representatividade usando-se uma fórmula que determina o tamanho mínimo necessário de uma amostra para representar a população de determinada categoria linguística ou de determinado objeto de estudo. Faz-se necessário determinar uma variedade X de tipos de texto para uma dada população; com base nessa determinação, vai-se aproximando do estabelecimento de sua representatividade. O texto seria, em suma, um conjunto de princípios para se atingir a representatividade de um *corpus*.

Embora sua publicação original tenha sido no *Literary and Linguistic Computing*, vol. 8, nº 4, em 1993, o texto que ora apresentamos continua atual e oferece um excelente embasamento em termos de metodologia de pesquisa e *representatividade*. De agora em diante, acreditamos que o texto possa circular mais ainda, visto que até agora, pelo que apuramos, esta é a sua primeira tradução acadêmica para o português do Brasil.

Referências:

BERBER SARDINHA, Tony (2004). *Lingüística de Corpus*. Barueri, SP: Manole.
BIBER, Douglas. (1998) *Corpus Linguistics*: Investigating language Structure and Use.Cambridge, UK: Cambridge University Press. 301 p.

Representatividade em planejamento de *corpus*[1]

Douglas Biber[2]

Tradução de Paula Marcolin
Revisão de Fabiano Bruno Gonçalves
e Susana de Azeredo Gonçalves
Revisão técnica de Maria José B. Finatto

Resumo

O presente artigo aborda uma série de questões relacionadas à obtenção da "representatividade" no planejamento de um *corpus* linguístico; essas questões englobam: a discussão sobre o que significa "representar" uma língua; a definição de população; a estratificação *versus* a amostra proporcional de uma língua; amostras dentro de textos; e questões relacionadas ao tamanho necessário da amostra (número de textos) de um *corpus*. O artigo destaca, dentre diversas maneiras, que características linguísticas podem ser distribuídas dentro de textos e entre textos diferentes; são analisadas as distribuições de diversas características particulares e são discutidas as implicações dessas distribuições para o planejamento do *corpus*.

Este artigo defende que a pesquisa teórica deveria ser anterior ao planejamento do *corpus*, identificando, então, os parâmetros situacionais que variam entre os textos de uma comunidade discursiva e também os tipos de características linguísticas que serão examinadas no *corpus*. Estas considerações teóricas devem ser complementadas por investigações empíricas da variação linguística de um *corpus*-piloto como base para decisões específicas de amostras. A construção efetiva de um *corpus* ocorreria, então, em ciclos: o planejamento original baseado em análises teóricas e de estudo-piloto seguido de uma coleta de textos, por investigações empíricas mais detalhadas da variação linguística, e por uma revisão do planejamento.

1 Original: **Representativeness in Corpus Design.** Literary and Linguistic Computing, Vol. 8, No. 4, 1993, p. 243- 257. Tradução para o Brasil com especial permissão do autor para esta publicação.

2 Departamento de Inglês, Northern Arizona University.

1. Considerações gerais

Algumas das primeiras considerações na construção de um *corpus* dizem respeito ao planejamento geral: por exemplo, os tipos de textos que o compõem, o número de textos, a seleção de determinados textos, a seleção de amostras retiradas de textos e o tamanho das amostras de texto. Cada uma, portanto, implica uma decisão sobre a amostra, seja ela consciente ou não.

A utilização de *corpora* eletrônicos oferece uma base empírica sólida para ferramentas e descrições linguísticas de uso geral, permitindo a análise de uma dimensão que, de outra forma, não seria possível. Todavia, um *corpus* deve ser "representativo" para ser usado adequadamente como base de generalizações acerca de uma língua como um todo; por exemplo, os dicionários baseados em *corpus*, as gramáticas e os etiquetadores gramaticais em geral são aplicativos que necessitam de uma base representativa (Biber, 1993b).

Normalmente, os pesquisadores priorizam o tamanho das amostras como o aspecto mais importante para obter representatividade: o número de textos que devem ser incluídos no *corpus* e o número de palavras por amostra. Os livros sobre teoria de amostras, no entanto, ressaltam que o tamanho da amostra não é o fator mais importante na seleção de uma amostra representativa; antes disso, uma definição completa sobre a população-alvo e as decisões a respeito dos métodos de amostragem deveriam ser considerações prioritárias. A representatividade se refere ao quanto uma amostra inclui de toda a gama de variabilidade de uma população. No planejamento do *corpus*, a variabilidade pode ser considerada a partir das perspectivas situacionais e linguísticas, pois ambas são importantes na determinação da representatividade. Assim, o planejamento de um *corpus* pode ser avaliado com base no quanto ele abarca dos seguintes itens: (1) a variedade dos tipos de textos em uma língua e (2) a variedade de distribuições linguísticas em uma língua.

