.

1.1.5 Padrões da Linguagem: conceituação e estudos

A determinação de padrões de linguagem pode contribuir de forma relevante para a investigação da padronização do léxico, ou da léxico-gramática. Dessa forma, padrões podem ser definidos "como todas as palavras e estruturas com as quais são regularmente associados e que contribuam para seu significado. Um padrão pode ser

identificado se uma combinação de palavras ocorre com relativa freqüência, se é dependente de uma palavra específica, e se há um significado claro associado" (Berber Sardinha 2004:39).

Segundo Berber Sardinha (2004:40), dentre os objetivos propostos pelas pesquisas que privilegiam a padronização busca-se responder a questões tais como:

1. Quais os padrões lexicais dos quais a palavra faz parte?
2. A palavra associa-se regularmente com outros sentidos específicos?
3. Em quais estruturas ela aparece?
4. Há uma correlação entre o uso/sentido da palavra e as estruturas das quais ela participa?
5. A palavra está associada com certa posição na organização textual?"

O estudo da padronização apóia-se teoricamente no princípio idiomático (*idiom principle*), segundo o qual o usuário de uma língua tem a sua disposição "um grande número de frases pré ou semiconstruídas que se constituem em escolhas únicas muito embora pareçam analisáveis em segmentos[6]" (Sinclair, 1987:320).

Assim, quando falamos em padrão léxico-gramatical, pressupomos que haja um espaço comum formado pelo léxico e pela sintaxe, destruindo dessa forma a dicotomia entre o léxico e a gramática (Sinclair, 1987). Em outras palavras, a escolha de um item lexical na língua implica a diminuição das escolhas dos itens lexicais e das categorias gramaticais que podem compô-lo. Da mesma forma, a escolha de uma classe gramatical também limita a escolha possível de classes gramaticais e itens lexicais que podem segui-la.

Nesta visão, a Lingüística de Corpus descreve as probabilidades de certos itens ocorrerem em co-textos específicos tornando supérflua, por conseguinte, a separação entre os níveis do léxico e da gramática, sendo esta uma questão de conveniência analítica, sem apoio empírico (Sinclair, 1991).

---

[6] "a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analyzable into segments."

1.1.6 Lingüística de Corpus e a visão probabilística da linguagem

A probabilidade lingüística é um tema que causa muitas discussões, devido ao fato de que alguns lingüistas ainda recorrem à tradicional prática estruturalista, influenciados pela teoria chomskyana.

Voltando a Chomsky (1957), reputamo-lo como o principal responsável pela crítica radical acerca da não-contribuição da probabilidade de uma sentença para o entendimento da linguagem. O exemplo utilizado pelo autor – acrescido de observação sarcástica – apresentava duas orações: '*I live in New York'* (Moro em New York) e *'I live in Ohio'* (Moro em Ohio). Baseando-se na comparação entre as duas declarações, Chomsky alegava que, por ser a primeira a mais freqüente, a concepção de que a relativa freqüência de um texto não se sustentava; em outras palavras, embora reconhecendo que a freqüência pudesse ter alguma significância teórica, não era suficiente para comprovar a importância de alguns tipos de dados quantitativos na análise lingüística; acrescentava, ainda, que um corpus em nada contribuiria para um melhor entendimento da linguagem, afirmando que comportamentos lingüísticos observados a partir de corpus poderiam ser afetados por variáveis, além da estrutura intrínseca peculiar da língua falada (Halliday, 1992).

A repercussão das idéias gerativistas e sua difusão e adoção em diversas comunidades acadêmicas, aliada às veementes críticas chomskyanas contra a visão probabilística da linguagem, revelaram-se bastante danosas ao avanço das pesquisas ligadas à questão das probabilidades; assim colocada em posição marginal, a visão probabilística da linguagem passou por um período de desprestígio, conseqüentemente refletido na redução das pesquisas da área.

Apesar do quadro sombrio aqui sintetizado, estudos baseados em probabilidades não cessaram de todo (Svartvik, 1966 apud Halliday, 1991). Pouco a pouco, verifica-se um número crescente de estudiosos que vêm adotando esse paradigma em suas pesquisas lingüísticas, tais como Biber (1998), Sinclair (1987, 1991), Halliday (1991, 1992, 1993), Berber Sardinha (2000, 2004) e outros.

Um dos maiores expoentes da visão probabilística é Halliday (1991, 1992, 1993), a partir de estudos sobre probabilidades em sistemas lingüísticos em meados de 1950,

quando preparava uma gramática da língua chinesa. Dez anos mais tarde, o autor volta-se para a língua inglesa, levando em conta as probabilidades de ocorrência de traços gramaticais ao tomar como amostras 2.000 orações de diferentes registros. Seu principal objetivo era o de descrever os sistemas não somente como escolhas *a* ou *b* ou *c*, mas o de verificar como as escolhas *a* ou *b* ou *c* estão ligadas a certas probabilidades de ocorrência (Berber Sardinha, 2006). Para Halliday, era evidente que alguns traços do sistema, tais como as 'polaridades' negativa/positiva ou os 'tempos verbais' presente/passado, poderiam ser mensurados. As 2.000 orações tomadas como amostra não formaram, porém, um corpus suficientemente grande que permitisse a conclusão do seu trabalho. Para um resultado fidedigno, centenas de milhares de orações – e não somente milhares de orações – seriam necessárias.

Em 1991, alguns pesquisadores - entre eles Halliday - dedicaram-se à compilação do Cobuild, grande corpus eletrônico da língua inglesa, projeto desenvolvido pela Universidade de Birmimgham[7]. Finalmente, foi então possível concluir o trabalho que permanecera inacabado. Para tanto, Halliday comparou os resultados das 2.000 orações iniciais aos grandes dados constantes do corpus Cobuild. O resultado foi surpreendente pois, ao analisar as 'polaridades' no corpus não-eletrônico, o autor encontrou a razão 0.9 em sentenças positivas e 0.1 em negativas (leia-se probabilidade de 90% para positivas e 10% para negativas); no corpus eletrônico, em contrapartida, Halliday obteve 0.87 para as positivas e 0.13 para as negativas. Quanto à análise dos tempos verbais, o primeiro corpus revelou 0.5 sentenças no presente e 0.5 no passado, ao passo que, na amostra do Cobuild, chegou-se a 0.4955 no presente e a 0.5041 no passado; sendo que tais variações (diferenças) foram consideradas como diferenças não-significativas (Bod, 2003:12 apud Berber Sardinha, 2006). Tal resultado apontou para uma tendência de distribuição de probabilidades das polaridades correspondente a 0.9 para 0.1 (sentenças positivas e negativas), e a 0.5 para 0.5 para os tempos verbais. Halliday nomeou tais distribuições 'skew' (enviesada) para os valores 0.9 para 0.1, e 'equiprobable' (equiprovável) para os resultados próximos de 0.5 para 0.5 (Halliday 1993:9). Conforme dito acima, os resultados também podem ser expressos em

[7] Grã-Bretanha.

porcentagem (Bod, 2003:12 *apud* Berber Sardinha, 2006). Desta forma, e tomando como exemplo os valores encontrados por Halliday (1993), as polaridades referentes às sentenças podem ser representadas como 90% positivas e 10% negativas; quanto aos tempos verbais, as polaridades revelam-se semelhantes, ou seja, 50% no presente e 50% no passado.

As pesquisas acima descritas mostram, dessa forma, que o fato de as probabilidades fazerem parte dos sistemas lingüísticos altera o entendimento do conceito de 'escolha' na língua em uso, revelando, em conseqüência, que a "livre escolha" é algo geralmente pouco provável (Halliday, 1993). Se assim fosse, os usuários de uma língua poderiam 'escolher' sentenças negativas às positivas, o que não ocorre normalmente.

Halliday (1993) destaca, ainda, dois pontos importantes a serem considerados na pesquisa probabilística: (1) As probabilidades mudam de acordo com o tempo, e essa mudança ocorre diacronicamente, de forma dinâmica; (2) os padrões de probabilidade variam de acordo com o número de diferentes situações. Por exemplo, a probabilidade de alguns traços em registros específicos normalmente varia se for comparada à língua em sua totalidade.

As pesquisas realizadas por Halliday (1991,1992,1993) somam-se a inúmeras outras que confirmam o potencial das probabilidades no enriquecimento das descrições lingüísticas (Biber et al, 1998), refletido no desenvolvimento de vários projetos em andamento, muitos deles voltados para o processo de ensino e aprendizagem.

1.1.7 Descrição lingüística: Qualitativa ou Quantitativa?

A tradicional prática de análises lingüísticas consiste em descrever traços gramaticais, léxicos, semânticos, entre outros; porém, ao discutirmos probabilidade e freqüência, uma questão se apresenta: trata-se, aqui, de pesquisa qualitativa ou quantitativa?

Segundo Biber (1998:8), as “técnicas quantitativas são essenciais para a pesquisa baseada em corpus[8]”. Em outras palavras, para sabermos como as palavras comportam-se em um determinado contexto, é preciso recorrer a métodos quantitativos que nos revelarão quais palavras são mais ou menos usadas numa dada língua, ou que tipos de padrão são mais freqüentemente empregados. Como ilustração, tomemos os adjetivos “grande” e “extenso”. Se a nossa finalidade, ao observarmos esses adjetivos, for a de comparar os padrões da língua em uso, duas questões devem ser respondidas: (1) Quantas vezes cada palavra ocorre no corpus? (2) Quais palavras co-ocorrem com freqüência perante tais adjetivos? Nesta última questão, estamos nos referindo aos colocados. Para responder a estas indagações, teremos que recorrer a dados quantitativos. Nesse aspecto, Halliday (1992:61) lembra que “os métodos quantitativos podem ser usados para estabelecer graus de associação entre diferentes sistemas gramaticais[9]”.

O método quantitativo respalda e enriquece as análises lingüísticas, pois o domínio da freqüência permite-nos estimar a probabilidade de ocorrência**.** No presente trabalho, optamos por utilizar método de natureza quantitativa; o material selecionado para análise será, portanto, avaliado em função da variável analisada com base nos objetivos centrais deste estudo. Ressaltamos, todavia, que embora a probabilidade esteja diretamente associada à quantificação, não é nosso propósito limitar-nos ao simples levantamento de dados quantitativos. Segundo Chizzotti (2001:98), o olhar qualitativo busca “compreender criticamente o sentido das comunicações, seu conteúdo manifesto ou latente, as significações explícitas ou ocultas”. Por conseguinte, o aspecto qualitativo emergirá naturalmente, durante o próprio levantamento e ao longo da análise dos dados, contribuindo para uma melhor descrição da linguagem-alvo deste trabalho - o *internetês*.

[8] Quantitative techiniques are essential for corpus-based studies.
[9] How quantitative methods could be used to establish degrees of association between different grammatical systems.

Capitulo 2 – Metodologia

Neste capítulo, propomo-nos inicialmente a reiterar o objetivo da pesquisa e a enumerar as questões que a norteiam. Em seguida, passaremos a detalhar a metodologia utilizada, descrevendo o corpus de estudo, as ferramentas utilizadas para análise e os dicionários adotados na investigação dos itens lexicais. Finalmente, apresentaremos os procedimentos de análise utilizados para a obtenção dos resultados constantes do Capítulo 3.

2.1 Objetivos e Questões de Pesquisa

2.1.1 Objetivos

Os objetivos deste trabalho são o de, primeiramente, utilizar os procedimentos e noções oferecidas pela Lingüística de Corpus para detectar as modificações gráficas que se verificam no internetês. Em um segundo momento, objetiva-se extrair as freqüências das formas dos itens lexicais e desta forma verificar os índices de freqüência das modificações observadas na linguagem da internet, visando estabelecer uma possível padronização dessa linguagem.

2.1.2 Questões da pesquisa

No decorrer da investigação, pretendemos responder às seguintes questões:

1. Quais palavras são encontradas com maior freqüência no corpus de estudo?

2. Quais palavras caracterizam a utilização do internetês?

3. Quais as modificações que ocorrem com maior freqüência na formação das palavras do internetês?
4. Quais padrões léxico-gramaticais são mais comumente verificados no internetês?

Para responder às questões acima relacionadas, tivemos acesso a considerável quantidade e diversidade de dados disponibilizados pela Lingüística de Corpus, o que nos permitiu investigar o sentido das palavras em contexto. Optamos por utilizar, como objeto de estudo, um corpus constituído de *blogs* de jovens que utilizam a internet como meio de comunicação.

Os procedimentos de análise detiveram-se na distinção das diferentes grafias, além de verificar, empiricamente, a freqüência, a maneira de uso e uma série de outros dados sobre os itens lexicais, sem deixar de considerar os padrões léxico-gramaticais do internetês.

2.2 Procedimentos para a coleta do corpus

2.2.1 Corpus de Estudo

O corpus objeto de estudo consiste em uma coletânea de blogs de jovens que se utilizam do internetês como linguagem de comunicação virtual. Notadamente, os blogs[10] têm características particulares: são páginas personalizadas, apresentando-se como diários públicos, nos quais são descritas experiências vividas diária ou periodicamente. Incluem, ainda, 'caixas' de comentários ou mensagens, onde amigos, familiares e pessoas em geral (usuários/visitantes) têm a opção de deixar mensagens, comumente também escritas em internetês. A coleta estendeu-se, portanto, não apenas às mensagens escritas pelos construtores dos blogs, mas também aos comentários postados pelos usuários/visitantes.

Vale salientar que o próprio blog facilita a coleta de corpus por meio de atalhos ou links, que por sua vez funcionam como canal de acesso a outros blogs, com grafia similar própria do internetês. Como exemplo, citamos o "Blog Máfia" (http://www.fotolog.com/valcaintheroom) que disponibiliza acesso aos links de vários outros blogs – na sua maioria, com mensagens em internetês.

[10]A definição e características dos blogs constam do capítulo Introdução.

A coleta do corpus compreendeu um período de dois meses, ou seja, entre outubro e dezembro de 2005. Alguns critérios de seleção foram adotados a fim de facilitar a compilação do corpus, tais como:

1. restringir a coleta a blogs em língua portuguesa;

salvar apenas páginas pessoais que apresentassem grafia em internetês;

Ao final da coleta, os corpora selecionados e utilizados na presente pesquisa totalizaram 98 blogs[11], com um total de ocorrências (*tokens*) levemente superior a 135 mil palavras. Berber Sardinha (2004) define "*token* como o número de itens - ou ocorrências - em que uma palavra aparece no corpus". Como exemplo de *token*, destacamos a frase: 'O pai e o filho viajaram'. Seis itens compõem a frase: O (1); pai (2); e (3); o (4); filho (5); viajaram (6), em um total de 6 tokens.

A coleta de um corpus também considera o total de *types* nele verificados. Voltando a Berber Sardinha (2004), "*types* são o resultado da quantidade de formas, ou o número de vocábulos presente em um dado corpus". Para melhor entendimento, no exemplo: "O pai e o filho viajaram", a frase é composta por cinco formas: 2 formas para 'o'; 1 forma para 'pai'; 1 forma para 'e'; 1 forma para 'filho' e 1 para 'bola'; a frase é, portanto, composta de 5 *types.* Como mostra o quadro 2, o total de palavras do corpus revelou-se superior a 15 mil types.

ConteÚdo	Tokens	Types
98 blogs.	135.021	15.552

Quadro 2: Total de blogs, de formas lexicais e suas ocorrências.

Em termos de tamanho, podemos classificar nosso corpus pequeno-médio, segundo a escala proposta por Berber Sardinha (2004) com base em quatro anos de conferências em Lingüística de Corpus (vide quadro1, seção 1.1.3).

[11] A lista de endereços dos blogs, utilizados na presente pesquisa, está relacionada no anexo I.

Uma análise que se proponha a utilizar os procedimentos e noções desenvolvidas pela Lingüística de Corpus demanda um corpus de estudo que possa atender ao tipo de pesquisa pretendido, que é o de explorar uma dada língua ou variedade de língua, por meio de evidências empíricas. Para tanto, fará uso de programas computacionais que servirão de suporte nas tarefas mais complexas, a fim de garantir dados mais consistentes. A ferramenta computacional utilizada especificamente para esta pesquisa será apresentada na próxima seção.

2.2.2 WordSmith: ferramenta computacional

Para a análise do nosso corpus, utilizamos a ferramenta WordSmith, que é um conjunto de programas integrados (Suíte) destinados à análise lingüística.

O WordSmith foi desenvolvido em 1996 por Mike Scott e publicado pela Universidade de Oxford University Press[12], e está atualmente em sua quinta versão. Para esse trabalho faremos uso da versão 3, cuja licença foi-nos gentilmente cedida pela Instituição.

O WordSmith dispõe de uma série de recursos extremamente úteis e poderosos para a análise de vários aspectos da linguagem (Berber Sardinha, 2004:86). Dentre esses aspectos, podemos destacar a composição lexical, a temática de textos selecionados e a organização retórica e composicional dos gêneros discursivos (Berber Sardinha, 2004). A disponibilidade de ferramentas computacionais que disponibilizem programas flexíveis e de fácil utilização - como é o caso do WordSmith - pode levar a uma maior aplicabilidade desse tipo de ferramentas em análises lingüísticas (Berber Sardinha, 2004:85-86).

O software do WordSmith pode ser obtido pela internet[13], em duas versões – paga (completa) ou gratuita (demo), sendo que esta última restringe a análise a 25 linhas de ocorrências, no máximo. Se o usuário optar por adquirir o programa completo, ele deverá pagar uma licença de uso que lhe dará direito a uma senha, cuja digitação transformará a versão “demo” em completa.

---

[12] Grã-Bretanha.

[13] www.liv.ac.uk/~ms2928/ ou www.lexically.net

O WordSmith compõe-se de ferramentas - *WordList, Keywords, Concord* [14] -, além de utilitários, instrumentos e funções. As ferramentas são utilizadas, em geral, para a verificação do comportamento das palavras no corpus.

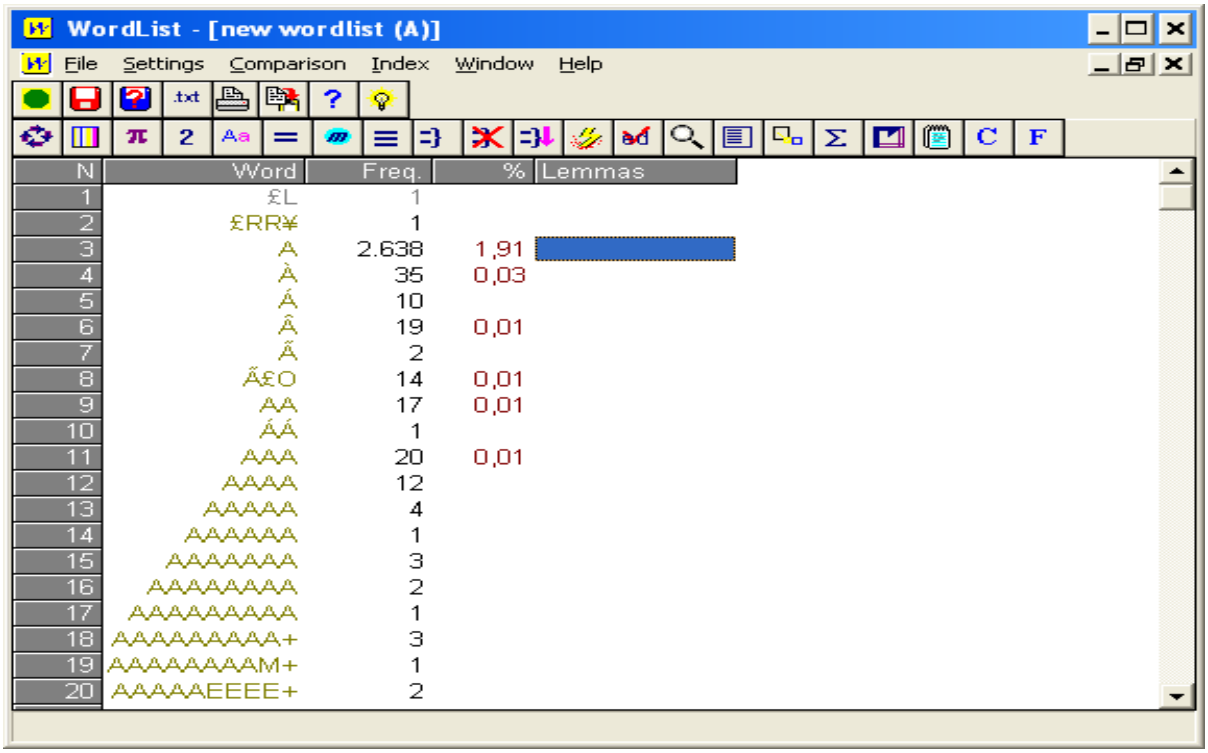
A seguir, explanaremos com mais detalhes as funções da WordList e do Concord utilizados na análise do nosso corpus. A KeyWords não foi utilizada posto que os nossos objetivos não incluíam a detecção de palavras-chave, tampouco a comparação de freqüências entre este corpus de estudo e um corpus de referência.

2.2.2.1 WordList

O programa WordSmith Tools oferece diversos tipos de estatísticas; para o presente trabalho, contudo, limitamo-nos apenas às estatísticas fornecidas pela ferramenta WordList, consideradas como fundamentais para o desenvolvimento da pesquisa.

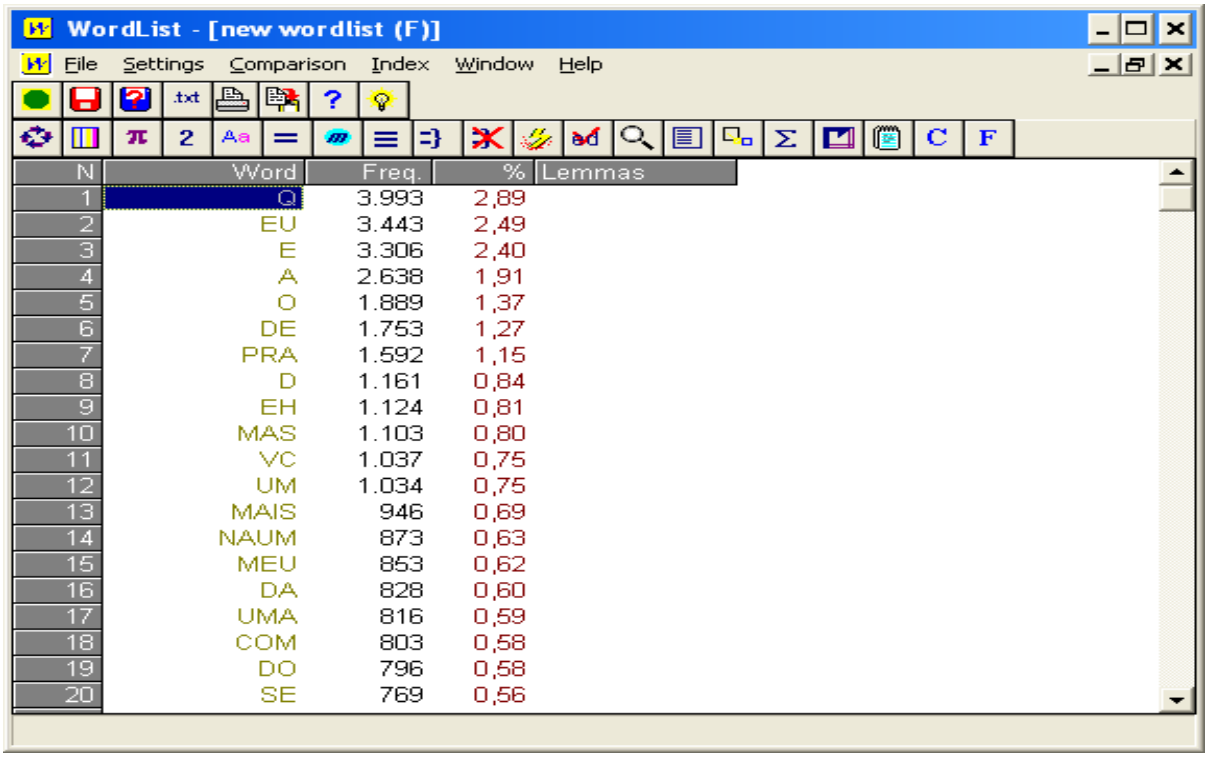
A ferramenta WordList propicia a criação de listas de palavras e, quando executada, exibe três janelas distintas: a primeira janela lista as palavras do corpus em ordem alfabética; a segunda mostra a mesma lista, ordenada por freqüência e encabeçada pela palavra com o maior número de incidências no corpus; e a terceira janela fornece a estatística relativa aos dados usados na produção da lista - *tokens* e *types*. As figuras 1 a 3 exemplificam os três tipos de configuração aqui descritos:

[14] Lista de palavras, palavras-chave, concordâncias.



N	Word	Freq.	%	Lemmas
1	£L	1		
2	£RR¥	1		
3	A	2.638	1,91	
4	À	35	0,03	
5	Á	10		
6	Â	19	0,01	
7	Ã	2		
8	Ã£O	14	0,01	
9	AA	17	0,01	
10	ÁÁ	1		
11	AAA	20	0,01	
12	AAAA	12		
13	AAAAA	4		
14	AAAAAA	1		
15	AAAAAAA	3		
16	AAAAAAAA	2		
17	AAAAAAAAA	1		
18	AAAAAAAAA+	3		
19	AAAAAAAAM+	1		
20	AAAAAEEEE+	2		

Figura 1 – Tela da WordList com as palavras em ordem alfabética.



N	Word	Freq.	%	Lemmas
1	Q	3.993	2,89	
2	EU	3.443	2,49	
3	E	3.306	2,40	
4	A	2.638	1,91	
5	O	1.889	1,37	
6	DE	1.753	1,27	
7	PRA	1.592	1,15	
8	D	1.161	0,84	
9	EH	1.124	0,81	
10	MAS	1.103	0,80	
11	VC	1.037	0,75	
12	UM	1.034	0,75	
13	MAIS	946	0,69	
14	NAUM	873	0,63	
15	MEU	853	0,62	
16	DA	828	0,60	
17	UMA	816	0,59	
18	COM	803	0,58	
19	DO	796	0,58	
20	SE	769	0,56	

Figura 2 – Tela da WordList com as palavras em ordem de freqüência.