Qualquer seleção de textos é uma amostra. Saber se uma amostra é ou não "representativa" depende, acima de tudo, da medida em que ela é selecionada a partir da variedade dos tipos de textos de uma população-alvo; assim, uma avaliação dessa representatividade depende de uma definição completa prévia da "população" que a amostra pretende representar e das técnicas usadas para selecionar a amostra daquela população. A definição da população-alvo apresenta pelo menos dois aspectos: (1) a delimitação da população – quais são os textos incluídos e excluídos; (2) a organização hierárquica dentro da população – quais categorias de texto estão incluídas na população e como são definidas. No planejamento dos *corpora*, essas preocupações muitas vezes não recebem a devida atenção, e as amostras são coletadas sem uma definição prévia da população-alvo. Consequentemente, não há maneira alguma de avaliar a adequação ou a representatividade desse *corpus* (porque não há concepção alguma bem definida do que a amostra pretenda representar).

Além disso, a representatividade de um *corpus* depende do quanto ele abarca a variedade das distribuições linguísticas da população; isto é, as diferentes características linguísticas são distribuídas de modo diferente (nos textos, entre textos e entre tipos diferentes de textos), e um *corpus* representativo deve permitir a análise dessas diversas distribuições. Essa condição de representatividade linguística depende da primeira condição; ou seja, se um *corpus* não representa a gama de tipos de texto de uma população, ele não representará a variedade de distribuições linguísticas. Além disso, a representatividade linguística depende de questões como o número de palavras por amostra de texto, o número de amostras por "texto" e o número de textos por tipo de texto. Essas questões serão abordadas nas seções 3 e 4.

Entretanto, a questão da definição de população é a primeira consideração no planejamento de um *corpus*. Para exemplificar, levemos em conta as definições de população utilizadas no *corpus* Brown (Francis e Kucera, 1964/79) e no *corpus* LOB (Johansson *et al.*, 1978). Essas populações-alvo foram definidas com relação a seus limites (todos os textos publicados em inglês em 1961, nos Estados Unidos e no Reino Unido, respectivamente), e suas organizações hierárquicas (quinze categorias principais de textos e diversas distinções de subgêneros dentro dessas categorias). Na construção desses *corpora*, os compiladores também tiveram boas "bases de amostragem", o que permitia a amostragem probabilística e aleatória da população. Uma base de amostragem é uma definição operacional da população, uma listagem detalhada em itens dos membros da população, na qual uma amostra representativa pode ser selecionada. O manual do *corpus* LOB (Johansson *et al.*, 1978) é bastante explícito sobre a base de amostragem utilizada: para livros, a população-alvo foi operacionalizada como contendo todas as publicações de 1961 listadas no *The British National Bibliography Cumulated Subject Index*, 1960-1964 (que é baseado nas divisões por assunto do sistema de Classificação Decimal de Dewey), e para periódicos e jornais, a população-alvo foi operacionalizada como todas as publicações de 1961 listadas no *Willing's Press Guide* (1961). No caso do *corpus* Brown, a base de amostragem foi a coleção de livros e periódicos da *Brown University Library* e da *Providence Athenaeum*; essa base de amostragem, retirada de textos impressos de 1961 é menos representativa do que as listas utilizadas na construção do *corpus* Lancaster Oslo/Bergen (LOB), mas foram estabelecidos limites bem definidos e uma listagem detalhada dos membros. Ao escolher e avaliar uma base de amostragem, aspectos de eficiência e de relação custo/benefício devem ser comparados com maiores graus de representatividade.

Dada uma base de amostragem adequada, é possível selecionar uma amostra probabilística. Há vários tipos de amostras probabilísticas, mas todos eles dependem de uma seleção aleatória. Em uma amostragem aleatória simples, todos os textos da população têm uma mesma chance de serem selecionados. Por exemplo, se todos os itens da Bibliografia Nacional Britânica *(British National*

Bibliography) fossem numerados sequencialmente, seria possível usar uma tabela de números aleatórios para selecionar uma amostra aleatória de livros. Outro método de amostragem probabilística, que aparentemente foi usado na construção dos *corpora* Brown e LOB, é a "amostragem estratificada". Nesse método, são identificados os subgrupos dentro da população-alvo (neste caso, os gêneros), depois cada um desses "estratos" seria objeto de amostra utilizando técnicas de amostragem aleatória. Essa abordagem tem a vantagem de garantir que todos os estratos estejam representados adequadamente e, ao mesmo tempo, selecionar uma amostra não tendenciosa dentro de cada estrato (ou seja, no caso dos *corpora* Brown e LOB, obteve-se 100% de representação no nível das categorias de gênero e uma seleção imparcial de textos dentro de cada gênero).