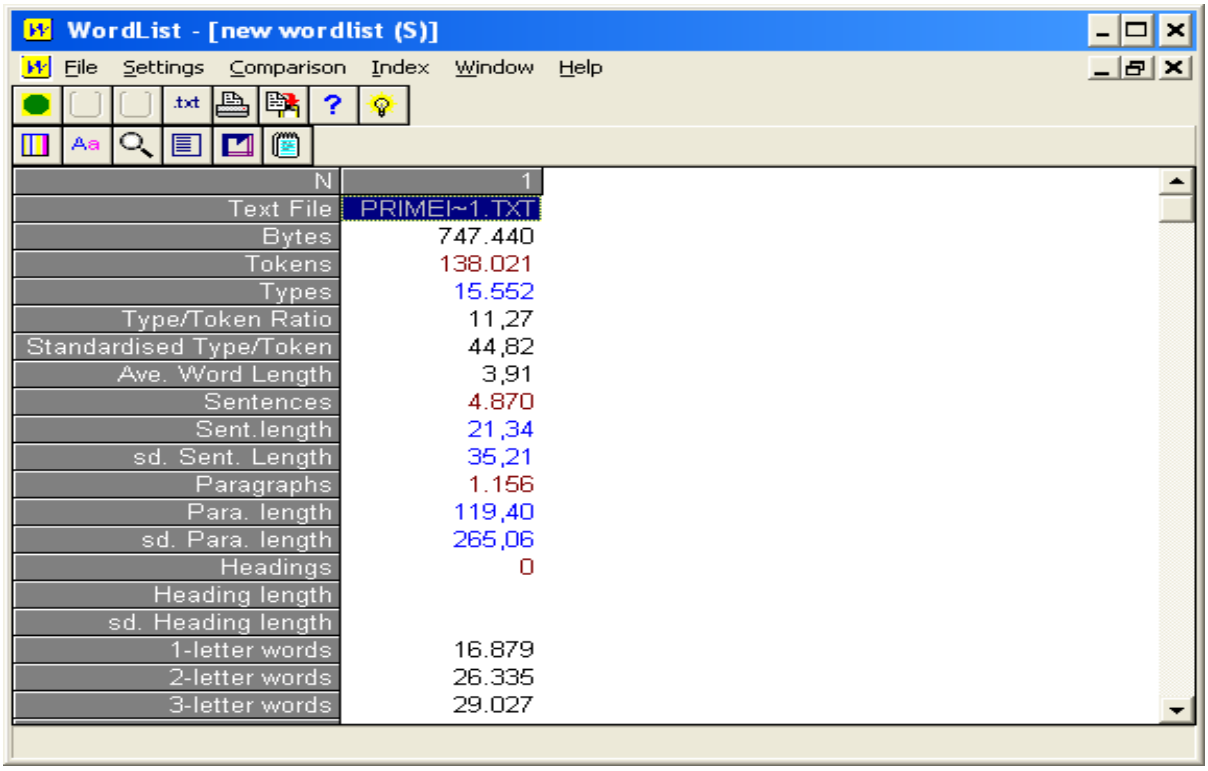


Figura 3 – Tela com as estatísticas fornecidas pela WordList.

A facilidade de uso da ferramenta está na economia de tempo e de trabalho por parte do pesquisador pois, enquanto os cálculos envolvidos – a contagem do número de palavras, ou das formas em um corpus constituído por vários textos - são executados em segundos, as mesmas contagens, feitas manualmente, demandariam extensa dedicação e considerável período de tempo, além da incerteza quanto à confiabilidade, comumente suscitada quando do manuseio de grande quantidade de dados.

Para melhor demonstrar o potencial de utilização da Lingüística de Corpus na viabilização e no tratamento analítico de bancos de dados extensos, optamos por selecionar um "recorte" com as 500 primeiras formas – ou types - mais freqüentes da lista de palavras sob investigação[15].

[15] A lista das 500 formas ou *types* mais freqüentes encontra-se no anexo II.

2.2.2.2 Concord

O próximo passo da pesquisa voltou-se para a produção das listas de 'concordância', termo definido por Berber Sardinha (2004) como "uma lista das ocorrências de um item específico". Para tanto, recorremos à ferramenta Concord que também integra o WordSmith Tools.

Na tela que exibe a lista de concordâncias, o item – ou nódulo - que se objetiva investigar destaca-se, centralizado e em tonalidade diferente, juntamente com as palavras co-ocorrentes – os colocados. Tal configuração permite-nos observar os usos distintos do mesmo item em contextos originais, bem como os diferentes sentidos a ele atribuídos.

Como já explicitado no item 1.1.4, os colocados são as palavras que cercam o nódulo (a palavra-alvo da pesquisa), tanto à direita quanto à esquerda, e que de alguma forma estão a ele associadas.

A importância da observação dos colocados no internetês reside, principalmente, no fato de estarem neles expressas as relações entre os textos e o uso da língua; em outras palavras, o sentido de um determinado termo em internetês depende do seu uso e dos colocados que o rodeiam em um dado contexto.

Dentre os recursos constantes da tela de Concordâncias está a coluna 'Set', de grande relevância para a nossa pesquisa. A coluna 'Set' foi utilizada para classificar as linhas de concordância por meio da diferenciação do sentido das ocorrências, utilizando para tanto as letras do alfabeto (a, b, c, d...), conforme o sentido existente em cada linha.

Para ilustrarmos tal recurso, destacamos a figura 4, cuja tela exibe, dentre outros dados, as linhas de concordância do item *'td'* (tudo), as posições de ocorrência do nódulo, os respectivos colocados e a coluna 'Set', com as classificações mencionadas.

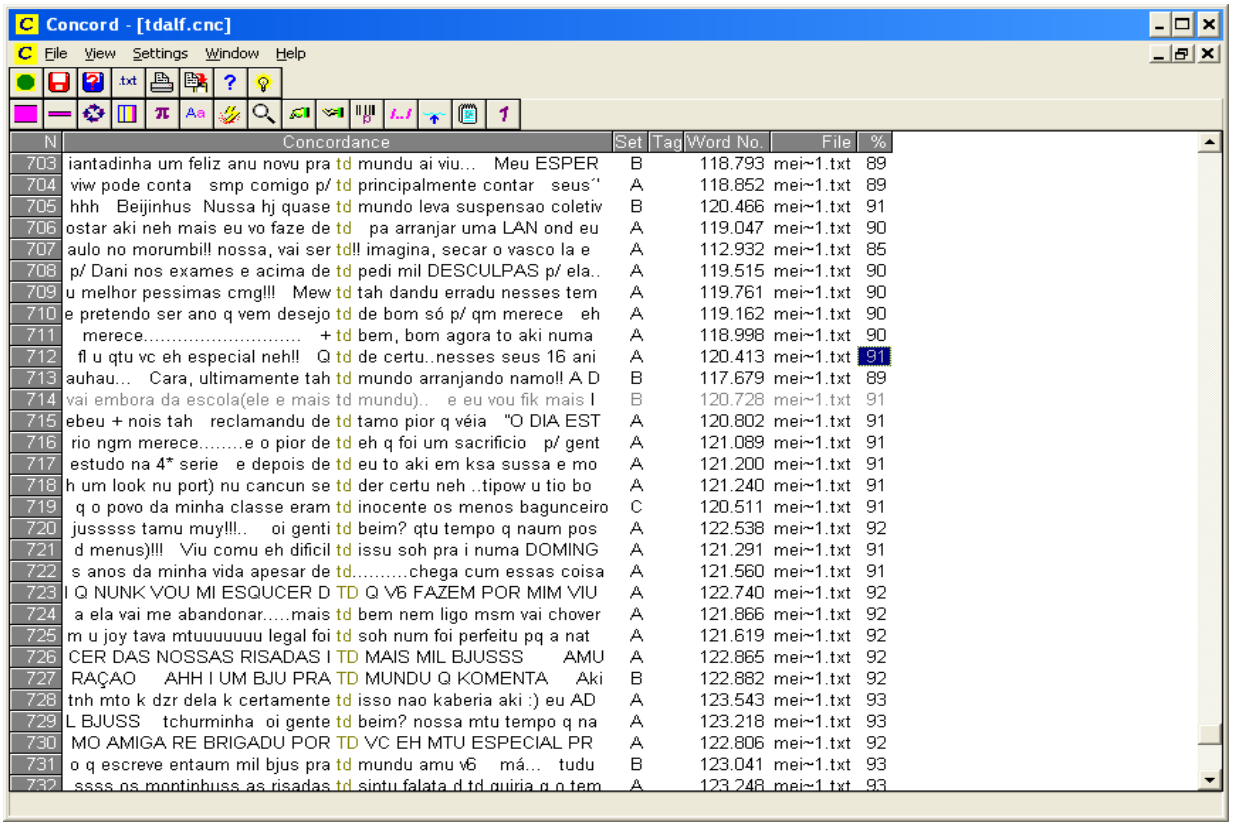


Figura 4 – Tela das Concordâncias do item '*td*', com a coluna 'Set' preenchida com as classificações.

Para exemplificar os diferentes sentidos do item selecionado para investigação (*'td'),* extraímos algumas linhas de concordâncias com as respectivas classificações, como mostra o quadro abaixo:

N	ConcordÂncias	Set
3	e resolvi posta akee \0/ okay marcelo ? =) td bem com vcs ? cmg ta td otimo !!X)	A
41	fim dia de folga aki... soh trampo a semana td, o dia inteiro... por isso q td mun	C
43	em escreve aki...entaum nem rola eu conta td as coisa neh....pq eh mta coisa.	E
49	e.. eu fiko vendo os meus amiguinho .. ele td falaum q saem com o papai de	D
515	ihkkikikikki....foi mto mico! Mto engraçado td mundo parou pra fikar olhando a	B

Quadro 3: diferentes sentidos do item 'td' e suas classificações.

O quadro 3 mostra-nos as posições de ocorrências no corpus retiradas com a finalidade de exemplificar os diferentes sentidos, quais sejam: tudo (A), todo (B), toda (C), todos (D) e todas (E).

Em internetês, vários itens apresentam diferentes sentidos, conforme o contexto onde aparecem. Em nossa investigação, optamos por solicitar à ferramenta que nos fornecesse a lista de concordância dos 500 itens mais freqüentes, com dois propósitos específicos:

1. Observar mais detalhadamente a relação entre o nódulo e seus colocados, permitindo-nos uma melhor compreensão do sentido de cada ocorrência. Como já mencionado anteriormente, é imprescindível a observação tanto dos itens estudados quanto das palavras que estão ao seu redor, pois no internetês a grafia de uma forma corresponde muitas vezes a vários sentidos ou a vários significados, que só se completam através dos colocados;

Buscar no corpus evidências que apontassem para uma possível padronização do léxico, da qual poderíamos extrair algumas respostas quanto ao estabelecimento de regularidades baseadas em diferentes tipos de associação, mesmo em se tratando do internetês. Tais pressupostos vão ao encontro da declaração de Berber Sardinha (2004:221) ao afirmar que “na língua, a menos que se prove o contrário, todas as palavras possuem padronização, escolhendo os padrões a que se associam, privilegiando alguns vizinhos e preterindo outros”.

Sob essa perspectiva e objetivando a observação dos padrões, seguimos os procedimentos detalhados a seguir. Primeiramente, determinamos um horizonte[16] de cinco palavras à direita e cinco palavras à esquerda da palavra-alvo; em outras palavras, estabelecemos a posição E5 (cinco palavras à esquerda do nódulo) até a posição D5 (cinco palavras à direta do nódulo) as quais incluem, conseqüentemente, as posições intermediárias E4, E3, E2, E1 – correspondentes respectivamente às

[16] O termo ‘horizonte’ pode ser entendido com a distância máxima entre a palavra de busca e a palavra de contexto.

posições 4, 3, 2, 1 à esquerda dos itens – e as posições D4, D3, D2, D1 – relativas às posições posicionadas à direita do nódulo.

Após os ajustes necessários, instruímos a ferramenta a fornecer as linhas de concordância do item selecionado, para uma verificação mais acurada dos padrões e respectivos sentidos. O horizonte permitiu-nos verificar os diversos sentidos daquele item lexical específico, constatar se havia regularidade nos tipos de associação a que se submetem as palavras do internetês e, finalmente, se os padrões exibidos contribuíam para o significado do item selecionado para o estudo.

2.3 Critérios para identificação de palavras do internetês

Para desenvolvermos a análise dos itens em internetês, foi preciso estabelecer alguns critérios que garantissem resultados confiáveis, posto que existem vários fenômenos e recursos gráficos envolvidos na composição dessa linguagem.

Como o presente estudo enfoca a grafia do internetês, optamos, então, por consultar o acervo lexical da Língua Portuguesa e compará-lo às formas do internetês. Para tanto, adotamos três dicionários tradicionais, de grande circulação nacional e constantemente atualizados, objetivando contemplar todos os vocábulos do nosso idioma e os sentidos a eles atribuídos. São eles:

1. “Dicionário Eletrônico Houaiss da Língua Portuguesa”: desenvolvido e atualizado pelo Instituto Antonio Houaiss; contém bibliografia vasta, abrangendo várias obras, datação e etimologia. A consulta aos verbetes pode ser realizada on-line, na página www.uol.com.br.

2. “Dicionário Aurélio Eletrônico” (versão em CD-ROM): elaborado inicialmente por Aurélio Buarque de Holanda Ferreira, é periodicamente atualizado sob a supervisão de Marina Baird Ferreira, viúva de Aurélio e responsável pelas edições do dicionário, juntamente com a professora Margarida dos Anjos. O dicionário é distribuído pela Editora ‘Positivo’ e conta com 435 mil verbetes,

definições, locuções e acepções que procuram acompanhar as alterações que ocorrem na língua ao longo do tempo.

3. “Michaelis Moderno Dicionário da Língua Portuguesa”: a elaboração do dicionário estendeu-se por dez anos e contou com a colaboração de 84 profissionais especializados; apresenta mais de 500.000 definições, distribuídas em mais de 200.000 verbetes e subverbetes. Seu planejamento consistiu em registrar o maior número possível de vocábulos, tanto da linguagem escrita quanto da oral. Esta obra baseia-se no banco de dados lexicográficos da ‘Melhoramentos’, e a consulta aos seus verbetes pode ser feita on-line, na página www.uol.com.br.

Os três dicionários acima foram utilizados com três propósitos: primeiramente, para verificar se os 500 itens mais freqüentes em nosso estudo faziam parte da grafia oficial da língua portuguesa; em seguida, para comparar os sentidos observados em contexto aos significados dicionarizados; e, finalmente, para detectar possíveis correspondências.

Além das fontes de consulta mencionadas, foi também necessário estabelecer algumas condições de identificação para os itens grafados em internetês, quais sejam:

1. Os itens em internetês não poderiam constar como palavras grafadas em Língua Portuguesa nos dicionários acima citados. Essa condição foi estabelecida de princípio, pois, ao consultarmos os dicionários mencionados, verificamos que incluíam vários itens idênticos a itens do internetês, porém com a observação de que correspondiam ao Tétum, língua falada no Timor Leste. Para ilustrar alguns desses exemplos, destacamos os itens ‘mundu’ (mundo) e ‘amu’ (amo), encontrados em verbetes do Dicionário Houaiss, mas definidos como termos do vernáculo Tétum. Tais exemplos foram, portanto, considerados como exemplos de internetês, visto que sua grafia não é reconhecida como integrante da norma-padrão da Língua Portuguesa;

2. Os itens em internetês não apresentassem recursos gráficos, como símbolos ou pontuação. Cabe observar, aqui, a existência de dois tipos básicos de recursos gráficos no internetês, dos quais esta pesquisa não tratará. O primeiro refere-se aos símbolos denominados *emoticons* ou *smileys,* usados para transmitir emoções ou sentimentos, visando assim dar às mensagens um cunho mais humano que possa ser facilmente perceptível pelo interlocutor. Por exemplo: para demonstrar alegria são freqüentemente utilizados os recursos **:-)** (dois pontos, hífen e parêntese de fechamento), os quais, observados em um ângulo de 90º à esquerda, assemelham-se a um rosto sorridente. Um outro tipo de recurso gráfico utilizado no internetês diz respeito ao emprego das pontuações; por exemplo, para demonstrar emoções, sentimentos ou dúvidas, os internautas pontuam palavras ou frases repetidamente, como em: *"num vai amanha!!!!!!!! pq???????"* (Não vai amanhã! Por quê?). Neste exemplo, os internautas demonstram surpresa diante de um determinado assunto através da repetição excessiva dos sinais gráficos;

3. Os itens em internetês não correspondessem a onomatopéias. Isto é, não correspondessem a figuras de linguagem representadas por palavras a partir da reprodução aproximada de um som a elas associado, através da utilização de alguns recursos disponíveis na língua (Faraco e Moura, 2000). Essa figura de linguagem é prática costumeira em textos produzidos pelos internautas. Habitualmente, os usuários empregam tais expressões para indicar gargalhadas e risos (hahaha, hehehe e kkkkk);

4. Os itens em internetês não correspondessem a abreviações ou acrônimos reconhecidos pela grafia padrão.

Em suma, o presente trabalho ater-se-á tão e somente às modificações na grafia de palavras do internetês.

2.4 Procedimento para a análise de dados

Na análise do corpus objeto desta pesquisa, utilizaremos os recursos dos programas WordList e Concord, mais especificamente a lista de palavras individuais com as respectivas freqüências, bem como as concordâncias e seus colocados. Como já mencionado no item 2.3.2, não utilizaremos os recursos da KeyWord, pois não é nosso objetivo comparar o corpus em questão com quaisquer corpora de referência, tampouco extrair palavras-chave do corpus de estudo.

Isto posto, os procedimentos de análise desenvolveram-se na seguinte seqüência:

1. Com o auxílio da ferramenta WordList, extraímos a lista de palavras em ordem de freqüência, isto é, ordenadas das mais freqüentes para as menos freqüentes, simultaneamente levantando as estatísticas sobre a freqüência de palavras e suas ocorrências. As palavras na WordList estão coligadas às concordâncias, podendo solicitá-las facilmente clicando no ícone [C] (concord) que se encontra na parte superior à direita da tela WordList;

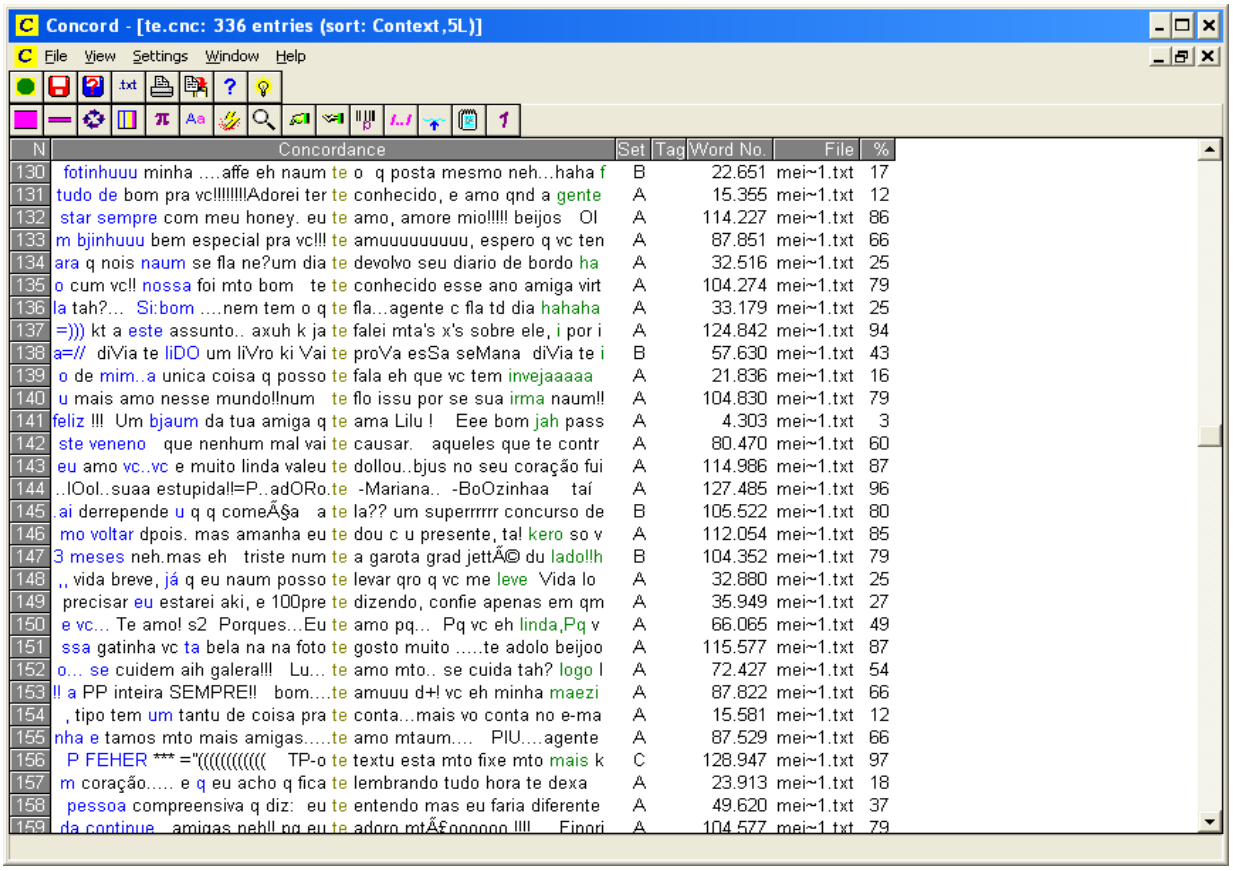
2. Por meio da ferramenta Concord, obtivemos as concordâncias através de uma lista contendo todas as ocorrências de uma dada palavra no contexto original. Cada lista corresponde a uma palavra investigada, destacada no centro de um fragmento de texto.

2.5 Análise de dados

2.5.1 Seleção da classe gramatical

Com o intuito de verificar algumas particularidades observadas nas formas do internetês, optamos, primeiramente, por levantar quais as classes gramaticais que apresentariam maior incidência de modificações.

Tendo em mente os itens em contexto e seus diferentes usos observados no corpus, buscamos classificá-los gramaticalmente utilizando, para isso, a lista de concordâncias, que nos forneceu indicações sobre quais funções gramaticais o item exercia em cada ocorrência. Vale salientar que muitos itens apresentaram duas ou mais classificações gramaticais, tanto na grafia padrão quanto no internetês. Um exemplo dessa ocorrência consta da figura a seguir:



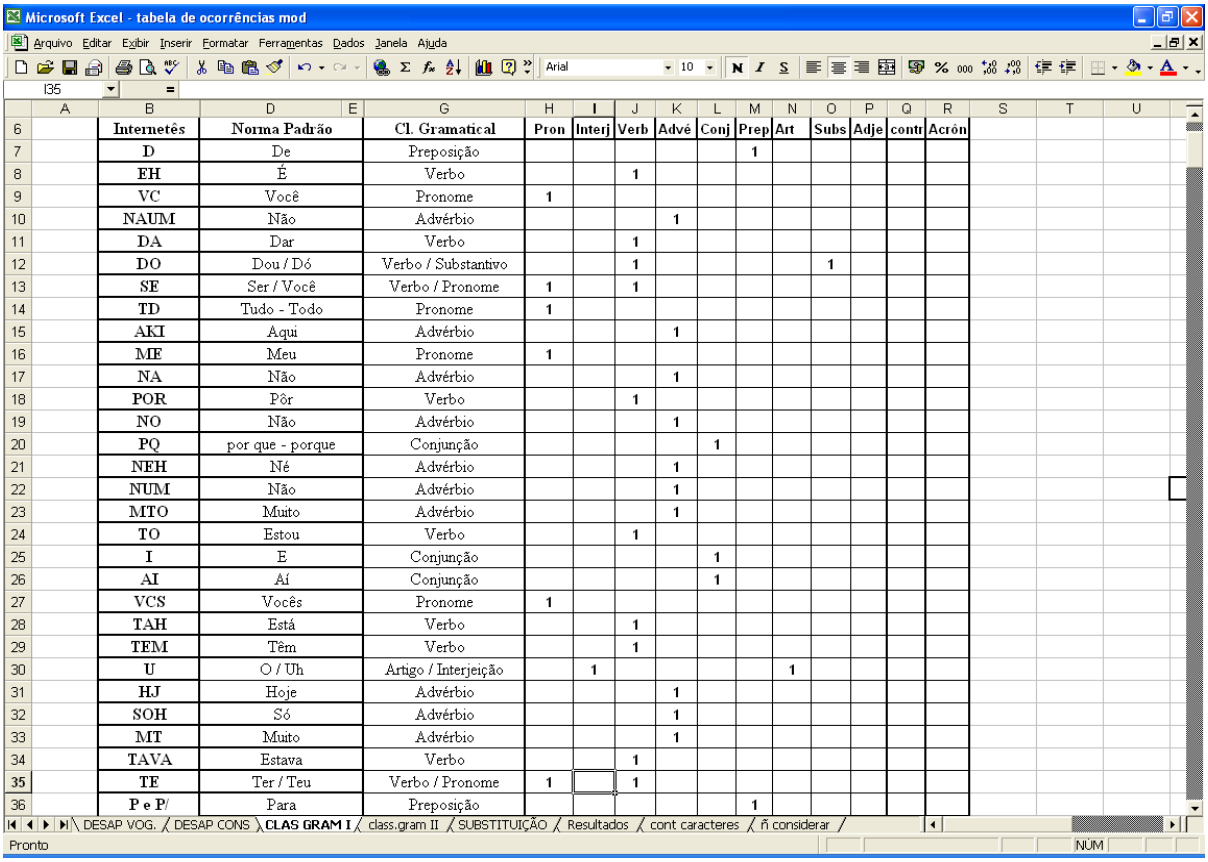
N	Concordance	Set	Tag	Word No.	File	%
130	fotinhuuu minha ....affe eh naum te o q posta mesmo neh...haha f	B		22.651	mei~1.txt	17
131	tudo de bom pra vc!!!!!!!!Adorei ter te conhecido, e amo qnd a gente	A		15.355	mei~1.txt	12
132	star sempre com meu honey. eu te amo, amore mio!!!!! beijos OI	A		114.227	mei~1.txt	86
133	m bjinhuuu bem especial pra vc!!! te amuuuuuuuuu, espero q vc ten	A		87.851	mei~1.txt	66
134	ara q nois naum se fla ne?um dia te devolvo seu diario de bordo ha	A		32.516	mei~1.txt	25
135	o cum vc!! nossa foi mto bom te te conhecido esse ano amiga virt	A		104.274	mei~1.txt	79
136	la tah?... Si:bom ....nem tem o q te fla...agente c fla td dia hahaha	A		33.179	mei~1.txt	25
137	=))) kt a este assunto.. axuh k ja te falei mta's x's sobre ele, i por i	A		124.842	mei~1.txt	94
138	a=// diVia te liDO um liVro ki Vai te proVa esSa seMana diVia te i	B		57.630	mei~1.txt	43
139	o de mim..a unica coisa q posso te fala eh que vc tem invejaaaaa	A		21.836	mei~1.txt	16
140	u mais amo nesse mundo!!num te flo issu por se sua irma naum!!	A		104.830	mei~1.txt	79
141	feliz !!! Um bjaum da tua amiga q te ama Lilu ! Eee bom jah pass	A		4.303	mei~1.txt	3
142	ste veneno que nenhum mal vai te causar. aqueles que te contr	A		80.470	mei~1.txt	60
143	eu amo vc..vc e muito linda valeu te dollou..bjus no seu coração fui	A		114.986	mei~1.txt	87
144	..lOol..suaa estupida!!=P..adORo.te -Mariana.. -BoOzinhaa taí	A		127.485	mei~1.txt	96
145	.ai derrepende u q q comeÃ§a a te la?? um superrrrrr concurso de	B		105.522	mei~1.txt	80
146	mo voltar dpois. mas amanha eu te dou c u presente, ta! kero so v	A		112.054	mei~1.txt	85
147	3 meses neh.mas eh triste num te a garota grad jettÃ© du lado!!h	B		104.352	mei~1.txt	79
148	,, vida breve, já q eu naum posso te levar qro q vc me leve Vida lo	A		32.880	mei~1.txt	25
149	precisar eu estarei aki, e 100pre te dizendo, confie apenas em qm	A		35.949	mei~1.txt	27
150	e vc... Te amo! s2 Porques...Eu te amo pq... Pq vc eh linda,Pq v	A		66.065	mei~1.txt	49
151	ssa gatinha vc ta bela na na foto te gosto muito .....te adolo beijoo	A		115.577	mei~1.txt	87
152	o... se cuidem aih galera!!! Lu... te amo mto.. se cuida tah? logo l	A		72.427	mei~1.txt	54
153	!! a PP inteira SEMPRE!! bom....te amuuu d+! vc eh minha maezi	A		87.822	mei~1.txt	66
154	, tipo tem um tantu de coisa pra te conta...mais vo conta no e-ma	A		15.581	mei~1.txt	12
155	nha e tamos mto mais amigas.....te amo mtaum.... PIU....agente	A		87.529	mei~1.txt	66
156	P FEHER *** ="(((((((((((TP-o te textu esta mto fixe mto mais k	C		128.947	mei~1.txt	97
157	m coração..... e q eu acho q fica te lembrando tudo hora te dexa	A		23.913	mei~1.txt	18
158	pessoa compreensiva q diz: eu te entendo mas eu faria diferente	A		49.620	mei~1.txt	37
159	da continue amigas neh!! pq eu te adoro mtÃ£ooooo !!!! Finori	A		104.577	mei~1.txt	79

Figura 5 – Linhas de concordâncias do item '*Te*' e suas classificações gramaticais.