Observe-se que, por duas razões, uma definição e uma análise cuidadosas das características não linguísticas da população-alvo são pré-requisitos fundamentais para as decisões de amostragem. Primeiro, não é possível identificar uma base de amostragem adequada ou avaliar em que medida uma determinada amostra representa uma população até que a própria população tenha sido cuidadosamente definida. Um bom exemplo é um *corpus* que se destina a representar os textos falados de uma língua. Como não existem catálogos ou bibliografias de textos falados, e uma vez que estamos constantemente expandindo o universo de textos falados em nossas conversas diárias, é difícil identificar, nesse caso, uma base de amostragem adequada. Todavia, sem uma definição prévia dos limites e parâmetros da produção oral em um idioma, a avaliação de uma determinada amostra não é possível.

A segunda motivação para uma definição prévia da população é que as amostras estratificadas são quase sempre mais representativas do que as amostras não estratificadas (e elas nunca são menos representativas). Isso ocorre porque os estratos identificados podem ser plenamente representados (amostragem de 100%) na proporção desejada, em vez de depender de técnicas de seleção aleatória. Em termos estatísticos, a variação entre grupos é geralmente maior do que a variância dentro do grupo, e, portanto, uma amostra que força a representação em todos os grupos de identificação será mais representativa no geral[3]. Voltando aos *corpora* Brown e LOB, uma identificação prévia das categorias de gênero (por exemplo, reportagem, discurso acadêmico, histórias de suspense) e subgêneros (por exemplo, Medicina, Matemática e Ciências Humanas dentro do gênero do discurso acadêmico) garantiu 100% de representação nesses dois níveis; ou seja, os planejadores do *corpus* tentaram compilar uma lista exaustiva das principais categorias de textos publicados em prosa em inglês, e todas essas categorias foram incluídas no planejamento do *corpus*. Desse modo, técnicas de

3 Além disso, no caso de *corpora* de língua, uma representação proporcional de textos geralmente não é desejável (ver Seção 3); em vez disso, a representação da gama de tipos de texto é necessária como base para análises linguísticas, fazendo que uma amostra estratificada seja ainda mais essencial.

amostragem aleatória foram necessárias apenas para obter uma seleção representativa de textos a partir de cada subgênero. A alternativa, uma seleção aleatória do universo de todos os textos publicados, dependeria de uma amostra grande e das probabilidades associadas à seleção aleatória para garantir a representação do intervalo de variação em todos os níveis (entre gêneros, subgêneros e textos dentro de subgêneros), uma tarefa mais complexa.

No presente trabalho, analiso em primeiro lugar as questões relativas às definições de população para os *corpora* de língua e tentarei desenvolver uma estrutura para a análise estratificada da população do *corpus* (Seção 2). Na Seção 3, retorno às questões de amostragem particular, incluindo amostragem proporcional *versus* não proporcional, a amostragem dentro de textos (o número de palavras por texto e amostragem estratificada dentro de textos) e questões relacionadas ao tamanho da amostra. Na Seção 4, descrevo as diferenças na distribuição das características linguísticas, apresentando as distribuições das diversas características particulares, e discuto as implicações dessas distribuições para planejamento de *corpus*. Por fim, na Seção 5, apresento um breve panorama de planejamento de *corpus* na prática.

2. Estratos de um *corpus* de texto: uma proposta operacional em relação aos parâmetros salientes de registro e variação

Conforme observamos na última seção, a definição da população do *corpus* exige a especificação das fronteiras e a especificação dos estratos. Se adotarmos o objetivo ambicioso de representar todo um idioma, os limites da população podem ser especificados como todos os textos do idioma. Especificar os estratos relevantes e identificar bases de amostragem são, obviamente, tarefas mais difíceis, que exigem uma especificação teoricamente motivada e completa dos tipos de textos. Nesta seção, apresento uma proposta preliminar para a identificação dos estratos para esse *corpus* e a operacionalização deles como bases de amostragem. A proposta se restringe às sociedades ocidentais (com exemplos dos Estados Unidos) e serve, sobretudo, como uma ilustração em vez de uma solução final, mostrando como um *corpus* desse tipo poderia ser planejado.