A figura 5 revela-nos que, em contextos diversos, o item '*Te'* exerce diferentes funções, tais como: pronome (te) classificado sob a letra 'A'; verbo (ter) classificado sob a letra 'B'; novamente pronome (teu), classificado sob a letra 'C'; foram, portanto, identificadas classificações distintas para um único item. Vale ressaltar que no exemplo citado do internetês, o item *'Te'* tem grafia idêntica nos três casos ('te', 'ter' e 'teu'), enquanto o pronome 'te' é forma reconhecida pela norma padrão da língua.

Desse modo, ao verificarmos que algumas formas em internetês apresentaram mais de um sentido em contexto e alguns itens podem exercer também, diferentes funções gramaticais, optamos por contabilizar apenas a classe gramatical que apresentou uma freqüência maior de ocorrências, tanto para as formas em internetês, como para as palavras da norma padrão.

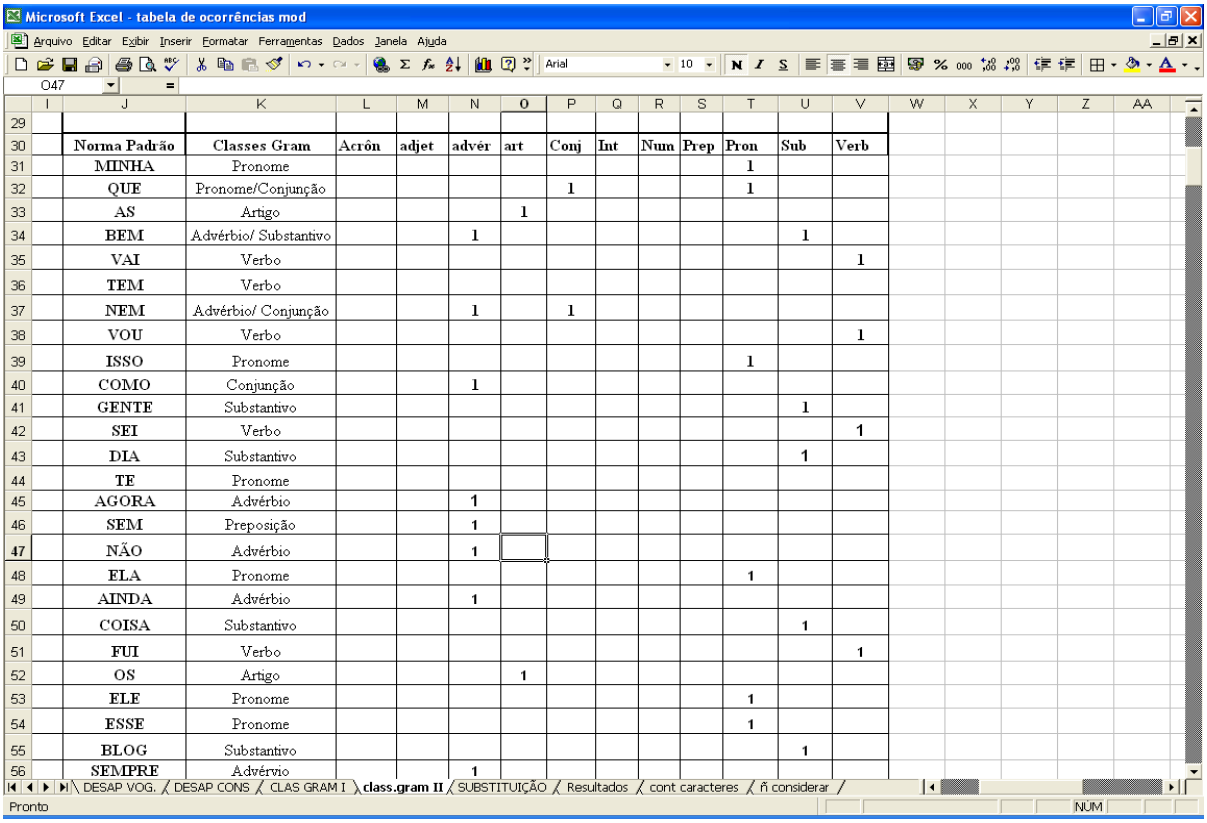
Em seguida, listamos manualmente as classes gramaticais dos itens selecionados para o estudo em duas planilhas do Excel, a primeira com itens em internetês e a segunda listando os itens da norma padrão, visando posterior determinação numérica dos percentuais de modificação por classe gramatical – e se, efetivamente, ocorreu modificação em cada uma das classes. As figuras 6 e 7 apresentam as planilhas com as classificações:



Microsoft Excel - tabela de ocorrências mod

Internetês	Norma Padrão	Cl. Gramatical	Pron	Interj	Verb	Advé	Conj	Prep	Art	Subs	Adje	contr	Acrón
D	De	Preposição						1					
EH	É	Verbo			1								
VC	Você	Pronome	1										
NAUM	Não	Advérbio				1							
DA	Dar	Verbo			1								
DO	Dou / Dó	Verbo / Substantivo			1					1			
SE	Ser / Você	Verbo / Pronome	1		1								
TD	Tudo - Todo	Pronome	1										
AKI	Aqui	Advérbio				1							
ME	Meu	Pronome	1										
NA	Não	Advérbio				1							
POR	Pôr	Verbo			1								
NO	Não	Advérbio				1							
PQ	por que - porque	Conjunção					1						
NEH	Né	Advérbio				1							
NUM	Não	Advérbio				1							
MTO	Muito	Advérbio				1							
TO	Estou	Verbo			1								
I	E	Conjunção					1						
AI	Aí	Conjunção					1						
VCS	Vocês	Pronome	1										
TAH	Está	Verbo			1								
TEM	Têm	Verbo			1								
U	O / Uh	Artigo / Interjeição		1					1				
HJ	Hoje	Advérbio				1							
SOH	Só	Advérbio				1							
MT	Muito	Advérbio				1							
TAVA	Estava	Verbo			1								
TE	Ter / Teu	Verbo / Pronome	1		1								
P e P/	Para	Preposição						1					

Figura 6 – Seleção das classes gramaticais dos itens em internetês.



Microsoft Excel - tabela de ocorrências mod

	Norma Padrão	Classes Gram	Acrôn	adjet	advér	art	Conj	Int	Num	Prep	Pron	Sub	Verb
31	MINHA	Pronome									1		
32	QUE	Pronome/Conjunção					1				1		
33	AS	Artigo				1							
34	BEM	Advérbio/ Substantivo			1							1	
35	VAI	Verbo											1
36	TEM	Verbo											
37	NEM	Advérbio/ Conjunção			1		1						
38	VOU	Verbo											1
39	ISSO	Pronome									1		
40	COMO	Conjunção			1								
41	GENTE	Substantivo										1	
42	SEI	Verbo											1
43	DIA	Substantivo										1	
44	TE	Pronome											
45	AGORA	Advérbio			1								
46	SEM	Preposição			1								
47	NÃO	Advérbio			1								
48	ELA	Pronome									1		
49	AINDA	Advérbio			1								
50	COISA	Substantivo										1	
51	FUI	Verbo											1
52	OS	Artigo				1							
53	ELE	Pronome									1		
54	ESSE	Pronome									1		
55	BLOG	Substantivo										1	
56	SEMPRE	Advérvio			1								

Figura 7 – Seleção das classes gramaticais das formas da norma padrão.

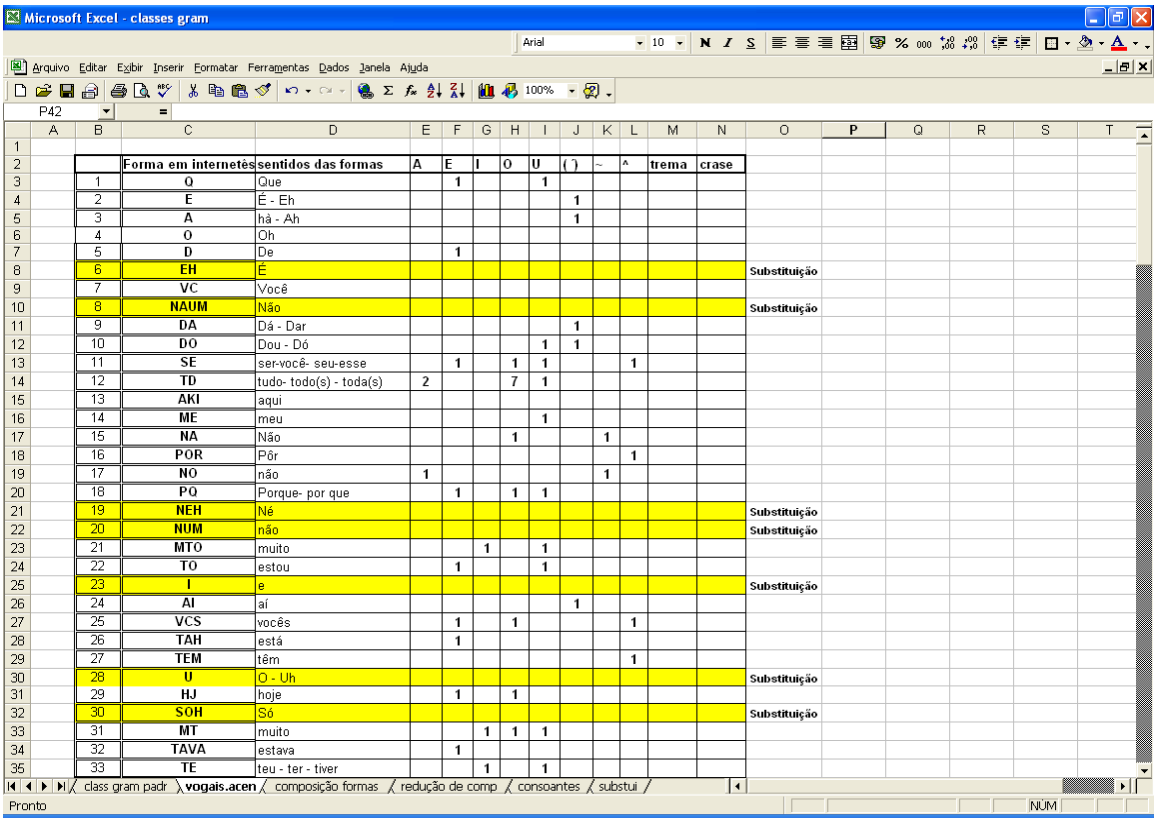
Cabe destacar que existe também, no mercado, uma ferramenta apta a classificar de classes gramaticais. Trata-se de etiquetadores morfossintáticos *(part of speech taggers)* que inserem automaticamente, no corpus, códigos que indicam a classe gramatical de cada palavra, de modo rápido e eficiente, facilitando assim a tarefa do analista. Um desses etiquetadores está disponível no site http://www.pucsp.br/pos/lael/cepril/cepril-info.php. Ainda não há, entretanto, etiquetador que reconheça a grafia do internetês e, por esta razão, não pudemos fazer uso da ferramenta para classificação gramatical dessa variedade lingüística.

2.5.2 Supressão das vogais, acentos e consoantes

Um fenômeno que ocorre tipicamente na grafia do internetês é a supressão de vogais, acentos gráficos e consoantes. Por exemplo, no item *'d'* (de) ocorre a supressão da vogal; já em *'da'* (dá) o acento agudo é omitido; em *'amanha'* (amanhã)

há a supressão do acento de nasalização e, em *'sabe'* (saber), a consoante final é suprimida.

Os procedimentos para levantamento de dados e análises relativos à supressão desses elementos seguiram a mesma ordem. Inicialmente, foi salva a lista de palavras com os termos em internetês em um arquivo do Excel. Foram então criadas colunas especificas para as supressões já observadas, a saber: uma coluna para a supressão da vogal 'a', outra para a do 'e', e assim sucessivamente (i, o, u), além de colunas para os diferentes acentos gráficos. O mesmo procedimento foi seguido em relação às possíveis supressões consonantais. A figura 5 sintetiza mais claramente os procedimentos adotados:



		Forma em internetês	sentidos das formas	A	E	I	O	U	( )	~	^	trema	crase	
1		Q	Que		1			1						
2		E	É - Eh						1					
3		A	há - Ah						1					
4		O	Oh											
5		D	De		1									
6		EH	É											Substituição
7		VC	Você											
8		NAUM	Não											Substituição
9		DA	Dá - Dar						1					
10		DO	Dou - Dó					1	1					
11		SE	ser-você- seu-esse		1		1	1			1			
12		TD	tudo- todo(s) - toda(s)	2			7	1						
13		AKI	aqui											
14		ME	meu					1						
15		NA	Não				1			1				
16		POR	Pôr								1			
17		NO	não	1						1				
18		PQ	Porque- por que		1		1	1						
19		NEH	Né											Substituição
20		NUM	não											Substituição
21		MTO	muito			1		1						
22		TO	estou		1			1						
23		I	e											Substituição
24		AI	aí						1					
25		VCS	vocês		1		1				1			
26		TAH	está		1									
27		TEM	têm								1			
28		U	O - Uh											Substituição
29		HJ	hoje		1		1							
30		SOH	Só											Substituição
31		MT	muito			1	1	1						
32		TAVA	estava		1									
33		TE	teu - ter - tiver			1		1						

Figura 8 – Arquivo do Excel exibindo as análises de supressões.

O arquivo constante da figura acima compreende análises de itens do internetês (coluna B), seguidos pelos itens correspondentes no vernáculo-padrão (coluna C). Vêem-se, à direita, as colunas relativas às vogais e aos acentos gráficos, bem como a indicação das supressões respectivas (colunas F-M).

Para determinação das ocorrências de supressão, foram observados todos os itens selecionados para a pesquisa; a cada desaparecimento de letra ou acento, era inserido o número 1 (ou acrescido mais um número) na coluna correspondente à letra ou acento suprimido. Dois pontos devem ser aqui ressaltados:

(1) freqüentemente, duas ou mais supressões de 'letras/acento' ocorrem num mesmo item, casos devidamente indicados nas respectivas colunas. Um exemplo é o item *'vc'* (você) – item número 7 da planilha acima - onde ocorrem simultaneamente o desaparecimento da vogal *'o',* o da vogal *'e',* e o do acento circunflexo **(^)**. Como mostra a figura 8, três entradas são dadas na mesma linha, nas colunas correspondentes (G, I e L);

(2) pode ocorrer, em um dado item, mais de uma supressão da mesma vogal/letra. Por exemplo, no item '*td*' (todo) – item número 12 da planilha acima - as duas vogais 'o' foram suprimidas, sendo então lançado o número 2 na coluna correspondente à supressão da vogal 'o' (coluna I).

A planilha do Excel oferece, ainda, a facilidade de calcular o total de cada uma das ocorrências, ao final de cada coluna.

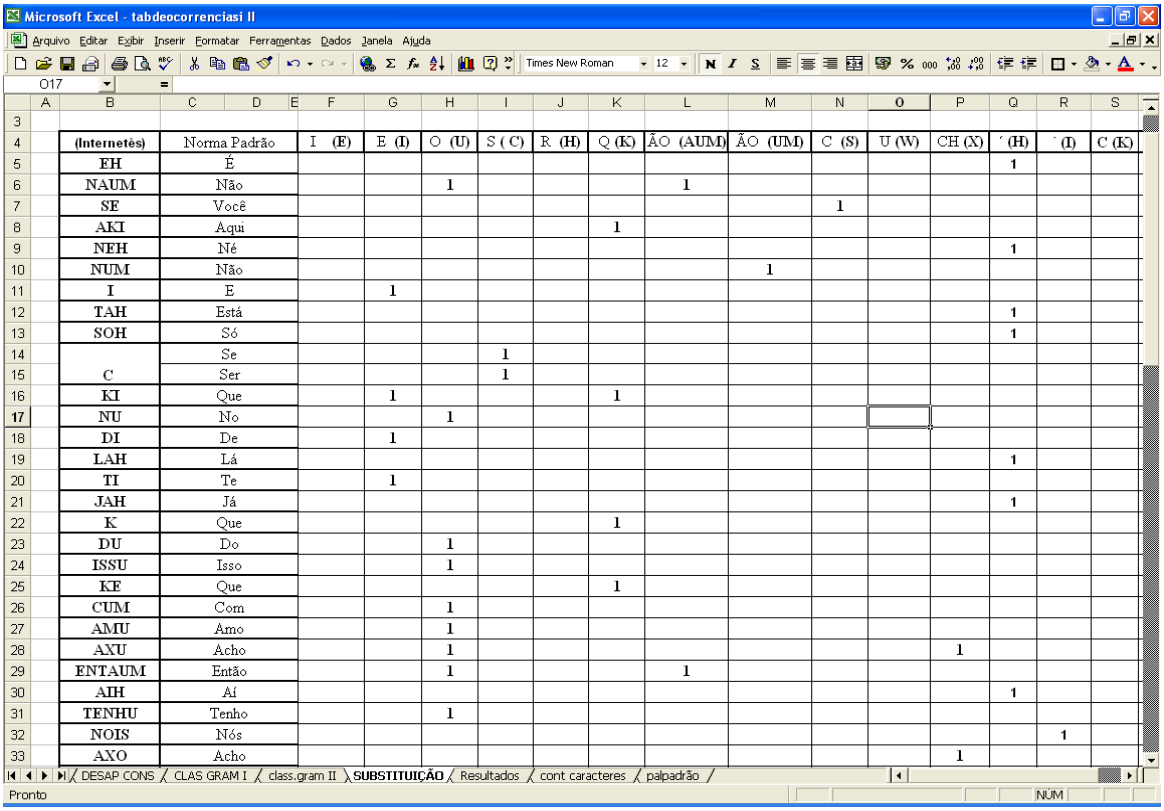
Finalmente, e ainda detendo-nos na figura 8, cabe observar que sete itens não apresentam indicações de supressão (grifo amarelo). Apesar disso, não foram eliminados das análises, pois serão enquadrados em outro fenômeno, a ser apresentado posteriormente.

2.5.3 Substituição de letras

Um outro fenômeno observado na grafia do internetês é a substituição de algumas letras por outras, ou ainda a substituição de acentos gráficos por outra forma de correspondê-los.

Para uma melhor observação desse fenômeno, nova lista de palavras em internetês foi criada na planilha do Excel e acrescida dos sentidos correspondentes em

língua-vernácula. Em seguida, inserimos colunas com a identificação das letras do alfabeto e sinais gráficos que poderiam estar sujeitos a substituições; em outras palavras, alocamos colunas para as letras ‘a’, ‘b’, ‘c’, (**~**), e assim por diante, acrescidas das respectivas substituições. Cada ocorrência de substituição era então sinalizada com a inserção do número 1 na coluna correspondente. Inicialmente, estimávamos que qualquer letra pudesse ser substituída na variedade internetês; no decorrer do levantamento, porém, percebemos que esse fenômeno não ocorre com todas as letras. Em conseqüência, excluímos as colunas das letras não utilizadas, restringindo a análise aos casos efetivamente observados. Para melhor exemplificar a substituição, tomemos a palavra ‘*amu*’ (amo) – item número 27 da planilha abaixo - onde, em vez de desaparecer, a vogal ‘o’ é substituída pela vogal ‘u’. Para esclarecer mais detalhadamente os procedimentos, apresentamos a figura abaixo:



(Internetês)	Norma Padrão	I (E)	E (I)	O (U)	S (C)	R (H)	Q (K)	ÃO (AUM)	ÃO (UM)	C (S)	U (W)	CH (X)	´(H)	´(I)	C (K)
EH	É												1		
NAUM	Não			1				1							
SE	Você									1					
AKI	Aqui						1								
NEH	Né												1		
NUM	Não								1						
I	E		1												
TAH	Está												1		
SOH	Só												1		
	Se				1										
C	Ser				1										
KI	Que		1				1								
NU	No			1											
DI	De		1												
LAH	Lá												1		
TI	Te		1												
JAH	Já												1		
K	Que						1								
DU	Do			1											
ISSU	Isso			1											
KE	Que						1								
CUM	Com			1											
AMU	Amo			1											
AXU	Acho			1								1			
ENTAUM	Então			1				1							
AIH	Aí												1		
TENHU	Tenho			1											
NOIS	Nós													1	
AXO	Acho											1			

Figura 9 – Planilha do Excel exibindo as substituições.

A figura 9 aponta as colunas com as letras/acentos usados na grafia da norma padrão, seguidos pelas substituições (em parênteses) ocorridas na grafia do internetês.

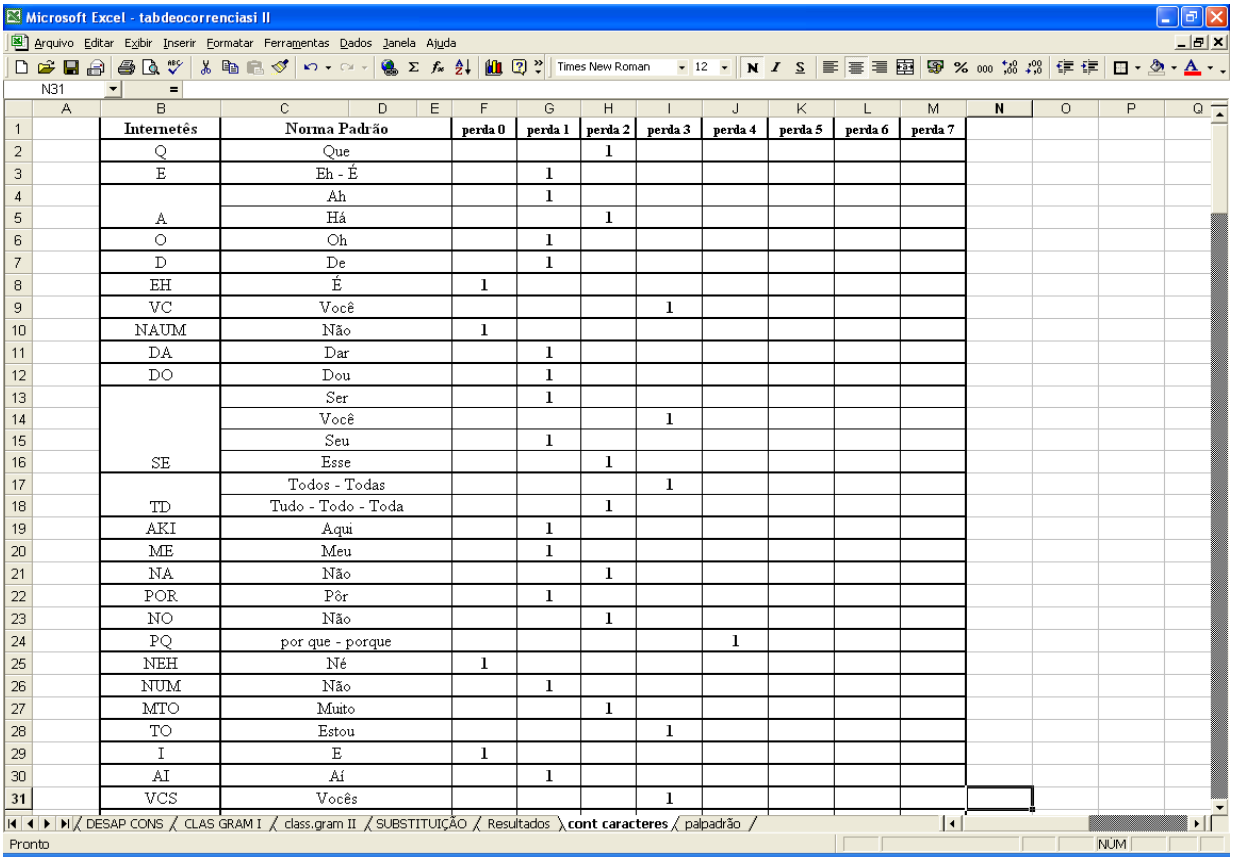
A planilha final forneceu-nos o total de ocorrências referente às letras substituídas, bem como suas respectivas substituições.

Reduções na grafia e nos toques (keystrokes)

Um fenômeno recorrente no internetês é a abreviação de palavras. A impressão geral é a de que as pessoas que se utilizam do internetês o fazem com o intuito de digitar a menor quantidade possível de caracteres, possivelmente com a intenção de economizar tempo.

Para verificar até que ponto a abreviação de palavras efetivamente reverte em um maior ganho de tempo, decidimos verificar se o resultado pretendido corresponde às expectativas dos usuários. Esse tipo de economia implica em abreviações que decorrem da supressão de letras ou acentos gráficos, conforme já discutido anteriormente no item 2.6.2 acima – *'vc', 'td'*, etc...

Passamos, então, a contabilizar quantos toques e, por conseguinte, quantas letras ou acentos gráficos foram subtraídos de algumas palavras. Para melhor ilustrar o procedimento, tomemos os itens '*vc'* (você) como exemplo: se observarmos o item 'você' na grafia padrão, temos uma palavra formada por 4 letras: 'v' (1), 'o' (2), 'c' (3), 'e' (4) mais um acento gráfico **(^)**; em internetês, a economia é de três toques, ou seja, duas letras e um acento gráfico, resultando em: *'v'* (1) e *'c'* (2). Portanto, as quatro letras e um acento gráfico da norma- padrão (cinco toques) são reduzidos a duas letras (ou dois toques) em internetês. A seguir, ilustramos o procedimento seguido:



Microsoft Excel - tabdeocorrenciasi II

Arquivo Editar Exibir Inserir Formatar Ferramentas Dados Janela Ajuda

N31

	Internetês	Norma Padrão	perda 0	perda 1	perda 2	perda 3	perda 4	perda 5	perda 6	perda 7
2	Q	Que			1					
3	E	Eh - É		1						
4		Ah		1						
5	A	Há			1					
6	O	Oh		1						
7	D	De		1						
8	EH	É	1							
9	VC	Você				1				
10	NAUM	Não	1							
11	DA	Dar		1						
12	DO	Dou		1						
13		Ser		1						
14		Você				1				
15		Seu		1						
16	SE	Esse			1					
17		Todos - Todas				1				
18	TD	Tudo - Todo - Toda			1					
19	AKI	Aqui		1						
20	ME	Meu		1						
21	NA	Não			1					
22	POR	Pôr		1						
23	NO	Não			1					
24	PQ	por que - porque					1			
25	NEH	Né	1							
26	NUM	Não		1						
27	MTO	Muito			1					
28	TO	Estou				1				
29	I	E	1							
30	AI	Aí		1						
31	VCS	Vocês				1				

DESAP CONS | CLAS GRAM I | class.gram II | SUBSTITUIÇÃO | Resultados | cont caracteres | palpadrão

Pronto NÚM

Figura 10 – Planilha do Excel exibindo a redução de toques.

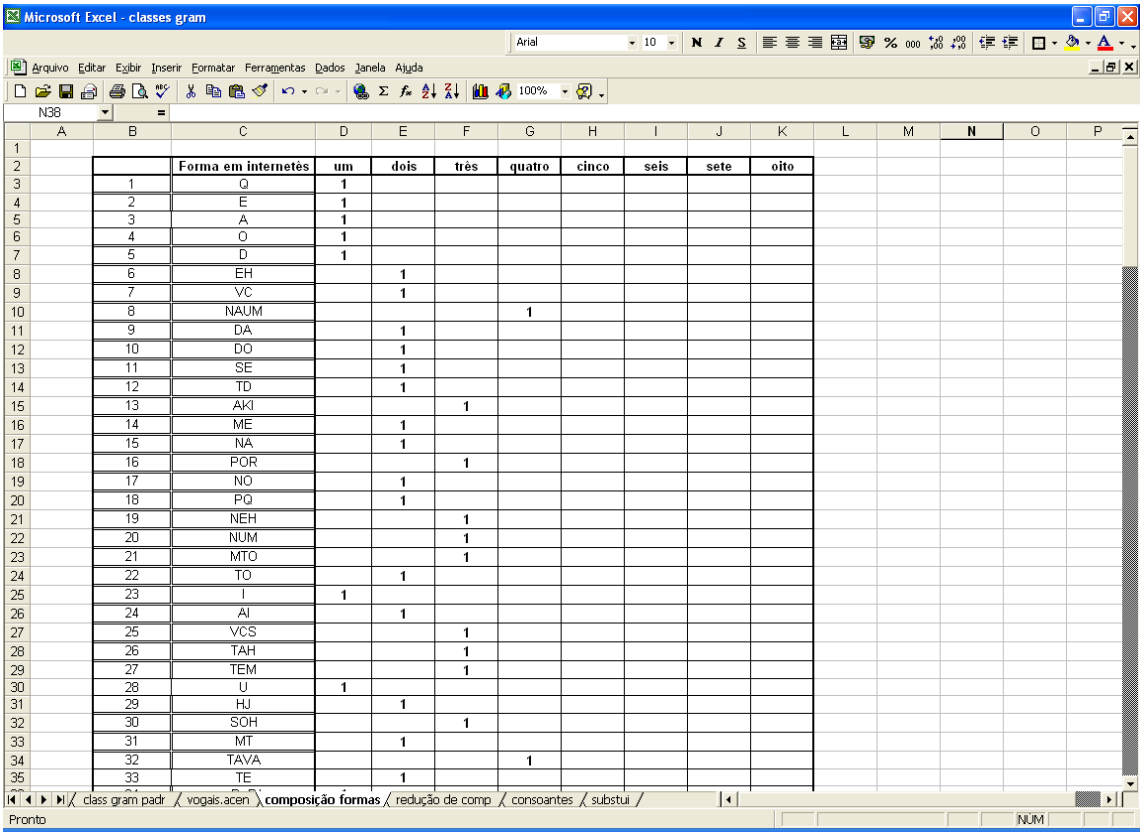
Esse procedimento comparativo entre as duas grafias – padrão e internetês - permitiu-nos claramente visualizar a subtração de ‘toques’ (*keystrokes*) ocorrida na segunda grafia, levando-nos a uma determinação mais acurada acerca da economia/perda de toques/letras/sinais na formação dos itens no internetês.

2.5.5 Quantidade de toques na formação dos itens em internetês

Como demonstrado na figura da seção anterior, as abreviações apontadas resultam, efetivamente, em uma economia de toques quando da utilização do internetês.

Com o intuito de determinar a quantidade de toques que foram necessários à composição desses itens, optou-se por levantar o número de toques comumente utilizados na formação dessas palavras. Para melhor exemplificar o procedimento adotado, tomemos as palavras ‘todo’ e ‘toda’ as quais, na formação do item

correspondente no internetês, são reduzidas a dois toques - *'td'*; em um outro exemplo, a palavra 'você' em internetês é representada por apenas dois toques - *'vc'*. Nos exemplos selecionados, a economia de toques verificada no internetês resultou na formação de itens formados por apenas dois toques. A figura 11 aplica o mesmo procedimento em um número maior de palavras:



Microsoft Excel - classes gram

	Forma em internetês	um	dois	três	quatro	cinco	seis	sete	oito
1	Q	1							
2	E	1							
3	A	1							
4	O	1							
5	D	1							
6	EH		1						
7	VC		1						
8	NAUM				1				
9	DA		1						
10	DO		1						
11	SE		1						
12	TD		1						
13	AKI			1					
14	ME		1						
15	NA		1						
16	POR			1					
17	NO		1						
18	PQ		1						
19	NEH			1					
20	NUM			1					
21	MTO			1					
22	TO		1						
23	I	1							
24	AI		1						
25	VCS			1					
26	TAH			1					
27	TEM			1					
28	U	1							
29	HJ		1						
30	SOH			1					
31	MT		1						
32	TAVA				1				
33	TE		1						

Figura 11 – Quantidade de toques na formação de palavras do internetês.

A figura 11 mostra-nos um fragmento das análises realizadas nos itens em internetês com intuito de verificar a quantidade de toques *(keystrokes)* foram necessários para a formação desses itens. Para tal, foram selecionadas colunas seguindo a ordem de um toque, dois toques, três toques e, assim por diante; possibilitando uma visão mais detalhada sobre a quantidade de toques freqüentemente são utilizados na formação dos itens. Ao final da planilha, pode-se calcular quantos toques mais freqüentemente são usados na formação dos itens do internetês.

Neste capítulo foi apresentada a metodologia de pesquisa empregada no estudo, incluindo a descrição das ferramentas utilizada para análise, bem como as divisões das análises. No capítulo, a seguir, serão apresentados e analisados os resultados da pesquisa.

CAPÍTULO 3 - Resultados

Neste capítulo, são apresentados e interpretados os resultados obtidos no decorrer deste trabalho, dados que respondem às questões que nortearam o estudo.

A interpretação de dados tenta relacionar o emprego das formas em internetês com fatores condicionadores, ou seja, quais fatores lingüísticos e/ou não-lingüísticos podem estar implicados nas modificações da grafia do internetês.

3.1. Análise da lista de palavras

Nesta seção, objetivamos responder à primeira questão da pesquisa:

1. Quais palavras são encontradas com maior freqüência nos corpora eletrônicos extraídos de blogs de jovens que utilizam o internetês?

Para respondê-la, foi solicitado ao programa WordSmith Tools (vide seção 2.3.2) a lista de palavras com todas as formas contidas no corpus de estudo. Em seguida, como não dispúnhamos de tempo hábil para empreendermos análises em todas as ocorrências do corpus, optamos, então, por 'recorte', analisando as 500 formas (types) mais freqüentes. Essa escolha resultou em 3% (500/15.552) do total de *types*. Da mesma forma, foram analisados os tokens equivalentes as 500 formas mais freqüentes. Essa análise resultou em 68% (93.083/138.021) de *tokens* do total do corpus.

A fim de demonstrar algumas palavras do corpus de estudo, relacionamos abaixo uma amostra com as 50 formas mais freqüentes:

N	PALAVRAS	FREQUÊNCIA	PORCENTAGEM
1	Q	3.993	2,89
2	Eu	3.443	2,49
3	E	3.306	2,40

4	A	2.638	1,91
5	O	1.889	1,37
6	De	1.753	1,27
7	Pra	1.592	1,15
8	D	1.161	0,84
9	Eh	1.124	0,81
10	Mas	1.103	0,80
11	Vc	1.037	0,75
12	Um	1.034	0,75
13	Mais	946	0,69
14	Naum	873	0,63
15	Meu	853	0,62
16	Da	828	0,60
17	Uma	816	0,59
18	Com	803	0,58
19	Do	796	0,58
20	Se	769	0,56
21	Td	760	0,55
22	Bom	747	0,54
23	Aki	742	0,54
24	Me	725	0,53
25	Na	708	0,51
26	Por	702	0,51
27	No	681	0,49
28	Foi	662	0,48
29	Pq	642	0,47
30	Neh	639	0,46
31	Num	638	0,46
32	Mto	625	0,45
33	To	623	0,45
34	Em	618	0,45
35	É	594	0,43

36	Minha	586	0,42
37	Que	581	0,42
38	As	579	0,42
39	I	568	0,41
40	Ai	554	0,40
41	Bem	542	0,39
42	Vcs	534	0,39
43	Tah	495	0,36
44	Vai	490	0,36
45	Tem	489	0,35
46	Nem	471	0,34
47	U	448	0,32
48	Vou	445	0,32
49	Hj	424	0,31
50	Isso	415	0,30

Quadro 04: primeiras 50 palavras mais freqüentes do corpus de estudo.

No quadro 04, apresentamos, primeiramente, a seqüência numérica de acordo com as palavras mais freqüentes do corpus de estudo; em seguida expomos as 50 formas mais freqüentes, tanto na grafia da norma padrão como em internetês. Ao observar superficialmente a amostra exposta acima, pode-se notar que embora seja uma variável lingüística presente em blogs, o internetês apresenta-se como um traço marcante e se sobressai com relação à grafia da norma padrão.

3.1.1. Análise das Linhas de Concordância

Essa seção responde a segunda questão da pesquisa, qual seja:

2. Quais palavras caracterizam o internetês?

Após o levantamento dos itens mais freqüentes, nosso objetivo foi o de confrontar as 500 palavras selecionadas para o estudo, com os dicionários de referência desta pesquisa (vide seção 2.6.2), com a finalidade de determinar quais palavras são grafadas de acordo com a norma padrão e quais palavras são grafadas em internetês.

Entretanto, ao confrontarmos os itens lexicais selecionados para a pesquisa com os dicionários selecionados para o cotejo, constatamos que os resultados não foram confiáveis, pois muitas palavras grafadas na norma padrão recebem um sentido/significado diferente quando verificadas em contexto. Para exemplificar esse fenômeno citamos a palavra '*num'* que de acordo com os dicionários utilizados como referência, trouxeram-nos conceitos, tais como: *(em + um) contração da preposição 'em' mais o artigo indefinido, masculino, singular 'um'*. – Acepção: indicação de lugar ('num determinado lugar').

Porém, ao investigarmos a palavra em seu contexto original, pudemos conferir que no internetês, a palavra '*num'* é utilizada numa freqüência maior com sentido de 'não'. A fim de exemplificar tal sentido, destacamos abaixo 10 linhas de concordância:

1 -	eu se ouvinte por 1 semestre (q aki	num	existe colegial eh td por semes
2 -	im vlw a pena pur 1ª coisa soh (q eu	num	possu fla aki) + q tm gent ae q j
3 -	e a Mar (28/09)... Mtus nivers (axu q	num	eskeci ng... se eskeci disculpa.
4 -	...eh soh 3 meses neh.mas eh triste	num	te a garota grad jettÃ© du lado!!
5 -	e refaze a burrada..... ... ai q saco...	num	falei naquele post a um tempo
6 -	omei a coca cum baunilha!hehehe ah	num	eh q eu num aguentei...mas
7 -	soh num foi perfeitu pq a nat e u btu	num	foram.mais tava da hora...
8 -	as músikas A Lu compro ingresso e	num	vai!!! =/ uma tosca mesmo!!! he
9 -	a mas entaum....to meio perdida...	num	sei qnd eu escrevi aki.....ah ta
10 -	oq tem acontecido na minha vida, pq	num	qru fikr lembrandu, fikou td q

Quadro 05: linhas de concordâncias do item lexical *'num'* com sentido de não.

A partir da observação das linhas de concordância, pudemos notar que vários itens lexicais são grafados conforme a norma padrão, porém em muitas ocorrências,

esses itens recebem um novo sentido/significado, caracterizando a grafia do internetês. A partir dessa constatação, tornou-se imprescindível à verificação do uso da palavra em seu contexto original, exigindo, assim, que extraíssemos as concordâncias de todos os itens selecionados para o estudo e, dessa forma, verificássemos os contextos de ocorrência; utilizando, para tanto, as listas de concordâncias e os colocados (vide seção 2.3.2.2).

Após as análises das linhas de concordância dos 500 itens selecionados para o estudo, obtivemos 35% (178/500) das formas grafadas com características do internetês e 62% (308/500) das formas grafadas de acordo com a norma padrão. Por conseguinte, foram excluídas das análises 3% (14/500) das formas, pois não atendiam às condições estabelecidas conforme mencionamos no capítulo 2 (vide seção 2.4).

O quadro abaixo apresenta de forma mais detalhada o resultado dessa investigação, destacando os itens em internetês que serão focados nas análises subseqüentes:

	Formas em internetês	**Palavra na norma padrão**	**Total de ocorrências da forma**	**Total de ocorrênci-as dos sentidos**	**Porcentagem para os sentidos**
1	Q	Que	3.993	3.993	100,00%
2	E	Eh!	3.306	6	0,18%
		É		42	1,27%
3	A	Ah!	2.638	13	0,49%
		Há		1	0,04%
4	O	OH!	1.889	2	0,11%
5	D	De	1.161	1.161	100,00%
6	EH	É	1.124	1.124	100,00%
7	VC	Você	1.037	1.037	100,00%
8	NAUM	Não	873	873	100,00%
9	DA	Dar	828	58	7,00%
		Dá		14	1,75
10	DO	Dou	796	3	0,38%
		Dó		2	0,36

11	SE	Ser	769	31	4,03%
		Você		14	1,82%
		Seu		1	0,13%
		Esse		1	0,13%
12	TD	Tudo – Todo (s), toda (s),	760	760	100,00%
13	AKI	Aqui	742	742	100,00%
14	ME	Meu	725	5	0,69%
15	NA	Não	708	2	0,28%
16	POR	Pôr	702	9	1,28%
17	NO	Não	681	1	0,15%
18	PQ	Por que-porque	642	642	100,00%
19	NEH	Né!	639	642	100,00%
20	NUM	Não	638	607	95,14%
21	MTO	Muito	625	625	100,00%
22	TO	Estou	623	620	100,00%
23	I	E	568	537	100,00%
24	AI	Aí	554	554	100,00%
25	VCS	Vocês	534	534	100,00%
26	TAH	Está	495	495	100,00%
27	TEM	Têm	489	33	6,75%
28	U	O	448	434	96,88%
		UH!		2	0,45%
29	HJ	Hoje	424	424	100,00%
30	SOH	Só	415	415	100,00%
31	MT	Muito	386	386	100,00%
32	TAVA	Estava	354	354	100,00%
33	TE	Ter	336	37	11,01%
		Tiver		1	0,30%
		Teu		1	0,30%
34	P e P/	P - Para	332	64	19,28%
		P/ - Para		172	51,81%

35	C e C/	Se	330	132	40,00%
		Com (C/)		145	43,94%
		Você		28	8,48%
		Ser		26	7,88%
36	TB	Também	326	326	100,00%
37	N.A.O	Não	322	322	100,00%
38	KI	Que	270	270	100,00%
39	NU	No	267	267	100,00%
40	LA	Lá	263	250	95,06%
41	VO	Vou	261	261	100,00%
42	DI	De	258	258	100,00%
43	TA	Está	247	247	100,00%
44	LAH	Lá	245	245	100,00%
45	TI	Te	241	241	100,00%
46	JAH	Já	228	228	100,00%
47	K	Que	226	226	100,00%
48	DU	Do	218	218	100,00%
49	MSM	Mesmo	215	215	100,00%
50	T – T+	Te/ até mais - ter	215	215	100,00%
51	MTU	Muito	213	213	100,00%
52	PARA	Pára	212	14	6,60%
53	TBM	Também	208	208	100,00%
54	J.A	Já	201	201	100,00%
55	SABE	Saber	189	9	4,76%
56	ISSU	Isso	184	184	100,00%
57	S.O	Só	171	171	100,00%
58	KE	Que	169	169	100,00%
59	N	Não	163	162	99,39%
60	CUM	Com	158	158	100,00%
61	ND	Nada	142	142	100,00%
62	PELO	Pêlo	142	1	0,70%
63	TÔ	Estou	142	142	100,00%

64	PA	Para	135	128	94,81%
		(Pah)!		2	1,48%
65	AMU	Amo	133	133	100,00%
66	AXU	Acho	133	133	100,00%
67	TDS	Todos	128	128	100,00%
68	ENTAUM	Então	126	126	100,00%
69	AGENTE	A Gente	123	126	100,00%
70	MTA	Muita	122	122	100,00%
71	AIH	Aí	120	120	100,00%
72	ATE	Até	120	120	100,00%
73	NOS	Nós	120	22	18,33%
74	TENHU	Tenho	120	120	100,00%
75	NOIS	Nós	119	119	100,00%
76	POST	postagem	113	113	100,00%
77	FLA	Fala	107	44	41,12%
		Falar		62	57,94%
78	AXO	Acho	102	102	100,00%
79	AE	Aí	100	100	100,00%
80	RS	Risos	99	99	100,00%
81	TAUM	Estão	97	28	28,87%
		Então		1	1,03%
		Tão		68	70,10%
82	FAZE	Fazer	95	95	100,00%
83	DAE	daí	92	92	100,00%
84	MOH	Maior	92	92	100,00%
85	NDA	Nada	92	92	100,00%
86	M	Me	88	88	100,00%
87	Ñ	Não	87	87	100,00%
88	ATEH	Até	86	86	100,00%
89	POKO	Pouco	85	85	100,00%
90	NOM	Não	81	82	100,00%
91	VOLTA	Voltar	81	10	12,35%
		Voltas		1	1,23%
92	GENT	Gente	80	80	100,00%

93	VE	Vê	79	22	27,85%
		Ver		57	72,15%
94	FIKEI	Fiquei	78	78	100,00%
95	KEM	Quem	77	77	100,00%
96	KSA	Casa	77	77	100,00%
97	S e S/	S/ - Sem	76	12	15,79%
		S – Se		31	40,79%
98	FICA	Ficar	75	24	32,00%
99	NIVER	Aniversário	75	75	100,00%
100	INTAUM	Então	73	73	100,00%
101	NET	Internet	69	69	100,00%
102	MIGA	Amiga	68	68	100,00%
103	BLZ	Beleza	67	67	100,00%
104	COMENTS	Comentário	67	67	100,00%
105	GENTI	Gente	67	67	100,00%
106	DPOIS	Depois	65	65	100,00%
107	QM	Quem	65	65	100,00%
108	TÁ	Está	65	56	86,15%
109	TOW	Estou	65	65	100,00%
110	BJAUM	Beijão	64	64	100,00%
111	US	Os	64	64	100,00%
112	BJUS	Beijos	63	63	100,00%
113	OQ	O que	63	63	100,00%
114	MTOOO	Muito	62	62	100,00%
115	FLOG	Fotoblog	61	61	100,00%
116	MI	Me	61	53	86,89%
		Mim		2	3,28%
117	QND	Quando	61	61	100,00%
118	TV	Estava	61	46	75,41%
		Tive		8	13,11%
119	DEXA	Deixa – deixar	59	59	100,00%
120	NUNK	Nunca	59	59	100,00%
121	PASSA	Passar	59	15	25,42%

122	TIVE	Tiver	59	4	6,78%
123	CMG	Comigo	57	57	100,00%
124	AGENTI	A gente	55	55	100,00%
125	MTOO	Muitíssimo	55	55	100,00%
126	S.A.O	São	55	55	100,00%
127	FIKAR	Ficar	54	54	100,00%
128	LAY	Layout	54	54	100,00%
129	MEW	Meu	54	54	100,00%
130	QDO	Quando	53	53	100,00%
131	TPO	Tipo	52	52	100,00%
132	SAUM	São	51	51	100,00%
133	ENTA.O	Então	51	51	100,00%
134	FICO	Ficou	50	23	46,00%
135	KRA	Cara	50	50	100,00%
136	OLHA	Olhar	50	2	4,00%
137	M.AE	Mãe	50	50	100,00%
138	DAH	Da	49	4	09,00%
		Dá		30	61,22%
		Dar		15	30,61%
139	PARECE	Parecer	48	1	2,08%
		Aparece		1	2,08%
140	AHH	Ah!	46	46	100,00%
141	AMANHA	Amanhã	46	46	100,00%
142	DPS	Depois	46	46	100,00%
143	MTAS	Muitas	44	44	100,00%
144	NEE	Né	44	44	100,00%
145	VLW	Valeu	44	44	100,00%
146	HR	Hora	43	43	100,00%
147	NGM	Ninguém	43	43	100,00%
148	GNT	Gente	42	42	100,00%
149	PAH	Pah!	42	5	16,67%
		Para		37	88,10%
150	DAKI	Daqui	41	42	100,00%

151	F.ERIAS	Férias	41	41	100,00%
152	FIKO	Fico - Ficou	41	41	100,00%
153	QNDO	Quando	41	41	100,00%
154	V	Ver	41	21	51,22%
		Vê		11	26,83%
155	AHHH	Ah!	40	40	100,00%
156	AKELA	Aquela	40	40	100,00%
157	BJOS	Beijos	40	40	100,00%
158	FLO	Falou	40	27	67,50%
		Falo		13	32,50%
159	PO	Pó!	40	35	87,50%
		Pro		3	0,80%
160	SAI	Sair	40	7	17,50%
		Saí		20	50,00%
161	TAO	Tão	40	33	82,50%
		Estão		7	17,50%
162	FIK	Ficar	39	20	51,28%
		Fica		14	35,90%
		Fique		5	12,82%
163	A.LGUEM	Alguém	39	39	100,00%
164	FDS	Fim de semana	38	39	100,00%
165	KERO	Quero	38	38	100,00%
166	MÓ	Maior	38	38	100,00%
167	SAB	Sabe Saber	38	38	100,00%
168	ESCREVE	Escrever	37	31	83,78%
169	POSTA	Postar	37	37	100,00%
170	COMENTA	Comentar	36	4	10,33%
171	DA.I	Daí -	36	36	100,00%
172	NINGUEM	Ninguém	36	36	100,00%
173	QD	Quando	36	36	100,00%
174	RSRS	Risos	36	36	100,00%
175	S.ABADO	Sábado	36	36	100,00%
176	VAMU	Vamos	36	36	100,00%
177	IH	Ir	35	33	94,29%

178	MUNDU	Mundo	35	35	100,00%

Quadro 06: lista de palavras com os itens mais freqüentes em internetês.