Uso os termos gênero ou registro para fazer referência às categorias de textos definidas de acordo com a situação (como ficção, transmissões desportivas, artigos de psicologia) e tipo de texto para se referir a categorias de texto definidas linguisticamente. Esses dois sistemas de classificação de texto são válidos, mas possuem bases diferentes. Embora os registros/gêneros não sejam definidos em termos linguísticos, há diferenças linguísticas estatisticamente importantes entre essas categorias (Biber, 1986, 1988), e as contagens das características linguísticas são relativamente estáveis entre textos de um mesmo registro (Biber, 1990). Em contrapartida, os tipos de texto são identificados com base nos padrões compar-

tilhados de coocorrência linguística, de modo que os textos dentro de cada tipo possuem características linguísticas o mais semelhantes possível, enquanto os tipos diferentes são o mais distinto possível uns dos outros (Biber, 1989).

Na definição de população para um *corpus*, as distinções de registro/gênero têm precedência sobre as distinções de tipo de texto. Isso ocorre porque os registros são baseados em critérios externos ao *corpus*, enquanto os tipos de texto são baseados em critérios internos; ou seja, os registros são baseados em diferentes situações, propósitos e funções do texto em uma comunidade discursiva e podem ser identificados antes da construção de um *corpus*. De outro modo, a identificação das distinções marcantes de tipo de texto em uma língua requer um *corpus* representativo de textos para análise; não há uma maneira *a priori* para identificar os tipos definidos linguisticamente. Mas, como apresento na Seção 4, os resultados de estudos anteriores, bem como a investigação em curso durante a construção de um *corpus*, podem ser usados para garantir que a seleção de textos seja representativa tanto linguística quanto situacionalmente.

A linguística de *corpus* concentrou-se, preferencialmente, nas diferenças de registro[4]. No planejamento de um *corpus*, no entanto, as decisões devem ser feitas ou para incluir uma gama representativa de dialetos ou se restringe o *corpus* a um único dialeto (por exemplo, uma variedade "padrão"). Os parâmetros dialetais especificam as características demográficas de quem fala e escreve, tais como a região geográfica, idade, sexo, classe social, etnia, educação e ocupação[5].

Diferentes planejamentos de *corpus* representam diferentes populações e satisfazem objetivos de investigação diferentes. Três dos possíveis planejamentos gerais são organizados em torno da produção do texto, a recepção do texto e textos como produtos. Os dois primeiros são demograficamente organizados no nível superior, ou seja, os indivíduos são selecionados a partir de uma população demográfica maior, e, em seguida, esses indivíduos são monitorados para registrar o seu uso da língua. Um planejamento de produção incluiria textos (falados e escritos) realmente produzidos pelos participantes da amostra; um planejamento para a recepção incluiria os textos ouvidos ou lidos. Essas duas abordagens tratariam do que as pessoas realmente fazem com a língua regularmente. A seleção demográfica pode ser estratificada segundo as subdivisões de ocupação, sexo, idade etc.

Um *corpus* orientado demograficamente não representaria toda a gama de tipos de texto em uma língua, uma vez que muitos tipos de linguagem são raramente usados, embora sejam importantes sob outras perspectivas. Por

4 Na verdade, muito pouco trabalho foi realizado sobre variação dialetal a partir de uma perspectiva baseada em texto. Ao contrário, estudos dialetais tendem a concentrar-se na variação fonológica, minimizando a importância das características discursivas e gramaticais.

5 Outros fatores demográficos caracterizam falantes e produtores de texto individuais, em lugar de grupos de usuários, que incluem características relativamente estáveis, como a personalidade, os interesses e as crenças, e características temporárias, como o humor e o estado emocional. Esses fatores provavelmente não são importantes para o planejamento do *corpus*, a não ser que a intenção do *corpus* seja a investigação das diferenças pessoais.

exemplo, poucas pessoas irão escrever uma lei ou um tratado, um contrato de seguro ou um livro de qualquer espécie, e alguns desses tipos de texto também são raramente lidos. Portanto, seria difícil estratificar um *corpus* demográfico de tal forma que assegurasse a representatividade da gama de categorias de texto. No entanto, várias dessas categorias são muito importantes na definição de uma cultura. Um *corpus* organizado em torno de textos como produtos se propõe a representar a gama de registros e tipos de texto em vez de padrões típicos de usos de vários grupos demográficos.