Acima, as formas apresentam-se antecedidas, primeiramente, pela classificação de acordo com a seleção dos itens selecionados para o estudo, ou seja, a palavra 1 corresponde à primeira palavra encontrada no corpus que apresentou características do internetês, a palavra dois corresponde à segunda forma e assim por diante. Na segunda coluna, apresentam-se os itens em internetês focados nesse estudo. Na terceira coluna, estão dispostos os sentidos apresentados em contexto, representados pelas palavras grafadas de acordo com norma padrão. Na quarta coluna, estão relacionadas à quantidade de ocorrência para cada item no corpus. A quinta coluna apresenta a quantidade de ocorrências em internetês, seguidos por suas respectivas porcentagens. Esse quadro fornece evidências sobre a diferença quantitativa entre o total de ocorrências da palavra no corpus e o total de ocorrências em internetês.

A partir dessas evidências, pudemos verificar que as 178 formas grafadas em internetês representaram 221 sentidos/significados de acordo com a grafia da norma padrão. Vale salientar que foram analisados 93.083 *tokens* do corpus de estudo. O resultado da análise dos *tokens* permitiu-nos comprovar que 33% (30.389/93.083) das ocorrências foram grafadas ou apresentaram sentido em internetês, enquanto 67% (62.694/93.083) das ocorrências foram grafadas conforme a norma padrão ou são ocorrências de itens excluídos das análises (vide capítulo 2, seção 2.3).

A fim de exemplificar alguns dos resultados[17] obtidos com essa investigação, arrolamos abaixo as 10 primeiras palavras mais freqüentes em internetês.

1. O item *'Q'* ocorreu 3.993 vezes no corpus de estudo e 100% dessas ocorrências são de 'que' (conjunção/pronome);

2. O item *'E'* ocorreu 3.306 vezes no corpus de estudo. Dessas ocorrências 0,18% (6/3.306) correspondem à interjeição 'eh'. Por outro lado, 1,27

[17] Esses resultados serão apresentados em letra maiúscula, com o propósito de melhorar a visualização dos itens lexicais, muito embora, eles possam ser encontrados no corpus de estudo tanto grafados em letras maiúsculas como em letras minúsculas.

(42/3.306) correspondem ao verbo ‘é’. Assim, 1,45% (48/3.306) dos usos de *‘E’* são em internetês e 98% (3.258/3.306) são grafados conforme a norma padrão;

3. O item *‘A’* ocorreu 2.638 vezes no corpus de estudo. Dessas ocorrências 0,49% (13/2.638) correspondem à interjeição ‘ah’. Por outro lado, 0,04 (1/2.638) correspondem ao verbo ‘há’. Assim, 0,53% (14/2.638) dos usos de ‘A’ são em internetês e 99% (2.624/2.638) de ‘a’ são grafados conforme a norma padrão;

4. O item *‘O’* ocorreu 1.889 vezes no corpus de estudo. Dessas ocorrências 0,11% (2/1.889) correspondem à interjeição ‘oh’. Portanto, 99,89% (1.887/1.889) de ‘o’ são grafados conforme a norma padrão;

5. O item *‘D’* ocorreu 1.161 vezes no corpus de estudo e 100% dessas ocorrências são de ‘de’ (preposição);

6. O item *‘Eh’* ocorreu 1.124 vezes no corpus de estudo e 100% dessas ocorrências são de ‘é’ (verbo);

7. O item *‘Vc’* ocorreu 1.037 vezes no corpus de estudo e 100% dessas ocorrências são de ‘você’ (pronome);

8. O item *‘Naum’* ocorreu 873 vezes no corpus de estudo e 100% dessas ocorrências são de ‘não’ (advérbio);

9. O item ‘*Da’* ocorreu 828 vezes no corpus de estudo. Dessas ocorrências 7% (58/828) correspondem ao verbo ‘dar’ (infinitivo). Por outro lado, 1,75% (14/828) correspondem ao verbo ‘dá’ (terceira pessoa do singular). Assim, 8,75% dos usos de *‘da’* são em internetês e 91% (756/828) de ‘da’ são grafados conforme a norma padrão;

10. O item *'Do'* ocorreu 796 vezes no corpus. Dessas ocorrências 0,38% (3/796) correspondem ao verbo 'dou'. Por outro lado, 0,36% (2/796) correspondem ao substantivo 'dó'. Assim, 0,74% dos usos de 'Do' são em internetês e 99% ou (791/796) de 'do' são grafados conforme a norma padrão;

Após a exploração minuciosa dos 10 primeiros itens mais freqüentes em internetês, obtivemos várias informações, tais como:

1. O resultado de *tokens* analisados (93.083) mostrou-nos que 33% (30.389) das ocorrências apresentaram características da grafia do internetês;

2. Alguns itens em internetês apresentam sentido de mais de uma palavra da norma padrão;

3. O item que ocupa o topo da lista, ou seja, a forma mais freqüente do corpus *'Q'* (que) é grafada sempre em internetês;

4. Entre os 10 primeiros itens mais freqüentes do corpus de estudo, 60% (6/10) das ocorrências apresentaram características do internetês; por outro lado, 40% (4/10) foram grafadas de acordo com a norma padrão. Ao estendermos essa observação às 20 primeiras palavras mais freqüentes, obtivemos 55% (11/20) grafadas em internetês e 45% (09/20) grafadas na norma padrão;

3.2. Freqüência de modificações na grafia

Nas próximas seções, vários dados foram analisados e interpretados com o objetivo de responderemos à terceira questão da pesquisa, qual seja:

3. Quais as modificações mais freqüentes ocorrem na formação de palavras do internetês?

Assim, cada seção tratará dos resultados obtidos com a análise do internetês e descreverá os tipos de modificações ocorridas, bem como sua freqüência.

3.2.1 Modificações nas classes gramaticais

Para descobrirmos quais classes gramaticais foram mais modificadas, primeiramente fizemos o levantamento dos itens em internetês (178/500). Em seguida, fizemos o levantamento das classes gramaticais das palavras grafadas de acordo com a norma padrão (308/500). Desse modo, pudemos confrontar a freqüência dessas modificações e assim estabelecermos quais classes foram mais modificadas.

A tabela abaixo apresenta os itens em internetês com as respectivas classes gramaticais:

	Forma em internetês	**vernáculos**	**classe gramatical**
1	Q	Que	Pronome
2	E	É/Eh	Verbo
3	A	Ah/Há	Interjeição
4	O	Oh	Interjeição
5	D	De	Preposição
6	EH	É	Verbo
7	VC	Você	Pronome
8	NAUM	Não	Advérbio
9	DA	Dá/Dar	Verbo
10	DO	Dou	Verbo
11	SE	Ser/Seu/Você	Verbo
12	TD	Todo/Toda/Tudo(s)/Todas(s)	Pronome
13	AKI	Aqui	Advérbio
14	ME	Meu	Pronome
15	NA	Não	Advérbio
16	POR	Pôr	Verbo

17	NO	Não	Advérbio
18	PQ	Porque/ Por que?	Conjunção
19	NEH	Né	Advérbio
20	NUM	Não	Advérbio
21	MTO	Muito	Advérbio
22	TO	Estou	Verbo
23	I	E	Conjunção
24	AI	Aí	Interjeição
25	VCS	Vocês	Pronome
26	TAH	Está	Verbo
27	TEM	Têm	Verbo
28	U	O/ Uh	Artigo
29	HJ	Hoje	Advérbio
30	SOH	Só	Advérbio
31	MT	Muito	Advérbio
32	TAVA	Estava	Verbo
33	TE	Ter/Tiver/Teu	Verbo
34	P e P/	Para	Preposição
35	C e C/	Com/Se/Seu/Ser	Preposição
36	TB	Também	Advérbio
37	NAO	Não	Advérbio
38	KI	Que	Pronome
39	NU	No	Preposição
40	LA	Lá	Advérbio
41	VO	Vou	Verbo
42	DI	De	Preposição
43	TA	Está	Verbo
44	LAH	Lá	Advérbio
45	TI	Te	Pronome
46	JAH	Já	Advérbio
47	K	Que	Pronome
48	DU	Do	Preposição

49	MSM	Mesmo	Advérbio
50	T	Te	Pronome
51	MTU	Muito	Advérbio
52	PARA	Pára	Verbo
53	TBM	Também	Conjunção
54	JA	Já	Advérbio
55	SABE	Saber	Verbo
56	ISSU	Isso	Pronome
57	SO	Só	Advérbio
58	KE	Que	Pronome
59	N	Não	Advérbio
60	CUM	Com	Preposição
61	ND	Nada	Pronome
62	PELO	Pêlo	Substantivo
63	TÔ	Estou	Verbo
64	PA	Para/ Pá	Preposição
65	AMU	Amo	Verbo
66	AXU	Acho	Verbo
67	TDS	Todos	Pronome
68	ENTAUM	Então	Advérbio
69	AGENTE	A gente	Pronome
70	MTA	Muita	Advérbio
71	AIH	Aí	Interjeição
72	ATE	Até	Advérbio
73	NOS	Nós	Pronome
74	TENHU	Tenho	Verbo
75	NOIS	Nós	Pronome
76	POST	Postagem	Substantivo
77	FLA	Fala/ Falar	Substantivo
78	AXO	Acho	Verbo
79	AE	E aí	Interjeição
80	NE	Né	Advérbio

81	RS	Risos/Riso	Substantivo
82	TAUM	Estão/Tão	Verbo
83	FAZE	Fazer	Verbo
84	DAE	E daí	Advérbio
85	MOH	Maior	Adjetivo
86	NDA	Nada	Pronome
87	M	Me	Pronome
88	Ñ	Não	Advérbio
89	ATEH	Até	Preposição
90	POKO	Pouco	Advérbio
91	NOM	Não	Advérbio
92	VOLTA	Voltar/ Voltas	Verbo
93	GENT	Gente	Substantivo
94	VE	Vê	Verbo
95	FIKEI	Fiquei	Verbo
96	KEM	Quem	Pronome
97	KSA	Casa	Substantivo
98	S e S/	Sem/ Se	Preposição
99	FICA	Ficar	Verbo
100	NIVER	Aniversário	Substantivo
101	INTAUM	Então	Advérbio
102	NET	Internet	Substantivo
103	MIGA	Amiga	Substantivo
104	BLZ	Beleza	Adjetivo
105	COMENTS	Comentário/ Comente	Substantivo
106	GENTI	Gente	Substantivo
107	DPOIS	Depois	Advérbio
108	QM	Quem	Pronome
109	TÀ	Está	Verbo
110	TOW	Estou	Verbo
111	BJAUM	Beijão	Substantivo
112	US	Os	Artigo

114	OQ	O que	Pronome
115	MTOOO	Muitíssimo	Advérbio
116	FLOG	Fotolog ou fotoblog	Substantivo
117	MI	Me	Pronome
118	QND	Quando	Advérbio
119	TV	Tiver/ Estava	Verbo
120	DEXA	Deixa	Verbo
121	NUNK	Nunca	Advérbio
122	PASSA	Passar	Verbo
123	TIVE	Tiver	Verbo
124	CMG	Comigo	Pronome
125	AGENTI	A Gente	Pronome
126	MTOO	Muitíssimo	Advérbio
127	SAO	São	Verbo
128	FIKAR	Ficar	Verbo
129	LAY	Layout/Leiaute	Substantivo
130	MEW	Meu	Pronome
131	QDO	Quando	Advérbio
132	TPO	Tipo	Substantivo
133	SAUM	São	Verbo
134	FICO	Ficou	Verbo
135	KRA	Cara	Substantivo
136	OLHA	Olhar	Verbo
137	MAE	Mãe	Substantivo
138	DAH	Da/ Dar	Preposição
139	PARECE	Aparece/parecer	Verbo
140	AHH	Ah	Interjeição
141	AMANHA	Amanhã	Advérbio
142	DPS	Depois	Advérbio
143	MTAS	Muitas	Advérbio
144	NEE	Né	Advérbio
145	VLW	Valeu	Verbo

146	HR	Hora	Substantivo
147	NGM	Ninguém	Pronome
148	GNT	Gente	Substantivo
149	PAH	Pá/ Para	Interjeição
150	DAKI	Daqui	Advérbio
151	F.E.R.IAS	Férias	Substantivo
152	FIKO	Fico/ Ficou	Verbo
153	QNDO	Quando	Advérbio
154	V	Vê/ Ver	Verbo
155	AHHH	Ah	Interjeição
156	AKELA	Aquela	Pronome
157	BJOS	Beijos	Substantivo
158	FLO	Falou	Verbo
159	PO	Pô	Interjeição
160	SAI	Saí/ Sair	Verbo
161	T.AO	Tão/ Estão	Advérbio
162	FIK	Ficar/ Fica	Verbo
163	A.LGUEM	Alguém	Pronome
164	FDS	Fim de Semana	Acrônimo
165	KERO	Quero	Verbo
166	M.O	Maior	Adjetivo
167	SAB	Sabe/ Saber	Verbo
168	ESCREVE	Escrever	Verbo
169	POSTA	Postar	Verbo
170	COMENTA	Comentar	Verbo
171	DA.I	Daí	Advérbio
172	NINGU.EM	Ninguém	Pronome
173	QD	Quando	Advérbio
174	RSRS	Risos	Substantivo
175	S.ABADO	Sábado	Substantivo
176	VAMU	Vamos	Verbo
177	IH	Ir	Verbo

178	MUNDU	Mundo	Substantivo

Quadro 07: classes gramaticais dos itens em internetês.

O quadro 07 relaciona os itens em internetês e seus respectivos correspondentes da norma padrão. Em seguida relacionamos as classes gramaticais dos itens, conforme o sentido apresentado em contexto. Como mencionado anteriormente, algumas formas em internetês apresentaram mais de um sentido em contexto e alguns itens podem exercer, também, diferentes funções gramaticais. Para ilustrar, citamos primeiramente o item *'taum'* que apresenta dois sentidos para a mesma forma, a saber: estão (verbo), como em: *'...vcs taum nu meu core...'* (vocês estão no meu coração...), e tão (advérbio), como em: '*essi ano foi taum legal foi taum emocionanti'* (esse ano foi tão legal, foi tão emocionante). Nestes exemplos, pode-se notar que o item *'taum'* apresenta duas classes gramaticais distintas, além de dois sentidos. Segundo, a forma *'q'* (que), exerce a função de pronome em várias ocorrências, como em: '*Q minina eh essa???*' (Que menina é essa?), em outras exerce a função de conjunção, como em: *"Ele si iskondeu s/ q ela visse!!!*', (Ele escondeu-se sem que ela o visse!)". Porém, como mencionado anteriormente, ao contabilizarmos os resultados, consideramos apenas os sentidos ou funções gramaticais que apresentaram uma freqüência maior de ocorrências.

Outro aspecto de igual importância, diz respeito à grafia diferente da convencional, aglutinando palavras ou utilizando letras para indicar uma expressão; isto é, incluímos nos resultados das classes gramaticais as contrações e a formação de acrônimos. Entre as contrações de itens (*items conjoined*), normalmente decorrentes da marca de oralidade, obtivemos exemplos tais como: '*agenti'* (a gente), *'pq'* (porque e por que) e *'oq'* (o que). Por outro lado, obtivemos apenas um item que indicou a formação de 'acrônimo', nesse caso 'fds' (fim/final de semana).

Essas análises contribuíram para determinar quais as classes gramaticais, mais freqüentemente, foram modificadas. O quadro abaixo apresenta estes resultados, considerando apenas, a classe gramatical que apresentou maior freqüência para o total de ocorrências.

Classes Gramaticais - Internetês	Total de Ocorrências	Porcentagem
Verbo	48	27%
Advérbio	40	23%
Pronome	30	17%
Substantivo	27	15%
Preposição	10	6%
Interjeição	09	5%
Contração	05	3%
Conjunção	04	2%
Adjetivo	02	1%
Artigo	02	1%
Acrônimo	01	1%
Numeral	00	0%

Quadro 08: classes gramaticais e freqüência das formas grafadas como internetês.

Segundo o quadro 8, as classes gramaticais que apresentaram uma modificação em uma freqüência mais alta foram os verbos, seguidos pelos advérbios e por fim, os pronomes. Esses resultados foram confrontados com as classes gramaticais correspondentes às palavras grafadas segundo a norma padrão, como segue:

Classes Gramaticais – norma padrão	Total de Ocorrências	Porcentagem
Verbo	87	28%
Substantivo	74	24%
Pronome	46	15%
Advérbio	35	12%
Adjetivo	22	7%
Preposição	16	5%
Artigo	08	3%
Conjunção	06	2%
Acrônimo	06	2%
Interjeição	04	1%

Numeral	04	1%
Interjeição	04	1%

Quadro 09: classes gramaticais e freqüência das formas da norma padrão.

Segundo os dados do quadro 9, os substantivos correspondem à classe mais freqüente das palavras grafadas pela norma padrão; seguido pelos verbos e advérbios. Os substantivos e verbos juntos respondem por mais da metade de todas as palavras da norma padrão que foram selecionadas para o estudo.

Passa-se, agora, à apresentação do levantamento contrastivo de ocorrências lexicais dos itens mais freqüentes do corpus, por meio de um gráfico. Para o eixo 'X' utilizamos letras do alfabeto que indicarão respectivamente:

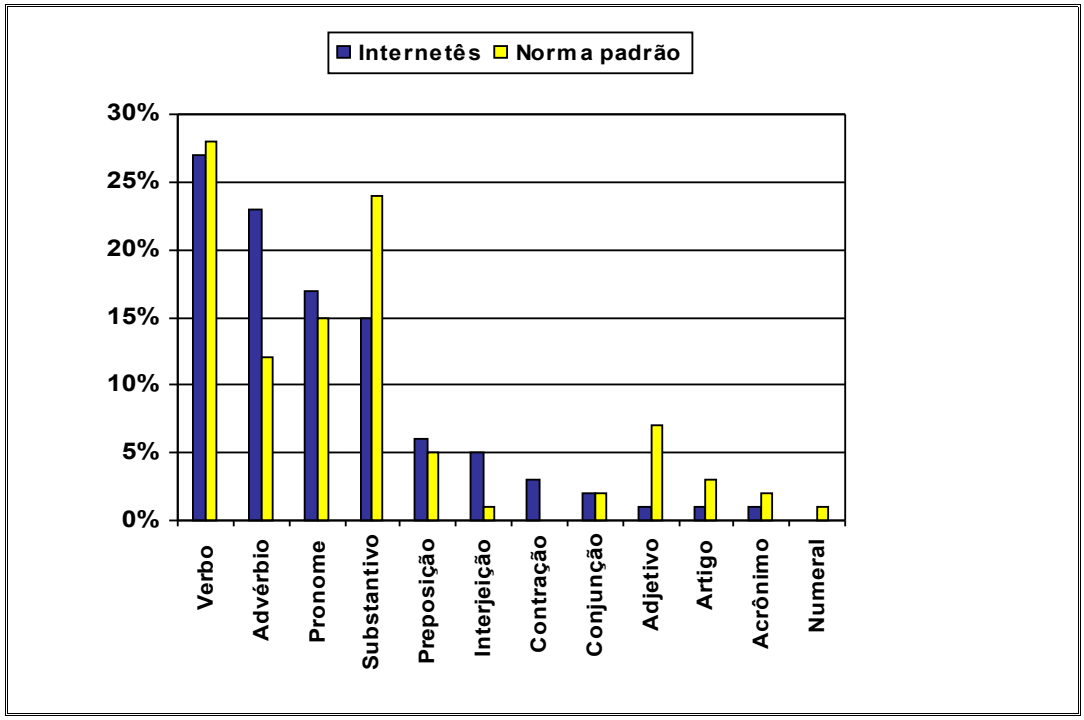


Gráfico 1 – resultado do confronto entre as classes gramaticais.

Ao confrontarmos os resultados das classes gramaticais dos itens em internetês, com as classes gramaticais da norma padrão, pôde-se notar que:

1. Os advérbios, preposições, pronomes e interjeições são as classes que são mais modificadas na grafia do internetês, pois o total de itens modificados ultrapassa o total de palavras grafadas conforme a norma padrão.

2. As conjunções são classes gramaticais que apresentaram uma percentagem equivalente, tanto na grafia padrão quanto em internetês. Dessa forma, há uma coincidência entre as classes mais modificadas e mais preservadas.

3. Os verbos apresentaram uma diferença percentual baixa de modificações na grafia, porém apresentaram uma maior incidência de preservação da norma padrão. Isso acontece porque os verbos são uma das classes que respondem pelo maior número de types do vocabulário da língua. Por outro lado, se compararmos as freqüências de ocorrências dos verbos tanto na norma padrão como no internetês, notamos que verbos no infinitivo tiveram sua grafia mais alterada do que preservada. A razão disso é a influência da oralidade na escrita, pois na 'fala' comumente não são pronunciados os 'r' finais dos verbos.

4. Os numerais não apresentaram ocorrência de modificação 0% (0/178) em internetês, embora tenha uma freqüência de 1% (3/309) na norma padrão. Isso indica que os internautas fazem uso de numerais na escrita, porém não são modificados. Além disso, muito provavelmente, os numerais são digitados como algarismos (10 em vez de 'dez', etc).

5. A criação de contrações ocorreram em 3% (5/178) no internetês. Tais contrações referem-se às palavras como: *'oq'* (o que), '*agenti*' (a gente), *'agente'* (a gente). Possivelmente, essas contrações acontecem por influência da oralidade na escrita, pois ao pronunciar tais palavras, comumente, os internautas pronunciam de forma aglutinada.

3.2.2 Supressão de vogais e acentos gráficos

À primeira vista tem-se a impressão de que o internetês é uma grafia que abrevia de forma indiscriminada. Com o intuito de verificar essas abreviações, primeiramente, observamos a supressão das vogais e acentos gráficos nos itens lexicais para estabelecermos a freqüência desses desaparecimentos, como mostra o quadro a seguir:

Vogais/ Acentos Suprimidos	Formas Modificadas Internetês	SENTIDOS ModificadOs	Total de Supressões
A	32	37	40
E	28	49	51
I	23	24	24
O	24	29	34
U	29	29	29
´ (acento agudo)	21	21	21
^ (acento circunflexo)	09	09	09
~ (til)	09	10	09
¨ (trema)	00	00	00
‘ (acento grave)	00	00	00

Quadro 10: supressão de vogais e acentos.

Os resultados expostos no quadro acima foram obtidos a partir da comparação entre as formas em internetês e os sentidos apresentados em contexto. Assim, na primeira coluna, estão relacionadas as vogais e os acentos que foram suprimidos das formas em internetês; na coluna dois, está relacionado o total de formas em internetês que ocorreram as eliminações. Na terceira coluna, estão relacionados à quantidade de sentidos que ocorreram as supressões. Por fim, a coluna quatro apresenta o resultado das vogais/acentos elididos.

Para exemplificar a disposição dos resultados à mostra no quadro 10, tomemos a vogal ‘e’; foram detectadas 51 vogais ‘e’ elididas em 28 formas do internetês. Essas formas, quando verificadas em contexto, resultaram em 49 sentidos (vide quadro 6,

seção 3.1.1). Dessa forma, para indicar os resultados das eliminações de vogais nas formas em internetês, tivemos que compará-las aos sentidos apresentados em contexto. Outra observação importante refere-se aos itens que apresentam a supressão de duas vogais iguais, tal como em *'gnt'* (gente) e '*blz'* (beleza). Essas palavras na norma padrão são grafadas com duas vogais 'e' que no internetês foram suprimidas. Nesses itens, foram contabilizadas as supressões de duas vogais 'e'; assim, o resultado da supressão de letras/acento, não foi o mesmo resultado do total de sentidos.

A partir das análises descritas acima, obtivemos os seguintes resultados:

Ocorreram eliminações de vogais/acentos em 66% (117/178) das formas; Por outro lado, ocorreram eliminações de vogais/acentos em 66% (146/221) dos sentidos;

1. Ocorreu o apagamento da vogal 'a' em 18% (32/178) das formas e 17% (37/221) dos sentidos;

2. Ocorreu o apagamento da vogal 'e' em 16% (28/178) das formas e 22% (49/221) dos sentidos;

3. Ocorreu o apagamento da vogal 'u' em 16% (29/178) das formas e um percentual igual para os sentidos;

4. Ocorreu o apagamento da vogal 'o' em 14% (24/178) das formas e 14% (29/221) dos sentidos;

5. Ocorreu o apagamento da vogal 'i' em 13% (23/178) das formas e 11% (24/221) dos sentidos;

6. Ocorreu o apagamento do acento agudo **(´)** em 12% (21/178) das formas e um percentual igual para os sentidos;

7. Ocorreu o apagamento do acento circunflexo **(^)** em 5% (9/178) das formas e um percentual igual para os sentidos;

8. Ocorreu o apagamento do 'til' **(~)** em 5% (9/178) das formas e 5% (10/221) dos sentidos;

9. Os apagamentos do trema e do acento grave ocorreram em 0% (0/178) das formas. Esse resultado foi obtido após observarmos que nenhuma das 500 primeiras palavras mais freqüentes do corpus apresentou o uso do trema; com relação à crase, verificamos que em muitas ocorrências, elas são substituídas por outras preposições ou pela letra 'h'. Também pudemos verificar que o item 'à' (fusão do artigo 'a' + preposição 'a'), consta entre as palavras mais freqüentes grafadas de acordo com a norma padrão.

Após analisarmos os resultados obtidos acima, contatamos que:

1. O resultado obtido para o apagamento das vogais *'a'*, *'e'*, *'i'* e *'u'* nas formas em internetês e nos sentidos podem ser considerados equiprováveis, pois a diferença oscilou entre 1% e 2%. Já com relação ao apagamento da vogal *'e'* os resultados apontaram uma diferença de 6% na comparação entre as formas e os sentidos, ou seja, os sentidos evidenciaram um apagamento mais freqüente da vogal *'e'*. Isso parece ocorrer, pois muitas letras do alfabeto (b, c, d, g, p, etc...) sonorizam vogal 'e'.

2. Os resultados discriminados acima mostraram que são favorecidas as supressões de todas as vogais nos itens em internetês, porém parece haver uma forte tendência para supressão da letra *'e'*, na qual as modificações ocorrem em maior freqüência e a porcentagem ultrapassa 20% dos sentidos. Além disso, observamos também que as vogais *'a'*, *'e'* e *'o'* são suprimidas mais de uma vez na mesma palavra, enquanto as letras 'i' e 'u' apresentam apenas uma supressão para cada palavra. Vale ressaltar que

as letras *‘e’*, *‘i’*, *‘o’* e *‘u’* foram incluídas também em outras análises que abordaremos posteriormente.

3. Todos os sinais gráficos são suprimidos, porém o acento **(´)** – acento agudo - apresenta maior freqüência de supressões com 12% (21/178) das formas modificadas e com resultado equivalente para os sentidos. Vale salientar que o acento agudo (´) foi incluído também em outras análises que abordaremos posteriormente.

4. O acento grave não apresentou ocorrências de supressões; entretanto, esse acento foi incluído também em outras análises que abordaremos posteriormente.

3.2.3 Supressão de consoantes

Além da eliminação de vogais e acentos gráficos, como relatados na seção anterior, o internetês apresenta também o desaparecimento de consoantes. Outro fato importante é que em muitas formas as consoantes não desaparecem, são apenas substituídas, porém estas análises serão apresentadas em uma seção subseqüente.

O quadro abaixo destaca as investigações sobre a eliminação das consoantes:

Consoantes Suprimidas	Formas Modificadas Internetês	Formas Modificadas Sentido	Total De Supressões
R	31	33	33
S	15	15	15
M	5	5	6
H	5	6	6
N	5	5	5
T	3	3	3
V	3	3	3
B	1	1	1
G	1	1	1

Quadro 11: supressões de consoantes.

O quadro acima apresenta, primeiramente, a relação de consoantes que foram eliminadas das palavras na grafia padrão. Segundo, está a quantidade de formas em internetês que ocorreram essas eliminações. Terceiro estão o total de sentidos que ocorreram as eliminações. Por fim, está o total de eliminações de consoantes para cada sentido. Como mencionado anteriormente, um fenômeno típico do internetês são duas ou mais supressões de letras iguais numa uma dada forma. Um exemplo disso é o item *'tb'* (também), cuja grafia subtrai duas vezes a consoante 'm'; assim, o resultado sobre a quantidade de consoantes eliminadas foi superior ao total de sentidos. Após as análises acuradas, obtivemos os seguintes resultados:

1. Ocorreram supressões de consoantes em 39% (69/178) das formas e 33% (72/221) para os sentidos;

2. Entre as supressões mais freqüentes encontramos o 'R' com 19% (33/178) dos itens e o 'S' com 9% (15/178) dos itens. A eliminação de consoantes nesses itens ocorre muitas vezes, por influência da oralidade, na qual alguns estudos (Tarallo, 2005; Mollica, 2007) relatam o apagamento da consoante *'r'* em finais de verbos e da consoante *'s'* nos finais de palavras no plural. No caso da internet, os blogueiros parecem buscar uma escrita mais informal e espontânea, despreocupados das formas ou conteúdos estabelecidos pelas normas que regem a grafia padrão, com a finalidade de manter o contato com a comunidade da qual faz parte, expressar seus sentimentos, pensamentos e parece fazê-lo com base na oralidade. Assim, tal como a fala, o emprego do 'R' e 'S' nas palavras torna-se desnecessário, pois o apagamento ocorre principalmente para indicar os plurais e os finais de verbos. Por exemplo, o item *'volta'* (voltas); ao consultá-lo no corpus de estudo pudemos verificar a expressão: '*A genti sai e da umas volta!!!'* (A gente sai e dá umas voltas!) ou na expressão *'Naum vo escreve nu blog dele...'* (Não vou escrever no blog dele...). No

primeiro exemplo, a indicação do plural (voltas) é expressa anteriormente pela palavra ‘umas’ deixando de ser grafado o *‘s’* no final da palavra. Já no segundo exemplo, o *‘r’* no final dos verbos, não é grafado tal como é reproduzido na fala;

3. A supressão da letra ‘m’ ocorre em 3% (5/178) dos itens e a supressão de ‘n’ em 3% (5/178). Essas supressões tipicamente ocorrem quando essas letras são usadas exercendo a função de nasalização. Para exemplificar destacamos os itens *‘tb*’ (também), *‘c/’* (com), *‘qdo’* (quando), *‘ngm’* (ninguém), nas quais as letras *‘m’* e *‘n’* foram elididas. Portanto, podemos considerar que quando as letras *‘m’* e *‘n’* exercem a função de nasalização, há uma forte tendência para que sejam elididas nas formas do internetês;

4. A supressão da letra ‘h’ ocorre em 3% (5/178) dos itens em internetês. Os sentidos apresentados em contexto que correspondem às supressões são as interjeições ‘ah’ ‘eh’ ‘oh’ ‘uh’ ‘pah’ e o verbo ‘há’, nos quais a letra ‘h’ não é reproduzida sonoramente.

5. As letras que apresentaram porcentagem relativa a 1%, ou seja, ‘B’ (1/178) ‘G’ (1/178), ‘T’ (3/178), V (3/178), foram elididas de palavras que fazem parte do vocabulário do meio internáutico, de interjeições ou por influência da oralidade na escrita. Para ilustrar, citamos os itens: *‘flog’* (fotoblog); *‘post’* (postagem); ‘*lay’* (layout); *‘c’* (você).

3.2.4 Substituição de letras e/ou acentos gráficos

Outro fenômeno típico do internetês é a substituição de letras e acentos gráficos por outros, considerados pelos internautas, equivalentes na grafia do internetês. Para ilustrar podemos citar o item *‘axu’* (acho), no qual as consoantes ‘ch’, grafadas de acordo com a norma padrão, foram substituídas pela letra *‘x’* no internetês.

Essa substituição também ocorre com as vogais e acentos. Como exemplo, tomemos os itens *'vamu'* (vamos), *'amu'* (amo), *'i'* (e – conjunção), nos quais foram substituídas as vogais por outras sonoramente equivalentes. Já entre a substituição de acentos citamos: *'soh'* (só), *'eh'* (é), *'saum'* (são) que tipicamente são substituídos por letras. O quadro abaixo traz os resultados dessas substituições:

Letras - Grafia da Norma Padrão	**Letras – na grafia do Internetês**	**Total de Substituições**
O	*U*	12
´ (acento agudo)	*H*	9
QU	*K*	9
E	*I*	9
C	*K*	6
ÃO	*AUM*	5
U	*W*	3
CH	*X*	2
I	*E*	2
ÃO	*UM*	1
C	*S*	1
^ (acento circunflexo)	*U*	1
´ (acento agudo)	*I*	1
‘ (acento grave)	*H*	1
R	*H*	3
S	*C*	1
ÃO	*OM*	1

Quadro 12: substituições de letras/acento.

No quadro 12, a primeira coluna relaciona as letras que constam nas palavras de acordo com a grafia da norma padrão. Na segunda coluna, relacionamos as letras na grafia do internetês. Na terceira coluna, relacionamos a quantidade de substituições referente às letras/acentos. Vale destacar que em muitas formas do internetês ocorre

a substituição de duas ou mais letras/acento; para ilustrar citamos o item *'Ki'* (que). As substituições ocorrem nos agrupamento de letras 'qu' (pela letra '*k')* e a vogal 'e' (pela letra '*i')*.

Após as análises, constatamos que 34% (60/178) das formas em internetês apresentaram substituições. Com relação aos sentidos, 29% (65/221) das palavras grafadas de acordo com a norma padrão apresentaram substituições de letras/acento.

A partir do levantamento desses dados, apresentaremos as substituições mais freqüentes, quais sejam:

1. A letra 'O' foi substituída por *'U'* em 7% (12/178) dos itens em internetês. Essa substituição parece indicar que ao digitar os internautas expressam muitas vezes a oralidade. Em outras palavras, há uma tendência em substituir o som da letra 'o' nos finais das palavras por 'u', como na fala (Molica, 2007:79);

2. O acento agudo **(´)** foi substituído por *'H'* em 5% (9/178) dos itens em internetês. Nesses casos, o princípio não é o da economia lingüística (não há uma redução na escrita), pois na digitação da palavra a quantidade de toques (keystrokes) é idêntica. Uma possível explicação para a falta de acentos seria a forte influência da oralidade na escrita internáutica, indicando que a vogal que precede imediatamente a letra 'h' tem um som aberto; por exemplo, /é/ (Othero, 2004);

3. As letras 'QU' foram substituídas pela letra *'K'* em 5% (9/178) dos itens em internetês. Essa substituição parece ocorrer, pois ao digitar o internauta visa ser o mais rápido possível, desta forma, para economizar tempo substituem o dígrafo 'QU' pela letra *'K'*. Assim, ao invés de grafarem a palavra 'aqui' eles escrevem *'aki'.* Ao contrário do exemplo anterior (*'H'* como substituto do acento agudo), temos um fenômeno considerado como 'simplificação ortográfica'. Outra explicação possível para a substituição do dígrafo 'QU' pela letra 'K' seria por influência de termos em inglês ou

termos da informática (oriundos do inglês) que apresentam a grafia de palavras com a letra ‘k’ (Othero, 2004); alguns exemplos são: ‘*OK’* , *cracker*, *hacker, kicka*r (*kick*), entre outros;

4. A letra ‘E’ foi substituída por *‘I’* em 5% (9/178) dos itens em internetês. Essa substituição parece indicar também uma forte influência da oralidade na escrita internáutica. Ao digitar os internautas o fazem de forma rápida e simulam uma ‘fala’, visto que, tendem a substituir o som da letra ‘e’ nos finais das palavras por ‘i’ no português brasileiro (Othero, 2004);

5. A letra ‘C’ foi substituída pela letra *‘K’* em 3% (6/178) dos itens em internetês. Na maior parte dos itens, essa substituição mostra que o fonema /k/ pode substituir duas letras, isto é, a letra ‘C’ e a letra ‘A’. Alguns exemplos dessa substituição incluem os itens ‘*ksa’* (casa), *‘Kra’* (cara), *‘nunk’* (nunca) e *‘fik’* (ficar e fica), na qual o nome da consoante ‘K’ supre a vogal que não é escrita (Possenti, 2006:31); enquadrado também no fenômeno de simplificação ortográfica. Por outro lado, pudemos constatar que os itens *‘fiko’* (fico, ficou) e *‘fikar’* (ficar) não ocorre a simplificação ortográfica, apenas a substituição da letra ‘C’ pela letra *‘K’*.

6. As letras/acento ‘ão’ foram substituídos pelas letras ‘*aum*’ em 3% (5/178) dos itens em internetês. Nesse caso, o princípio também não é o da economia lingüística (não há uma redução na escrita) e pode ser considerado análogo ao caso da substituição do acento agudo pela letra *‘H’*. Tal substituição decorre da substituição de acento pela grafia de letras, tornando possível à pronúncia idêntica da palavra. De forma similar, o ditongo nasal ‘ão’ foi substituído pela sílaba semelhante *‘aum’*, pois em muitas palavras da língua portuguesa a letra ‘M’ funciona como nasalizador. Já o som da letra ‘o’ na fala freqüentemente é substituído por */u/*; portanto para grafar tais palavras, os internautas substituem a letra ‘o’ pelo som de /u/, igualmente para o sinal de nasalização **(~)** pela letra ‘M’.

Vale ressaltar que se a palavra fosse grafada com as letras '*aom*', substituindo apenas o sinal de nasalização (~) pela letra 'm', o som reproduzido "ao" seria completamente diferente do que faz parte da palavra.

7. A letra 'U' foi substituída por '*W*'em 2% (3/178) dos itens em internetês, como em '*vlw*' (valeu). Embora seja pouco freqüente, essa substituição parece ocorrer, pois em algumas palavras reconhecidas pelos falantes da língua portuguesa, a letra 'w' tem o som /u/. Tais palavras apresentam-se freqüentemente como nomes próprios, ou seja: Wilson, Wellington e Willian. Assim, por reconhecerem que estes nomes são grafados com a letra 'W', mas tem o som da letra 'U', os internautas fazem essas substituições.

8. As letras 'CH' foram substituídas por *'X'* em 1% (2/178) dos itens em internetês. Tal como o fenômeno do dígrafo 'QU', os internautas substituem duas letras por uma. Isso parece suscitar pelo desejo de digitar mais rápido; assim, ao invés de grafarem a palavra 'acho', eles escrevem *'axo'* ou *'axu'.* Também é um fenômeno considerado uma "simplificação ortográfica" (Othero, 2004).

9. A letra 'I' foi substituída pela letra '*E'* em 1% (2/178) dos itens em internetês. Essa substituição ocorreu nos itens '*ae*' (e aí) e '*dae*' (e daí); nesses itens parece haver a influência da oralidade, pois algumas vezes, 'ai' é pronunciado 'aê', de modo informal, na fala. O mesmo ocorrendo com a pronúncia 'daê'.

10. As letras/acento 'ÃO' foram substituídas por 'UM' em 1% (1/178) dos itens em internetês. Embora, o resultado apresente baixa freqüência, pudemos contatar ao observar os contextos de ocorrência que 95,14% (607/638) o item *'num'* corresponde ao sentido/significado da palavra 'não'. Essa

substituição, possivelmente, ocorre por dois motivos: Primeiro, o princípio da economia lingüística semelhante a relatados anteriores. Segundo, a influência da oralidade, pois ao pronunciar a palavra ‘não’ comumente o som é reproduzido como ‘num’.

Portanto, se agruparmos os resultados conforme se apresentam os fenômenos, obteremos os seguintes totais:

1. 18% (32/178) para os fenômenos que possivelmente ocorrem por influência da oralidade na escrita (casos 1, 2, 4, 9);

2. 9% (15/178) para os fenômenos que ocorre a simplificação ortográfica (casos 3, 5, 8);

3. 5% (8/178) para os fenômenos de substituição de letras, sem eliminação de letras/acento ou influência da oralidade na escrita (casos 6 e 7);

4. 1% (1/178) para o fenômeno que, possivelmente, apresenta tanto a economia lingüística quanto à influência da oralidade na escrita (caso 10);

Assim, obtivemos uma maior incidência para os fenômenos que possivelmente ocorrem pela influência da oralidade na escrita. Em seguida, os fenômenos que apresentam em suas formas a simplificação ortográfica ou a economia lingüística.

3.2.5 Eliminação de toques (Keystrokes)

As abreviações costumeiras na grafia do internetês implicam não apenas na economia lingüística, como também, na redução de toques[18] (keystrokes) na digitação de palavras. Nesta seção, determinaremos a freqüência da redução do número de

[18] Referimo-nos a toque como a ação que indicará individualmente a impressão na tela de letra ou acento gráfico.

toques (keystrokes) na transcrição da norma padrão para o internetês. Para tanto, foram contabilizados os sentidos (221/500), pois desta forma conseguiremos obter com mais fidelidade os resultados sobre essa passagem, já que a grafia internáutica muitas vezes representa vários sentidos; para exemplificar os vários sentidos, destacamos o item *'td'* (tudo, toda, todo, todos e todas).

A tabela abaixo mostra o resultado dessas reduções:

Toques na grafia padrão	Toques em internetês	Toques elididos	Total de sentidos	Porcentagem
2	1	1	10	5%
3	1	2	7	4%
3	2	1	28	13%
4	1	3	2	1%
4	2	2	17	8%
4	3	1	27	13%
5	1	4	1	1%
5	2	3	9	3%
5	3	2	19	9%
5	4	1	18	8%
6	2	4	3	1%
6	3	3	8	4%
6	4	2	8	4%
6	5	1	7	3%
7	2	5	1	1%
7	3	4	1	1%
7	5	2	1	1%
7	6	1	5	2%
8	2	6	1	1%
8	3	5	2	1%
8	4	4	2	1%
8	7	1	3	1%
8	6	2	1	1%

11	3	8	1	1%
11	5	6	1	1%
11	7	4	1	1%
12	5	7	1	1%
*1	1	0	2	1%
*2	2	0	8	4%
*3	3	0	11	5%
*4	4	0	5	2%
*5	5	0	4	1%
*6	6	0	3	1%
2*	3	0	2	1%
2*	4	0	1	1%

Quadro 13: quantidade de toques na grafia da norma padrão e do internetês.

O quadro acima mostra-nos, primeiramente, os itens que sofreram redução no número de toques, conforme os sentidos apresentados em contexto. As reduções ocorreram em 84% (185/221) das palavras. Por outro lado, 15% (33/221), dos itens[19] não apresentaram eliminações de toques, apenas modificações na grafia; e finalmente em 1% (3/221) houve a extensão da grafia das palavras[20], como por exemplo, nos itens: *'Ahh'* e '*Ahhh'* (ah!).

1. Os resultados mostraram uma freqüência majoritária de 45% (98/221) para eliminação de um toque e, por conseguinte, de uma letra/acento na passagem da grafia padrão para o internetês. Essa freqüência majoritária indica que parece haver uma forte tendência na eliminação de 1 toque, pois a eliminação ocorre em quase 50% das palavras.

[19] Alguns itens não apresentaram reduções na passagem da grafia da norma padrão para o internetês. Estes itens foram sinalizados no quadro 13, com asterisco (*) precedendo a quantidade de toques na grafia padrão – primeira coluna.
[20] Alguns itens apresentaram aumento na quantidade de toques na passagem da norma padrão para o internetês. Estes itens foram sinalizados no quadro 13, com asterisco (*), em negrito, sucedendo a quantidade de toques na grafia padrão – primeira coluna.

2. Segundo, estão à eliminação de 2 toques e conseqüentemente duas letras/acento, apontando 24% (53/221) dos sentidos em contexto.

3. Terceiro, estão os casos que eliminam 3 toques das palavras da norma padrão, apresentando 9% (19/221) de palavras modificadas.

4. Porém, dois casos, em particular, apresentaram baixa freqüência 1% (2/221), ou seja, a eliminação de 7 em *'niver'* (aniversário) e a eliminação 8 toques no acrônimo *'fds'* (fim de semana).

3.2.6 Formação dos itens em internetês

A redução do tamanho das palavras parece ser um critério marcante da grafia do internetês, lhe conferindo o caráter abreviado, estenográfico ('shorthand') que é talvez sua marca registrada. Podemos destacar aqui as palavras *'e'* (eh), *'a'* (ah), *'d'* (de) e *'o'* (oh), que são as mais freqüentes com redução de um toque, e *'a'* (há), '*td'* (tudo, toda, todo), *'mto'* (muito), as mais freqüentes com redução de dois toques. Entretanto, conforme mostrado na seção anterior, há casos cuja extensão da palavra é equivalente à quantidade de toques, como em: '*naum*' (não) e *'soh'* (só).

Dessa forma, com a finalidade de verificar a predominância da quantidade de toques para a formação mais freqüente dos itens em internetês, contabilizamos a quantidade de letras/acentos utilizados para a composição desses itens. Para tal análise, consideramos os 178 itens grafados em internetês, tanto os que foram abreviados como os que não sofreram abreviações (aumento da palavra ou mesma extensão). O quadro abaixo destaca essas análises:

Toques na formação dos itens	Total de formas	Porcentagem
1	15	8%
2	43	25%
3	61	34%
4	33	18%

5	13	7%
6	09	5%
7	04	3%

Quadro 14: quantidade de toques na formação das palavras em internetês.

Ao observarmos as análises apresentadas acima, pudemos constatar que das 178 formas em internetês selecionadas para o estudo, em 34% (61/178) há uma predominância de palavras grafadas com 3 toques. Em segundo, com 25% (43/178) estão as palavras grafadas com 2 toques.

Assim, constatamos que no internetês, há uma predominância de palavras grafadas com 2 e 3 toques, isto é, dos 178 casos observados 59% (104/178), as palavras foram grafadas com 2 e 3 letras/acento.

3.3 Análise dos padrões

Ao observarmos os types mais freqüentes do corpus de estudo, percebemos que alguns itens apresentavam vários sentidos. Optamos, então por analisar os padrões de um dado item do internetês, com a finalidade de buscar evidências que constatem se o léxico é padronizado, observando os contextos de ocorrência, utilizando, para tanto, a lista de concordância e os colocados. Portanto, essa seção objetiva responder a quarta pergunta da pesquisa, qual seja:

1. Quais padrões léxico-gramaticais são encontrados com mais freqüência em um item do internetês?

Para respondermos a essa questão optamos pelo item '*td'* que possui freqüência de 760 ocorrências, sendo a 12ª palavra mais freqüente entre as do internetês. Foram feitas concordâncias de *'td'* no corpus, com o WordSmith Tools Concord, para verificar os usos dessa palavra em contexto. A escolha deste item foi motivada pelo fato de ser potencialmente ambígua, pois podem significar 'tudo', 'todo', 'toda', 'todos' e 'todas'.

O quadro abaixo resume as freqüências de ocorrência de cada sentido da palavra *'td'*:

Sentido	Total de ocorrências	porcentagem
Tudo	571	75%
Todo	137	18%
Toda	36	5%
Todos	08	1%
Todas	08	1%

Quadro 15: total de ocorrências para cada sentido do item *td.*

Assim, pudemos observar que *'td*' com sentido de 'tudo' apresentou uma freqüência maior de ocorrências com 75% (571/760). Já *'td*' com sentido de 'todo' apresentou 18% (137/760) das ocorrências. Portanto, para verificarmos se há padronização na linguagem da internet e se esses padrões contribuem para os sentidos do item *'td'* optamos por abordar somente os dois padrões mais freqüentes encontrados no corpus, ou seja, *'td'* com sentido de 'tudo' e 'todo'.

Em seguida foram analisados os agrupamentos[21] (*clusters*) selecionados, com a finalidade de classificá-los em porções de acordo com a sua freqüência. Essa observação incluiu a necessidade da descrição dos sentidos associados com as colocações recorrentes.

3.3.1 Padrões do item *'td'* com sentido de 'tudo'

Em seguida, apresentaremos os colocados, isto é, as palavras que estão ao redor do nódulo, neste caso, as palavras que ladeiam o item '*td'* quando esse apresentou o sentido de 'tudo'. Há uma maior incidência de padrões para esse sentido, já que a palavra 'tudo' tem um sentido mais abrangente. A tabela a seguir

[21] Esses agrupamentos foram formados pelas cinco palavras (colocados) antepostas ao nódulo e as cinco palavras (colocados) pospostas ao nódulo.

mostra os principais colocados que apresentaram uma freqüência maior de ocorrências:

Item	Total	E	D	E5	E4	E3	E2	E1	*	D1	D2	D3	D4	D5
Q	208	94	114	24	16	15	25	14	0	32	16	25	27	14
E	147	74	73	10	15	18	12	19	0	14	14	14	19	12
EU	97	35	62	7	15	8	5	0	0	8	21	16	5	12
O	81	35	46	9	9	11	5	1	0	7	12	11	5	11
DE	76	44	32	4	5	4	4	27	0	3	5	4	11	9
BOM	75	13	62	0	3	4	6	0	0	42	11	6	3	0
PRA	75	40	35	9	3	5	8	15	0	5	10	10	5	5
MAS	61	42	19	6	4	6	3	23	0	8	5	1	0	5
BEM	53	29	24	5	6	9	1	8	0	4	10	3	4	3
BAUM	51	14	37	0	2	5	6	1	0	28	5	2	0	2

Quadro 16: freqüência dos colocados em torno do nódulo '*td'* (tudo).

O quadro acima apresenta primeiramente, os colocados em torno do nódulo *'td'* que apresentaram sentido de 'tudo', em seguida apresenta-se a soma do total de ocorrências tanto na posição à direita e à esquerda do nódulo. Na coluna 'E' apresenta-se o total do colocado à esquerda do nódulo e na coluna 'D' apresenta-se o total de ocorrências à direita do nódulo. Por fim, apresentam-se o total para cada posição, ou seja, quinta palavra à esquerda do nódulo (E5), quarta palavra à esquerda do nódulo (E4) e assim por diante. O mesmo ocorrendo à direita do nódulo: quinta palavra à direta do nódulo (D5), quarta palavra à direita do nódulo (D4) e assim por diante. A seguir trataremos dos agrupamentos '*td* bem' (tudo bem), *'td* bom[22]' (tudo bom) e *'td baum*[23]*'* (tudo bem/bom) que apresentaram uma freqüência maior de ocorrências.

Padrões - *'td* bem', *'td* bom' e *'td baum'*

[22] Nota-se que a palavra 'bom' consta dicionarizada como adjetivo, mas os jovens utilizam em lugar de advérbio.
[23] '*Baum'* é uma variante da grafia de bom/bem.

Os colocados agrupam-se em vários conjuntos lexicais, sendo que o maior deles é formado pelo advérbio 'bem/*bom/baum'* posposto ao nódulo com 21% (120/571) das ocorrências. Estes conjuntos comumente aparecem com intuito de interagir com os participantes, cumprimentando os usuários ou concordando com uma dada situação. O quadro a seguir destaca alguns exemplos dos agrupamentos:

14	do meu fofuxo! olá pessoal,	td	bem? nossa, andei sumida
16	entem, hein!!!!! oi genteee,	td	bom? nossa meu, vcs viraram
23	gostem::: oi genteeeeee!!!!	td	baum com vcs??? cumigu td b
35	naum era neto dela * outra,	td	bem, q o bentinho era meu mel
57	d bem com vcs? comigo tah	td	bem, bom sabe aquela menina
61	pessoalzinhu do meu bloguxo	td	baum com vcs? espero q sim.....
93	confiram olá pessoaaaaal,	td	baum? nossa gente, mais q triste
120	carinhosos, fike c deus oie	td	bem??? parabens pelo seu niver
131	e 7 anos do meu sobrinho...	td	bem eu tava meio excluída da fes
154	virou festa neh?! huahuahua...	td	bem, eu sou chata mesmo e daí...
174	maluca, d kem nem leu o post!	td	bem q vai ter mt idiotice naquela
193	a casa?? huahauahuaahu.....	td	bem, td bem... td pelo césar e o
201	e vai faze mta falta pra mim!	td	bem q tem icq , tel , carta, msn e
209	welcome!! aguardem! oii gnt!!	td	bom?esse link é da minha kerida
225	!!!! desse meu viver! oii gente	td	bom com vc?? poxaa to mtooo
250	uma de menina ballet sem vc??	td	bem q eh soh 3 meses na espera
299	hihi.. uses for all .... olá pessoal	td	bom com vcs? espero q sim já q
310	uss dedicados...ahan logicoo!!	td	bem q o fime deu um pouco de s
327	falam seu nome, sua idade..... .	td	bem, eu jah colokeias no meu perf
344	vidar eu... teve festinha ontem..	td	bem q uma tia mal comida é foda

Quadro 17: linhas de concordância com os padrões 'td bem', 'td bom' e 'td baum'.

Padrões - 'mas/+ *td* bem*', '*mas/+ *td* bom', 'mas/+ *td baum*'

Destacamos esses padrões para mostrar casos em que o colocado pode ser uma palavra ('mas') ou sinal ('+'). Note que o sinal de adição é em si um caso interessante,

pois é tanto uma forma reduzida da palavra ‘mais’ e significa ‘mas’; sendo, portanto, ao mesmo tempo um caso de redução (de toques) quanto de ampliação, pois se tomarmos sua escrita por extenso (‘mais’ – considerando a pronúncia da palavra), temos o acréscimo de um toque (dos quatro caracteres de ‘mais’ para os três de ‘mas’). Porém, como estamos tratando do processo de grafia (e não de pronúncia), é mais coerente tratá-lo como um caso de redução. Assim, os colocados freqüentes antepostos ao nódulo encontrado no corpus são ‘mas/**+** (‘mas’ em internetês pode ser substituído por *mais* ou o sinal de ‘*+’)*. Esses padrões apresentaram 13% (71/571) dos casos e está relacionado a um conformismo, ironia sobre algo ou alguém, ou adversidade diante de uma situação, alguns dos quais aparecem nas linhas de concordância a seguir:

25	q vou fazer (qr dizer, tenhu +-).. +	td	bem pq tem tempo ainda.....
38	GI..mas nenhuma comentou...mas	td	bem nunca + deixo msg aki pra nenh
41	... ñ fui com a cara... mas ateh aih	td	bem.. pq eu (normalmente) ñ dexo d
83	manha ela vai me abandonar, mais	td	bem nem ligo msm vai chover os quatro
101	do meu padrasto q eu adoro! mas	td	bem....mt legal e amanha eu vo viaja!!!
106	nha migaa..meio distante assim. +	td	bemmmmm.... bju pra todo mundo.....
125	tendo muito bem esses dias...mas	td	bem....te amodoro mtoooo.....
127	eee pessoalzinho do meu bloguxo	td	bem com vcs? Ai espero q sim né p
128	manha ela vai me abandonar...mais	td	bem nem ligo msm vai chover os qua
130	r vagas se inscrevam!!!! Bjusss!!	Td	bom?? Comigo td.. Eai novidades
133	nda comparadu a 1 ano atrás mas	td	bem, apesar do stress q rolou
134	ninguem merece.....+	td	bem, bom agora to aki numa Lan
137	...nd dura pra sempre, neh...=(.. +	td	baum, vc volta logo, neh?!!!...(diz q s
157	mer e me arrumar ainda. uiaaa.. +	td	bem pq qse ngm comenta msm =(...
162	...Daninha: nem me liga mais , mais	td	bem te adoru do msm jeito... Licca:
178	haha...vc roubou minha kminha, mas	td	bem...ti doluuu... Juju e Gi...eu sou
199	gent bunita... mais pakeras... mais	td	.. hehehe... Imagina se eu num fikei
203	barrada lá fora do q convidado, mas	td	bem... faz parte...rs Enfim, por incriv
204	bom as férias em fala seriu!!? mais	td	bem é preciso!! Nossa eu não tenhu
208	or e carinhu q eu tinha por elas mas	td	baum neh.,..... Ai eu vou te falar eu to

Quadro 18: linhas de concordância com os padrões ‘mas/+ *td* bem*’, ‘*mas/+ *td* bom’, ‘mas/+ *td baum*’.

Ao observarmos as linhas de concordância acima pudemos notar que a posição desses padrões não é aleatória, isto é, encontram-se frequentemente antepostas ao nódulo.

3.3.2 Padrões do item '*td'* com sentido de 'todo'

Após analisarmos as linhas de concordância e observarmos os sentidos em contexto, obtivemos para o item *'td'* com sentido de 'todo', os colocados mais freqüentes, quais sejam:

item	Total	E	D	E 5	E4	E3	E2	E1	*	D1	D2	D3	D4	D5
MUNDO	84	6	78	0	3	0	1	2	0	78	0	0	0	0
Q	68	30	38	10	7	6	0	7	0	2	16	6	6	8
EU	28	14	14	2	4	3	5	0	0	0	1	5	4	4
PRA	28	20	8	0	1	3	2	14	0	0	1	2	3	2
A	24	15	9	5	2	3	1	4	0	0	3	3	2	1
MUNDU[24]	23	1	22	1	0	0	0	0	0	22	0	0	0	0
O	23	12	11	1	3	2	6	0	0	3	1	4	0	3
DE	21	16	5	2	3	3	2	6	0	0	2	1	2	0
EH	21	11	10	2	2	3	4	0	0	0	2	2	2	4
DIA	16	6	10	1	0	0	1	4	0	7	2	1	0	0

Quadro 19: freqüência dos colocados em torno do nódulo '*td'* ('todo).

O quadro acima apresenta primeiramente, os colocados em torno do nódulo *'td'* que apresentaram sentido de 'todo', seguidos pelo total de ocorrências nas posições à direita e à esquerda do nódulo. Por fim, apresenta-se separadamente o total para cada posição.

[24] Em internetês a grafia da palavra mundo é comumente encontrada como '*mundu'.*

Ao observarmos as linhas de concordância, a primeira constatação é que há um padrão freqüente que é formado por um colocado que sucede *'td'* que apresenta o sentido de 'todo', neste caso, referimo-nos a '*mundu*/mundo'. Conforme mencionado, das 137 linhas de ocorrências obtidas no corpus com sentido de 'todo', o padrão '*td* + *mundu*/mundo' abrange 73% (100/137) das ocorrências. Esse padrão vem imediatamente antecedido pela preposição '*pra'* (para) em 15% (20/137) das ocorrências.

589	boa sort lah nu show du LP...!! Pra	td	mundu q entra aki- continuem comen
590	meu kbelu di novu !! =D hehe agora	td	mundu tah flnd q meu kblu tah LARA
591	..fuui la na ksa do meu avo, visitei	td	mundo la...nem fikei mt tempo entao
592	CraZy´s, Lucca, Kalilzinhu, bj pa	td	mundu q eu flei hj na net i q sintiu mi
593	adiantadinha um feliz anu novu pra	td	mundu ai viu... Meu ESPERU q u a
594	meuu, eu num vo aguentahhh.. tah	td	mundo indu emboraaaaaaa!! drogaaa,,
595	ajudaram a crescer... Obrigado por	td	mundo q fez parte da minah vida no
625	o q escreve entaum mil bjus pra	td	mundu amu v6 má... tudu
626	do q d brigas. Feliz Natal pra	td	mundo! Bjos especiais pra
627	vo mais!! Entaum Bjusss*** pra	td	mundo!! e especialll hehe pr
628	cisa nem convidar.. . isso eh pra	TD	mundo! haha... Bom vo indu.
630	ai da escola (ou seja praticamente	td	mundu) Mais ctz q uns migu
631	o tds vcs... E um feliz 2005 pra	td	mundo!!!! Morzinhuuuuuuuu te
632	mo nos presentear??? =] quase	td	mundo fazendo vest... diske o
633	hhh Beijinhus Nussa hj quase	td	mundo leva suspensao coletiva
634	lher dele!! hehehe bem simpático	td	mundo!! e o serginho grijsakj
636	antes de eu bota recadinhus pra	td	mundo eu soh queria fla assi
637	dexo recadus de final de ano pra	td	mundo!!!hehehehe Ah entau
638	e eu ja to morrendu d saudadi de	td	mundo!!! Meu...to super "enrol
639	q o Brasil não vai pra frente! Pq	td	mundo sabe q naum eh por isso

Quadro 20: linhas de concordância com os padrões do item 'td' com sentido de 'todo'.

A explicação para o constante uso do padrão '*td* + *mundu*/mundo' e '*pra* + *td* + *mundu*/mundo' é clara, pois os usuários referem-se a todos que acessam o *blog* para redigir uma mensagem ou normalmente quando os blogueiros mencionam um acontecimento, no qual amigos e/ou familiares fazem parte. Vale salientar que '*Td*

m*undu*/mundo' normalmente é aplicado ao "mundo particular" ou a todos que fazem parte do convívio dos blogueiros, porém ao mencionar assuntos políticos, econômicos e sociais referindo-se a todas as pessoas do globo terrestre a freqüência é menor.

Padrões '*td* + dia' e 'dia *td*'

Um outro conjunto vasto na lista de colocados é o que indica idéia de tempo. Embora a tabela de colocados não apresente uma freqüência alta para esses colocados, achamos importante destacá-los. Os agrupamentos são formados pelo colocado 'dia', anteposto ao nódulo e posposto ao nódulo. Os padrões *'td* dia' e 'dia *td'* referem-se frequentemente aos relatos do cotidiano, abordando temas que indicam a temporalidade limitada ou extensa, ou seja, o padrão de posposição possui sentido de um tempo extenso, considerado mais que um dia. Já no padrão de anteposição, o colocado 'dia' vem precedido do artigo definido 'o/*u*[25]' (o + dia + *td*). Esse padrão exprime sentido limitado há apenas um dia. A seguir apresentamos 12 linhas com os respectivos padrões:

637	msm, se já não bastava eu ter aula	td	dia d noite e aos sábados d tarde,
640	do risadas com as mininas...show	td	dia... bota e saia... huahauha
645	ontem...Bia e Suellen (q m atura	td	dia...brigado pela sua pacienci
646	. .nem tem o q te fla...agente c fla	td	dia hahahaha....bjus.. Mari: va
647	a verdade foi uma eternidade...	td	dia a mesma coisa... trampo,
657	...q ouviram as minhas besteiras	td	dia...durante 4 hrs! Lú...a pes
658	sábado, dormi o dia praticamento	td	! E hj vi as minhas vidas! Minha
679	...e tb eh uma chatice ir no hospital	td	santo dia neh! Ainda tipo
728	o, num tenhu nd pra fazer o dia	td	, ngm me convida pra sair +...
733	e os professores, poder fikr o dia	td	sem ter q se preoucupar com
734	o e axu q vo fica em casa u dia	td	sem faze nda!!hehehehe ainda
735	casa hiper cansada, andei o dia	td	e to meio sem grana tb. Sem

Quadro 21: linhas de concordância com os padrões '*td* + dia' e 'dia *td*'

[25] O artigo definido singular "o" é grafado em internetês como '*u*'.

Para o colocado ‘dia’ anteposto ao nódulo, as ocorrências apontaram para 8% (10/137), enquanto o colocado ‘dia’ anteposto ao nódulo apresentou 5% (6/137). Esse resultado mostra-nos que há uma tendência maior para o uso do colocado ‘dia’ anteposto ao nódulo. Isso ocorre pois, o blog é um registro sobre os relatos cotidianos que expressam tempo limitado, ou seja, descreve diariamente as experiências vividas pelos blogueiros.

Nesse capítulo, apresentamos uma amostra das palavras mais freqüentes do corpus de estudo e as palavras mais freqüentes em internetês selecionadas para as análises. Esse levantamento possibilitou-nos a observação dos fenômenos ocorridos na grafia do internetês, incluindo a freqüência das modificações e a verificação sobre a padronização léxico-gramatical.

Em seguida, apresentaremos as Considerações Finais, incluindo a discussão dos resultados.

Considerações Finais

Nesse capítulo, apresentaremos as considerações finais relativas a esta pesquisa, seguidas por algumas ponderações críticas e algumas sugestões para futuras pesquisas.

Conforme se postulou na Introdução, a realização deste estudo motivou-se pela necessidade de compreender as modificações ortográficas ocorridas no internetês, assim como, fornecer alternativas para explicar essas modificações no léxico, visto que não há pesquisa empírica a respeito desse assunto.

Estudos prévios apenas enfocam alguns itens (Possenti, 2006; Thurlow and Brown, 2006; Crystal, 2001; Jaffe, 2006; Othero, 2004), enquanto outros se voltam à questão do erro causado na norma padrão por interferência do internetês que os jovens internautas costumam cometer na escrita, ora pelas costumeiras abreviações, ora pela grafia fonética. Dentre as inúmeras críticas ao internetês, podemos citar Eduardo Martins, (2006:24) quando alega que "o aprendizado da escrita depende da memória visual: muita gente escreve uma palavra quando quer lembrar sua grafia. Se bombardearmos por diferentes grafias, muitos jovens, ainda em formação, tenderão à duvida". Outro argumento que reforça essa crítica é sustentado por Sergio Nogueira (2007) quando se refere, principalmente, à grafia fonética, defendendo que "temos uma ortografia vigente, que foi estabelecida em 1943, com uma pequena reforma em 1971, e é o que está valendo até hoje. É ali que estão as palavras como devem ser escritas. Quando você começa a inventar uma grafia, isso cria o pior dos vícios, que é o da memória visual. (...). O problema do internetês é que ele pode causar vícios irreversíveis em relação à ortografia".

Não há, porém, uma clara preocupação em abordar o assunto de forma mais sistemática verificando as freqüências dessas modificações e a freqüência de uso desses itens.

Por outro lado, o trabalho ora descrito pretendeu enfocar um o aspecto que a nosso ver é central a essa problemática, qual seja: quais as modificações ocorrem com mais freqüência nos itens em internetês e verificar se os padrões léxicos gramaticais

dos itens grafados em internetês estão associados a sentidos distintos, evitando assim a ambigüidade que teoricamente pode surgir quando um mesmo item significa vários outros; como por exemplo, o item *'td'* que substitui os sentidos de: 'tudo', 'todo', 'toda', 'todos' e 'todas.

Fizemos uso da Lingüística de Corpus para atingir nossos objetivos, bem como, estabelecermos se os padrões léxico-gramaticais são padronizados.

A seguir, são apresentados os resultados de acordo com cada questão da pesquisa.

Em primeiro lugar, objetivamos responder à primeira questão da pesquisa; quais as palavras são encontradas com maior freqüência nos corpora eletrônicos extraídos de blogs de jovens que utilizam o internetês? Para respondê-la, solicitamos à ferramenta WordSmith Tools a lista de palavras do corpus de estudo. Em seguida, realizamos um 'recorte' selecionando para a análise os 500 primeiros types mais freqüentes.

Em seguida, objetivamos responder a segunda questão da pesquisa; quais palavras caracterizam a utilização do internetês? Extraímos as linhas de concordâncias das 500 formas selecionadas para o estudo. Assim, obtivemos 35% das palavras grafadas em internetês e 62% das palavras grafadas conforme a norma padrão. Os 3% restantes foram descartados das análises, pois não atendiam aos critérios estabelecidos para a identificação das palavras em internetês (capítulo 2 seção 2.4).

Em terceiro lugar, após caracterizarmos as palavras do internetês, levantamos quais as modificações ocorrem com maior freqüência na grafia desses itens. A partir dessas análises, pudemos constatar que há vários fenômenos ocorrendo simultaneamente, porém a investigação focou somente nas abreviações (supressão de vogais, acentos e consoantes), nas substituições (letras e acentos), na extensão de palavras, nas classes gramaticais frequentemente modificadas, na economia de toques (keystrokes) e na quantidade de toques para a formação dos itens em internetês.

Os resultados demonstraram que dentre os fenômenos analisados, há uma incidência maior para as supressões de vogais, consoantes e acentos gráficos, isto é,

as abreviações em geral. Essas abreviações são tão notáveis no internetês que pode até ser verificada sem acesso a corpora e à metodologia da Lingüística de Corpus, porém, a verdadeira extensão e probabilidade dessa perda de letras/acento só puderam ser apreciadas a partir da evidência disponibilizada pelos corpora.

Dentre essas supressões, constatamos que há uma incidência maior para a eliminação de vogais e acentos gráficos. As consoantes são também eliminadas, porém em menor freqüência. De acordo com Thurlow and Brown (2006:19) as supressões de vogais/acentos ocorre, principalmente, pois "os usos de agrupamentos de consoantes têm um valor semântico maior que as vogais[26]". Para exemplificar tal fenômeno, destacamos o item *'kd'* (Cadê?), na qual a letra *'k'* sonoriza a sílaba 'ca' e a letra *'d'* sonoriza a sílaba 'de'. Por outro lado, pudemos também verificar que em muitas formas do internetês ocorre um aumento da extensão da palavra (número maior de caracteres), em comparação com a norma padrão. Isso parece estar relacionado à reprodução do som da palavra. Neste caso a motivação para mudança gráfica é menos clara, pois a primeira vista o internetês tem a função de simplificar e utilizar menos tempo para a digitação.

As substituições de letras/acentos, costumeiras no internetês, são formadas por uma grafia que, segundo alguns estudos (Thurlow and Brown, 2006; Androutsopoulos, 2000; Crystal, 2001) partilha de características da língua oral e da língua escrita, sempre adequado-a ao contexto. Isto é, muitos itens apresentam uma aproximação fonológica, caracterizando uma forte influência da oralidade na escrita; por exemplo, os itens *'vamu'* (vamos) e *'axu'* (acho). Além disso, verificamos que as substituições visam também à simplificação ortográfica, ou seja, a digitação de uma quantidade menor de caracteres (lei do menor esforço). Por exemplo, os itens '*ke*' (que) e '*axo'* (acho).

Com relação às classes gramaticais, pudemos observar que há uma incidência maior para as modificações ocorrerem em advérbios, pronomes e interjeições. Por outro lado, há várias criações de contrações (aglutinação de palavras) criadas na grafia do internetês, tipicamente, resultantes da influência da oralidade, tais como: '*agenti*' (a gente) e '*oq*' (o que). Vale ressaltar que classificamos o item '*fds*' (final/fim

[26] The use of consonant clusters (…), recognizing that consonants (.) usually have more detail/value than vowels.

de semana) como acrônimo, porém o mesmo fenômeno em inglês (Thurlow, 2006) é classificado como 'Initialism' (inicialismo ou iniciais[27]). O fenômeno das iniciais (inicialism) aparentemente é formado por palavras/expressões comuns, tal como: 'asap' (as soon as possible) que significa 'tão logo quanto possível'. Já com relação aos acrônimos (acronyms), Thurlow (2006:24) considera a formação a partir de nomes próprios, tal como: 'DI' (Detective Inspector) que significa 'Detetive Inspetor'.

A investigação considerou também a eliminação de toques (keystrokes) na passagem da grafia padrão para o internetês, bem como a quantidade de toques para a composição desses itens. Com relação à eliminação de toques, notamos que entre os 221 sentidos das formas do internetês, 45% foram eliminados um toque, ou seja, foi eliminado um acento/letra das palavras da norma padrão. Já com relação à composição dos itens, os resultados apontaram uma incidência maior para a formação de itens compostos por 3 toques.

Enfim, objetivamos responder a quarta questão da pesquisa; quais padrões léxico-gramaticais são mais comumente verificados no internetês? Para tanto, optamos pelo item *'td*, pois os contextos de ocorrência apresentaram vários sentidos. Desse modo, sendo um item ambíguo, deveria haver, no texto, sinais que indicassem qual o sentido que se deva dar a *'td'*. Ou seja, de acordo com a Lingüística de Corpus, deveria haver padrões de linguagem que tornassem claro qual o sentido pretendido pelo escritor do blog, para que não existisse ambigüidade, nem para quem escreve, nem para quem lê. Em outras palavras, os diferentes sentidos de *'td'* (tudo, todo, toda, todos e todas) aparecem sinalizados no texto/contexto por meio de padrões distintos.

Assim, em conformidade com a Lingüística de Corpus, pudemos verificar que, em geral, há um sentido 'default', 'não marcado' de *'td'*, que é o mais provável (tudo), ou seja, a ambigüidade potencial é dissipada de antemão. Esse sentido parece ser a escolha mais provável dos usuários de blog. Isso é o que na Lingüística de Corpus, chamamos de sistema probabilístico (Halliday, 1992, 1993; Beber Sardinha, 2004, 2006) que é o fato de que as formas e estruturas da língua terem uma certa probabilidade de ocorrência no contexto. Os usuários da língua conhecem essas probabilidades conscientemente ou não, de acordo com os gêneros a que têm acesso.

[27] Tradução da autora.

Desse modo, os usuários dos blogs devem ter em mente, pelo menos de modo subconsciente, as probabilidades de *'td'*, o que lhes permite processar (ler e escrever) os textos de blogs de modo rápido e eficiente, como o meio exige.

Após despendermos todas as análises mencionadas acima, poderíamos, talvez, inferir que as abreviações estejam baseadas no princípio da economia lingüística. Essa economia está ligada às máximas da brevidade e velocidade que parece ser motivada mais especialmente por demandas discursivas, tais como, a espontaneidade e a fluidez na interação social e, menos por obstáculos tecnológicos (Thurlow and Brown, 2006).

Essas abreviações ocorrem como um contrato instituído entre os grupos de participantes, destacando aqui os blogueiros, que ao reduzirem/modificarem a grafia estabelecida pela norma padrão, buscam uma identificação e interatividade através da escrita, formando assim uma comunidade lingüística. Além disso, o internetês é utilizado por indivíduos que pretendem defender o 'espírito de grupo', afirmando suas identidades sociais, desviando-se das formas convencionais e, dessa forma, diferenciam-se dos adultos.

Outro aspecto relevante diz respeito à influência da oralidade na escrita. Em consenso com Thurlow and Brown (2006:19) "os jovens 'escrevem tal como falam[28]' para estabelecer um registro mais informal" que ajude por sua vez na interação e, desse modo, criam uma identificação com o outro. Por isso, os internautas procuram expressar risadas (hehehe, hahaha) enfatizar partes de palavras para dar a idéia de entonação mais forte (mtooo, amuu).

Em suma, nossos resultados sobre as abreviações/modificações na grafia ratificam as observações gerais de Possenti (2006), segundo o qual o internetês "é apenas um conjunto de soluções ortográficas, (..) não ameaça a língua portuguesa". Por outro lado, nossa análise revelou algo interessante sobre a linguagem dos blogs e, sobretudo com relação à linguagem em geral que não foram antecipadas por Possenti; nesse caso, os padrões associados aos sentidos da forma *'td'* (tudo e todo). As transformações que verificamos ocorrer na linguagem, muitas vezes omitindo as vogais ou modificando a escrita não parecem alterar os sentidos dos agrupamentos e

[28] young people 'write it as if saying it' to establish a more informal register (Thurlow and Brown, 2006).

seus respectivos padrões em relação a seus equivalentes da norma culta. Os padrões verificados podem ser claramente remetidos a formas de expressão ratificadas pela norma padrão.

Apresentam-se, agora, algumas ponderações críticas sobre esse trabalho.

Como todo e qualquer trabalho de pesquisa este também possui algumas limitações. A primeira delas refere-se à escolha do registro. Uma vez que se pretende investigar a linguagem dos meios digitais, isto é, o internetês, existe a necessidade de utilização de vários registros (Chats, MSN, mensagens via celular, etc..) que atinjam o objetivo geral de estabelecer as freqüências dessas modificações. Um corpus mais variado poderia fornecer mais dados, na medida em que se objetivasse uma investigação mais sistemática das freqüências de modificações ou de uso e dos padrões léxico-gramaticais encontrados no internetês.

A segunda reside no fato de que a comunicação digital ocorre todo o tempo e que essas modificações, possivelmente, podem apresentar uma freqüência diferente da detectada. Dessa forma, consideramos necessário o desenvolvimento de outras pesquisas que abordem o mesmo tema, porém que a coleta consista de várias fontes.

Algumas propostas para futuras pesquisas:

Em primeiro lugar, poder-se-ia, por exemplo, explorar as freqüências de modificações ou características léxico-gramaticais de itens a partir de outros registros, como mencionado anteriormente, com a finalidade de compará-los para verificar se há uma incidência menor ou maior dessas modificações. Tal investigação contribuiria para a descoberta de novas modificações ocorridas na passagem da norma padrão para o internetês, bem como a freqüência em que elas ocorrem.

Em segundo lugar, poder-se-ia realizar uma descrição mais sistemática analisando todos os itens lexicais presentes no corpus, não somente dos quinhentos primeiros, como realizado nesse trabalho.

Finalmente, poder-se-ia, comparar o fenômeno da grafia internáutica de duas ou mais línguas diferentes; por exemplo, o internetês e o netspeak; o internetês e o chateo. Tal comparação demonstraria se ocorrem os mesmo fenômenos lingüísticos e se as modificações ocorrem na mesma freqüência.

Esperamos que este trabalho venha preencher uma lacuna importante ao analisar empiricamente as modificações na grafia do meio internáutico e a colaborar para um melhor entendimento do Internetês.

Referências Bibliográficas

Androutsopoulos, J.K., *Non-standard spellings in media texts: The case of German fanzines*, Journal of Sociolinguistics, 4(4), 514-533, 2000.

Aurélio, B. H., *Novo Dicionário Aurélio da Língua Portuguesa*. Curitiba: Positiva, 2ª edição – versão em CD- ROM.

Berber Sardinha, A. P. (org.). *A Língua Portuguesa no Computador*. Campinas-SP: Mercado das Letras, Fapesp, 2005.

Berber Sardinha, T. *Usando WordSmith Tools na investigação da linguagem*. DIRECT Paper 40. São Paulo / Liverpool, 1999.

Berber Sardinha, T., *Computador, corpus e concordância no ensino de léxico-gramática de língua estrangeira.* In V. Leffa (org.) *As palavras e sua companhia: o léxico na aprendizagem*. Pelotas: EDUCAT, Universidade Católica de Pelotas, 45-72 (2000a).

Berber Sardinha, T., *Lingüística de Corpus: histórico e problemática*. D.E.L.T.A., 16 (2), 323-367, (2000b).

Berber Sardinha, T., Computador, corpus e concordância no ensino de léxico-gramática de língua estrangeira. In V. Leffa (org.) *As palavras e sua companhia: o léxico na aprendizagem*. Pelotas: EDUCAT, Universidade Católica de Pelotas, 45-72, (2000a).

Berber Sardinha, T. *Lingüística de Corpus.* Barueri-S.P: Manole, 2004.

Biber, D. S. Conrad & R. Reppen. *Corpus linguistics: investigating language structure and use.* Cambridge: Cambridge University Press, 1998.

Biderman, M. T. C*., Dicionário Didático de Português*. São Paulo: Ática, 2ª edição, 1998.

Biber, D., *Longman grammar of spoken and written English.* Harlow: Longman, 1999.

Camara Jr., J. Mattoso. *Dicionário de lingüística e gramática referente à língua portuguesa*. Petrópolis-RJ: Vozes, 1998.

CHIZZOTTI, A., *Pesquisa em Ciências Humanas e Sociais*. São Paulo: Cortez Editora, 5ª edição, 2001.

Coutinho, I. L. *Pontos de gramática histórica*. Rio de Janeiro: Liv. Acadêmica, 5ª edição, 1962.

Crystal, D., *Language and the internet*, Cambridge: Cambridge University Press, 2001.

Faraco e Moura. *Língua e literatura*. São Paulo: Ática, 2000.

Firth, J., *Papers in linguistics – 1934-1951.* Oxford: Oxford University Press, 1957.

Halliday, M., *Corpus studies and probabilistic grammar*. In K. Aijmer & B. Altenberg (org.). *English corpus linguistics: studies in honour of Jan Svartvik*.London: Longman, 1991.

Halliday, M., *Quantitative studies and probabilities in grammar.* In: M. Hoey (Ed.), **Data* Description Discourse - Papers on the English language in Honour of John McH Sinclair on his Sixtieth Birthday* (1-25). London: HarperCollins. (1993).

Houaiss, A., *Dicionário eletrônico Houaiss da Língua Portuguesa.* Rio de Janeiro: Objetiva. Disponível no site: www.uol.com.br

Hunston, S. *Corpora in Applied Linguistic*. Cambridge: Cambridge University Press, 2002.

Jaffe, A., *Introduction: Non-standard orthography and non-standard speech*, Journal of Sociolinguistics, 4(4), 497-513, 2000.

Marcushi, L. A. e Xavier, A. C. (Orgs.). *Hipertexto e gêneros digitais: novas formas de construção do sentido.* Rio de Janeiro: Lucerna, 2005.

Marcushi, L. A. Gêneros textuais: definição e funcionalidade. In Dionísio A. P. et al Gêneros textuais & ensino; Rio de Janeiro: Lucerna, 2002.

Martins, Eduardo. *Revista Língua Portuguesa – a Revolução do Internetês.* Segmento, nº 5, 28, 2006.

McEnery, T. & A. Wilson. *Corpus Linguistics.* Edinburgh: Edinburgh University Press, 1996.

Michaeli, *Moderno Dicionário da Língua Portuguesa.* São Paulo: Melhoramentos. Disponível no site: www.uol.com.br.

Moita Lopes, L. P. *Contextos Institucionais em Lingüística Aplicada:* Novos Rumos. Intercâmbio, vol. 05, 1996. p. 03-14.

Othero, Gabriel A. *A Língua Portuguesa nas salas de b@te-p@po.* Berthier, 2005.

Pavão, Jadyr. *Revista Época – Qual é o melhor.* Globo, ed. 182, 2001 ou disponível no site: http://epoca.globo.com/edic/20011112/cult1b.htm.

Possenti, Sírio. *Revista Língua Portuguesa - a Revolução do Internetês.* Segmento, nº 5, 2006. p. 22-27.

Possenti, Sírio. *Discutindo a Língua Portuguesa – A Revolução do Internetês.* Escala Educacional, nº 2, 2006, p 28-33.

Schittine, Denise. *Blog: comunicação e escrita íntima na internet.* RJ, Civilização Brasileira, 2004.

Scott, M. (1997) *WordSmith Tools*. *Versão 3.* Oxford: Oxford University Press.

Sinclair, J. M. *Corpus, Concordance, Collocation***.** Oxford: Oxford University Press, 1991.

Stubbs, M. (1993). *British traditions in text analysis*: from Firth to Sinclair. In M. Baker, G. Francis & E. Tognini-Bonelli (orgs.), *Text and technology.* Philadelphia: Benjamins, 1-33.

Sztajnberg, A & Denise Del Re Filippo. *Bem Vindo à Internet.* Rio de Janeiro – RJ: Editora Brasport, 1999.

Thurlow, Crispin & Brown, Alex *Generation Txt? The sociolinguistics of young people's text-messaging.* Department of Communication, University of Washington, Seattle ou disponível no site http://extra.shu.ac.uk/daol/articles/v1/n1/a3/thurlow2002003-paper.html#backnote1.

Anexo I - Relação de Blogs Coletados para compor o corpus de estudo

1	www.rouuusinha.blogger.com.br
2	www.ndadahcerto.blogger.com.br
3	www.legally_lilu.blogger.com.br
4	http://legallyfashiongirl.zip.net
5	Http://concurso_blog_patty.zip.net
6	www.danixinhamiga.weblogger.terra.com.br/200405_danixinhamiga
7	http://my-sweet-die.bigflogger.com.br/?11
8	http://crisflora.blogs.sapo.pt/arquivo/948594.html
9	http://www._nine_.blogger.com.br/
10	http://www.arydelonge182.blogger.com.br/
11	http://www.carolzitcha.blogger.com.br/
12	http://www.bruninhasdream.blogger.com.br/
13	http://www.duas_amigas.blogger.com.br/
14	http://nihnesblog.weblogger.terra.com.br/
15	http://www.smilebehappy.blogger.com.br/
16	http://fotolog.terra.com.br/lakacurti
17	http://www.intsjulinha.blogger.com.br/
18	http://www.lally_pink.blogger.com.br/
19	http://www.allstar.blogger.com.br/
20	http://www.natyenroladinha.weblogger.terra.com.br/
21	http://sweetbymilly.weblogger.terra.com.br/
22	http://www.lonelygirlbylele.blogger.com.br/
23	http://www.paulinha_kevin.weblogger.terra.com.br/
24	www.cherry_prizinha.weblogger.terra.com.br/200506_cherry_prizinha_
25	http://hanagumi.weblogger.terra.com.br/index.htm
26	http://www.tabs.weblogger.terra.com.br/
27	http://www.sekaianime.blogger.com.br/
28	http://www.concurso_legally_cute.blogger.com.br/
29	http://chezim.flogbrasil.terra.com.br/
30	http://www.orkut.com/Community.aspx?cmm
31	http://concurso-cutedream.weblogger.terra.com.br/
32	http://www.bloglorena.weblogger.terra.com.br/
33	http://www.19_ruivinha_19.blogger.com.br/
34	http://www.aartedeamar.blogger.com.br/
35	http://pleasepleaseme.blogger.com.br/
36	http://www.algumasmentiras.weblogger.terra.com.br/
37	http://mahzinha_.weblogger.terra.com.br/
38	http://www.just-le.blogger.com.br/
39	http://www.inmid-air.blogger.com.br/

40	http://www.lus-blog.blogspot.com/
41	http://eu_e_eu_mesmo.weblogger.terra.com.br/index.htm
42	http://www.amobeijar.blogger.com.br/
43	http://ivaninhamg.blogspot.com/
44	http://www.mitie.blogger.com.br/
45	http://laravedder.blogspot.com/2002_01_06_laravedder_archive.html
46	http://www.tsuki_no_ohimesama.blogger.com.br/
47	http://www.minasdocatalano.blogger.com.br/
48	http://www.nessa-girl.blogger.com.br/
49	http://www.place-of-dreams.blogger.com.br/
50	http://www.maiara_nunes.weblogger.terra.com.br/index.htm
51	http://www.vidalokask8.weblogger.terra.com.br/
52	http://pequenoponei.blogspot.com/
53	http://www.noixehfoda.weblogger.terra.com.br/
54	http://www.meninjas--.weblogger.terra.com.br/
55	http://jeanherbert.blogspot.com/
56	http://jaqzinha.weblogger.terra.com.br/
57	http://douglasoda.weblogger.terra.com.br/
58	http://hiromi.weblogger.terra.com.br/index.htm
59	http://www.karollokynha.weblogger.terra.com.br/
60	http://www.efeitotequila.blogger.com.br/
61	www.lunasolista.weblogger.com.br
62	http://www.merzinha_4ever.blogger.com.br/
63	http://www.marcelawalois.weblogger.terra.com.br/
64	http://www.shadows2.blogger.com.br/
65	http://karoll_angel.bigblogger.com.br
66	www.flogs.com.br/vanessinhajoplin
67	http://essenciagotica.weblogger.com.br
68	www.flogs.com.br/vanessinhajoplin
69	http://essenciagotica.weblogger.com.br
70	www.flogs.com.br/crisbh
71	http://karoll_angel.bigblogger.com.br
71	www.gotikaforever.weblogger.com.br
73	www.nikita-skate.theblog.com.br
74	www.flogs.com.br/crisbh
75	http://karoll_angel.bigblogger.com.br
76	www.shadowsnight.weblogger.terra.com.br
77	http://www.rocknroll.bigblogger.com.br
78	http://www.clanif.blogger.com.br/
79	http://www.conexaobonjovi.blogger.com.br/
80	www.brumaximus.blogger.com.br
81	http://www.nataliaoliveira.weblogger.terra.com.br/
82	http://www.lokinhamor.blogger.com.br/
83	http://msablog.blig.ig.com.br
84	http://www.prizinha12.weblogger.terra.com.br/index.htm
85	http://www.baianablog.blogger.com.br/

86	http://www.legally_tatha.blogger.com.br
87	http://www.marieamigos.blogger.com.br/
88	http://putons.weblogger.terra.com.br/
89	http://www.tsblog.blogger.com.br/
90	http://www.gatasagente_magina_rs.blogger.com.br/
91	http://www.mandinha_patty.blogger.com.br/
92	http://mazinha-love.weblogger.terra.com.br/
93	http://www.thepudim.weblogger.terra.com.br/index.htm
94	http://www.evemiga.weblogger.terra.com.br/
95	http://prendinhamari.flogbrasil.terra.com.br/
96	http://www.isaisacao.weblogger.terra.com.br/index.htm
97	http://www.artedassombras.weblogger.terra.com.br/
98	http://www.manuzinhahh.weblogger.terra.com.br/index.htm

Anexo II – As 500 palavras mais freqüentes do corpus de Estudo

N	Word	Freq.	%	N	Word	Freq.	%	N	Word	Freq.	%
1	Q	3.993	2,89	168	Até	114	0,08	334	Nós	55	0,04
2	Eu	3.443	2,49	169	Post	113	0,08	335	Outra	55	0,04
3	E	3.306	2,40	170	Adoro	112	0,08	336	Sao	55	0,04
4	A	2.638	1,91	171	Boa	112	0,08	337	Domingo	54	0,04
5	O	1.889	1,37	172	Deu	112	0,08	338	Essas	54	0,04
6	De	1.753	1,27	173	Dias	112	0,08	339	Falando	54	0,04
7	Pra	1.592	1,15	174	Sim	112	0,08	340	Fikar	54	0,04
8	D	1.161	0,84	175	Dos	111	0,08	341	Fotos	54	0,04
9	Eh	1.124	0,81	176	Ia	111	0,08	342	Lay	54	0,04
10	Mas	1.103	0,80	177	Fla	107	0,08	343	Mew	54	0,04
11	Vc	1.037	0,75	178	Amiga	106	0,08	344	Preciso	54	0,04
12	Um	1.034	0,75	179	Acho	105	0,08	345	Voltar	54	0,04
13	Mais	946	0,69	180	Meio	105	0,08	346	Cada	53	0,04
14	Naum	873	0,63	181	Nada	104	0,08	347	Contar	53	0,04
15	Meu	853	0,62	182	Menos	103	0,07	348	Férias	53	0,04
16	Da	828	0,60	183	Seja	103	0,07	349	Ficar	53	0,04
17	Uma	816	0,59	184	Axo	102	0,07	350	Fim	53	0,04
18	Com	803	0,58	185	Pessoa	102	0,07	351	Fora	53	0,04
19	Do	796	0,58	186	Fala	101	0,07	352	Indo	53	0,04
20	Se	769	0,56	187	Ae	100	0,07	353	Qdo	53	0,04
21	Td	760	0,55	188	Galera	100	0,07	354	Sair	53	0,04
22	Bom	747	0,54	189	Né	100	0,07	355	Amigas	52	0,04
23	Aki	742	0,54	190	Rs	99	0,07	356	São	52	0,04
24	Me	725	0,53	191	Tipo	99	0,07	357	Tpo	52	0,04
25	Na	708	0,51	192	Dizer	98	0,07	358	Dessa	51	0,04
26	Por	702	0,51	193	Minhas	98	0,07	359	Filme	51	0,04
27	No	681	0,49	194	Amor	97	0,07	360	Saum	51	0,04
28	Foi	662	0,48	195	Faz	97	0,07	361	Weblogger	51	0,04
29	Pq	642	0,47	196	Pode	97	0,07	362	Disso	50	0,04
30	Neh	639	0,46	197	Taum	97	0,07	363	Entao	50	0,04
31	Num	638	0,46	198	Cara	96	0,07	364	Fico	50	0,04
32	Mto	625	0,45	199	Amigos	95	0,07	365	Hahahahaha	50	0,04
33	To	623	0,45	200	Faze	95	0,07	366	Kra	50	0,04
34	Em	618	0,45	201	Escola	94	0,07	367	Mae	50	0,04
35	É	594	0,43	202	Dae	92	0,07	368	Olha	50	0,04
36	Minha	586	0,42	203	Dar	92	0,07	369	Tão	50	0,04
37	Que	581	0,42	204	Moh	92	0,07	370	Dah	49	0,04
38	As	579	0,42	205	Nda	92	0,07	371	Enfim	49	0,04
39	I	568	0,41	206	Saudades	92	0,07	372	Então	49	0,04
40	Ai	554	0,40	207	Comigo	91	0,07	373	Estão	49	0,04
41	Bem	542	0,39	208	Linda	89	0,06	374	Grande	49	0,04

42	Vcs	534	0,39	209	Todo	89	0,06	375	Novidades	49	0,04
43	Tah	495	0,36	210	M	88	0,06	376	Pelos	49	0,04
44	Vai	490	0,36	211	Ñ	87	0,06	377	Esses	48	0,03
45	Tem	489	0,35	212	Uns	87	0,06	378	Lu	48	0,03
46	Nem	471	0,34	213	Ateh	86	0,06	379	Nome	48	0,03
47	U	448	0,32	214	Br	85	0,06	380	Nova	48	0,03
48	Vou	445	0,32	215	Poko	85	0,06	381	Parece	48	0,03
49	Hj	424	0,31	216	Show	85	0,06	382	Voltei	48	0,03
50	Isso	415	0,30	217	Vi	85	0,06	383	Algo	47	0,03
51	Soh	415	0,30	218	Foto	84	0,06	384	Concurso	47	0,03
52	Como	398	0,29	219	Hahahaha	84	0,06	385	Final	47	0,03
53	Gente	396	0,29	220	Hein	83	0,06	386	Primeiro	47	0,03
54	Mt	386	0,28	221	Deus	82	0,06	387	Ahh	46	0,03
55	Sei	385	0,28	222	Nom	82	0,06	388	Amanha	46	0,03
56	Dia	369	0,27	223	Pai	82	0,06	389	Dps	46	0,03
57	Tava	354	0,26	224	Pela	82	0,06	390	Obrigada	46	0,03
58	Te	336	0,24	225	Quero	82	0,06	391	Pensar	46	0,03
59	Agora	333	0,24	226	Dela	81	0,06	392	Pessoal	46	0,03
60	Sem	333	0,24	227	Super	81	0,06	393	Povo	46	0,03
61	P	332	0,24	228	Volta	81	0,06	394	Quando	46	0,03
62	C	330	0,24	229	Gent	80	0,06	395	Amizade	45	0,03
63	Tb	326	0,24	230	Prova	80	0,06	396	Claro	45	0,03
64	Nao	322	0,23	231	Http	79	0,06	397	Disse	45	0,03
65	Não	320	0,23	232	Ve	79	0,06	398	Natal	45	0,03
66	Ela	318	0,23	233	Aula	78	0,06	399	Você	45	0,03
67	Ainda	310	0,22	234	Fikei	78	0,06	400	X	45	0,03
68	Coisa	300	0,22	235	Kem	77	0,06	401	Amigo	44	0,03
69	Fui	291	0,21	236	Ksa	77	0,06	402	Diferente	44	0,03
70	Os	291	0,21	237	Lindo	77	0,06	403	Mtas	44	0,03
71	Ele	285	0,21	238	Aqui	76	0,06	404	Nee	44	0,03
72	Esse	274	0,20	239	Noite	76	0,06	405	Template	44	0,03
73	Blog	270	0,20	240	S	76	0,06	406	Valeu	44	0,03
74	Ki	270	0,20	241	Fica	75	0,05	407	Verdade	44	0,03
75	Nu	267	0,19	242	Gosto	75	0,05	408	Vlw	44	0,03
76	La	263	0,19	243	Niver	75	0,05	409	Blogger	43	0,03
77	Vo	261	0,19	244	Hahaha	74	0,05	410	Fazendo	43	0,03
78	Sempre	260	0,19	245	Oi	74	0,05	411	Hr	43	0,03
79	Di	258	0,19	246	Pois	74	0,05	412	Ngm	43	0,03
80	Só	254	0,18	247	Quase	74	0,05	413	Aulas	42	0,03
81	Tempo	253	0,18	248	Antes	73	0,05	414	Gnt	42	0,03
82	Hehe	252	0,18	249	Especial	73	0,05	415	Nessa	42	0,03
83	Ta	247	0,18	250	Intaum	73	0,05	416	Pah	42	0,03
84	Lah	245	0,18	251	Pena	73	0,05	417	Realmente	42	0,03
85	Mim	243	0,18	252	Vezes	73	0,05	418	Seus	42	0,03
86	Ti	241	0,17	253	Fez	72	0,05	419	Sobre	42	0,03
87	Fazer	237	0,17	254	Hora	72	0,05	420	Daki	41	0,03
88	Hehehe	228	0,17	255	Nunca	72	0,05	421	Embora	41	0,03
89	Jah	228	0,17	256	Queria	72	0,05	422	Escrever	41	0,03
90	Ser	228	0,17	257	Hehehehe	71	0,05	423	Ferias	41	0,03
91	Ou	227	0,16	258	Lado	71	0,05	424	Fiko	41	0,03
92	K	226	0,16	259	Mal	71	0,05	425	Foda	41	0,03

93	Viu	225	0,16	260	Eles	70	0,05	426	Merece	41	0,03
94	Nossa	224	0,16	261	Triste	70	0,05	427	Onde	41	0,03
95	Vida	224	0,16	262	Net	69	0,05	428	Outros	41	0,03
96	Ano	221	0,16	263	Fomos	68	0,05	429	Qndo	41	0,03
97	Ver	221	0,16	264	Heheh	68	0,05	430	Sendo	41	0,03
98	Assim	219	0,16	265	Lugar	68	0,05	431	Será	41	0,03
99	Du	218	0,16	266	Miga	68	0,05	432	Terra	41	0,03
100	Essa	215	0,16	267	Ok	68	0,05	433	V	41	0,03
101	Msm	215	0,16	268	Teve	68	0,05	434	Vontade	41	0,03
102	T	215	0,16	269	Blz	67	0,05	435	Ahhh	40	0,03
103	Mtu	213	0,15	270	Coments	67	0,05	436	Akela	40	0,03
104	Para	212	0,15	271	Festa	67	0,05	437	Bjos	40	0,03
105	Depois	210	0,15	272	Genti	67	0,05	438	Deixa	40	0,03
106	Tbm	208	0,15	273	Tanto	67	0,05	439	Duas	40	0,03
107	Ter	207	0,15	274	Vim	67	0,05	440	Flo	40	0,03
108	Amo	204	0,15	275	Www	67	0,05	441	Hoje	40	0,03
109	Ah	201	0,15	276	Aí	66	0,05	442	Po	40	0,03
110	Ja	201	0,15	277	Todas	66	0,05	443	Ruim	40	0,03
111	Feliz	198	0,14	278	Dele	65	0,05	444	Sai	40	0,03
112	Semana	196	0,14	279	Dpois	65	0,05	445	Tal	40	0,03
113	Sabe	189	0,14	280	Posso	65	0,05	446	Tao	40	0,03
114	Coisas	188	0,14	281	Qm	65	0,05	447	Alguem	39	0,03
115	Mundo	185	0,13	282	Tá	65	0,05	448	Estar	39	0,03
116	Tudo	185	0,13	283	Tow	65	0,05	449	Feira	39	0,03
117	Issu	184	0,13	284	Adorei	64	0,05	450	Fik	39	0,03
118	Tinha	177	0,13	285	Bjaum	64	0,05	451	Msn	39	0,03
119	Pro	172	0,12	286	Estava	64	0,05	452	Sabem	39	0,03
120	Estou	171	0,12	287	Us	64	0,05	453	Bastante	38	0,03
121	So	171	0,12	288	Bjus	63	0,05	454	Chega	38	0,03
122	Tenho	170	0,12	289	Oq	63	0,05	455	Consegui	38	0,03
123	Espero	169	0,12	290	Outro	63	0,05	456	Fds	38	0,03
124	Ke	169	0,12	291	Parte	63	0,05	457	Jeito	38	0,03
125	Mesmo	168	0,12	292	Praia	63	0,05	458	Kero	38	0,03
126	Era	167	0,12	293	Alguma	62	0,04	459	Mó	38	0,03
127	Muito	167	0,12	294	Algumas	62	0,04	460	Passou	38	0,03
128	Casa	164	0,12	295	Mtooo	62	0,04	461	Sab	38	0,03
129	N	163	0,12	296	Passei	62	0,04	462	Boas	37	0,03
130	Seu	160	0,12	297	Segunda	62	0,04	463	Causa	37	0,03
131	Cum	158	0,11	298	Tarde	62	0,04	464	Escreve	37	0,03
132	Das	158	0,11	299	Flog	61	0,04	465	Faço	37	0,03
133	Melhor	157	0,11	300	Mi	61	0,04	466	For	37	0,03
134	Todos	156	0,11	301	Qnd	61	0,04	467	Importante	37	0,03
135	Novo	152	0,11	302	Saber	61	0,04	468	Mãe	37	0,03
136	Ao	151	0,11	303	Sexta	61	0,04	469	Mary	37	0,03
137	Pessoas	148	0,11	304	Tv	61	0,04	470	Monte	37	0,03
138	Ontem	147	0,11	305	Dani	60	0,04	471	Nosso	37	0,03
139	Nd	142	0,10	306	Está	60	0,04	472	Outras	37	0,03
140	Pelo	142	0,10	307	Falta	60	0,04	473	Posta	37	0,03
141	Tô	142	0,10	308	Pc	60	0,04	474	Saudade	37	0,03
142	Lá	138	0,10	309	Apesar	59	0,04	475	Sinto	37	0,03
143	Legal	137	0,10	310	Dexa	59	0,04	476	Carro	36	0,03

144	Meus	135	0,10	311	Elas	59	0,04	477	Comenta	36	0,03
145	Pa	135	0,10	312	Nunk	59	0,04	478	Dai	36	0,03
146	Amu	133	0,10	313	Passa	59	0,04	479	Entra	36	0,03
147	Axu	133	0,10	314	Tive	59	0,04	480	Fosse	36	0,03
148	Fiz	128	0,09	315	Deixar	58	0,04	481	Maior	36	0,03
149	Tds	128	0,09	316	Logo	58	0,04	482	Momento	36	0,03
150	Postar	127	0,09	317	Vamos	58	0,04	483	Ninguem	36	0,03
151	Entaum	126	0,09	318	Anos	57	0,04	484	Numa	36	0,03
152	Sua	126	0,09	319	Cmg	57	0,04	485	Passado	36	0,03
153	Agente	123	0,09	320	Esta	57	0,04	486	Qd	36	0,03
154	Falar	122	0,09	321	Xd	57	0,04	487	Rsrs	36	0,03
155	Ir	122	0,09	322	Aconteceu	56	0,04	488	Sabado	36	0,03
156	Mta	122	0,09	323	Amanhã	56	0,04	489	Talvez	36	0,03
157	Vez	122	0,09	324	Certo	56	0,04	490	Tanta	36	0,03
158	Aih	120	0,09	325	Passar	56	0,04	491	Vamu	36	0,03
159	Ate	120	0,09	326	Pouco	56	0,04	492	À	35	0,03
160	Nos	120	0,09	327	Agenti	55	0,04	493	Dificil	35	0,03
161	Tenhu	120	0,09	328	Fiquei	55	0,04	494	Ih	35	0,03
162	Umas	120	0,09	329	Foram	55	0,04	495	Meninas	35	0,03
163	Vem	120	0,09	330	Girls	55	0,04	496	Mundu	35	0,03
164	Nois	119	0,09	331	Merda	55	0,04	497	Nas	35	0,03
165	Quem	119	0,09	332	Mtoo	55	0,04	498	Pior	35	0,03
166	Sou	117	0,08	333	Nesse	55	0,04	499	Posto	35	0,03
167	Já	115	0,08					500	Sabia	35	0,03

Anexo II: listas de palavras contendo as 500 formas em ordem de freqüência no corpus de estudo, seguidas pelo número de ocorrências e pela porcentagem relativa ao total do corpus.