Há trabalhos sobre os parâmetros de variação de registro realizados por linguistas antropológicos, como Hymes e Duranti, e pelos linguistas funcionais, tais como Halliday (cf. Hymes 1974; Brown e Fraser, 1979; Duranti, 1985; Halliday e Hasan, 1989). Em Biber (1993a), tento desenvolver uma estrutura relativamente completa, argumentando que o "registro" deve ser especificado como uma noção contínua (ao invés de discreta) e distinguir entre a gama das variações situacionais que foram consideradas nos estudos de registro. Essa estrutura é por demais especificada para o trabalho de planejamento do *corpus* – os valores de alguns parâmetros são decorrentes de valores em outros parâmetros, e alguns parâmetros são específicos de determinados tipos de texto. A tentativa de amostragem nesse nível de especificidade, então, seria extremamente difícil. Por esse motivo, proponho, na Tabela 1, um conjunto reduzido de amostragem de estratos, equilibrando a viabilidade operacional com o intuito de definir a população-alvo o mais completamente possível.

Tabela 1: parâmetros situacionais listados como estratos hierárquicos de amostragem
1. *Canal primário*. Discurso escrito/falado/roteiro
2. *Formato*. Publicado/não publicado (+ vários formatos dentro de "publicado")
3. *Ambiente*. Institucional/outras entidades públicas/privadas ou pessoais
4. *Destinatário*
(a) Diversos. Não enumerados/vários/individual/pessoal
(b) Presença (lugar e tempo). Presente/ausente
(c) Interatividade. Nenhuma/pouca/ampla
(d) Conhecimento compartilhado. Geral/especializado/pessoal
5. *Remetente*
(a) *Variação demográfica*. Sexo, idade, ocupação etc.
(b) *Reconhecimento*. Indivíduo ou instituição reconhecidos
6. *Tipo de informação*. Factual-informacional/intermediária ou indeterminada/imaginária
7. *Objetivos*. Persuadir, entreter, ensinar, informar, instruir, explicar, narrar, descrever, registrar, autorrevelar-se, expressar atitudes, opiniões ou emoções, melhorar o relacionamento interpessoal...
8. *Tópicos*...

TABELA 1

O primeiro dos parâmetros acima divide o *corpus* em três componentes principais: a escrita, a fala e o roteiro de discurso. Cada um desses três exige

diferentes considerações de amostragem, e, portanto, nem todos os parâmetros situacionais subsequentes são relevantes para cada componente.

Em se tratando de escrita, a primeira distinção importante é a publicação[6]. Isso ocorre porque a população de textos publicados pode ser operacionalmente limitada, e vários catálogos e índices discriminados fornecem listas com um rol de membros. Por exemplo, é possível utilizar os seguintes critérios para a definição operacional de textos "publicados": (1) são impressos em várias cópias para distribuição; (2) têm registro de direitos autorais ou são registrados por um importante serviço de indexação. Nos Estados Unidos, uma lista de livros e periódicos com direitos autorais registrados está disponível na Biblioteca do Congresso (*Library of Congress*). Outros textos "publicados" que não possuem registro de direitos autorais são relatórios e documentos do governo, relatórios e documentos jurídicos, determinadas revistas e jornais e algumas dissertações; nos Estados Unidos, esses textos são indexados em fontes como o Catálogo Mensal de Publicações Governamentais dos EUA (*Monthly Catalog of US Government Publications*), o Índice de Periódicos do Governo dos EUA (*Index to US Government Periodicals*), todo um sistema de relatórios jurídicos (por exemplo, *o Pacific Reporter*, os *Supreme Court Reports*), índices de periódicos (por exemplo, *Reader's Guide to Periodical Literature, Newsbank*) e resumos de dissertações (indexados pela *University Microfilms International*).

Um terceiro estrato de textos escritos e publicados poderia ser esses "formatos" representados pelos diversos sistemas de indexação e catalogação. Juntos, esses índices fornecem uma lista discriminada dos textos escritos publicados e poderiam, portanto, ser utilizados como uma base de amostragem adequada. Com uma amostra suficientemente grande (ver seção seguinte), essa base de amostragem ajudaria a alcançar a "representatividade" dos vários tipos de escritos publicados. No entanto, sabemos por motivos teóricos que existem vários substratos importantes nos escritos publicados (por exemplo, finalidades e diferentes áreas temáticas), e é, portanto, melhor especificá-los ainda mais no planejamento do *corpus*. Essa abordagem é mais conservadora, na medida em que garante a representatividade nas proporções desejadas para cada uma dessas categorias de texto, e, ao mesmo tempo, permite amostras menores (uma vez que técnicas aleatórias exigem amostras maiores do que as técnicas estratificadas).

Ambiente e formato são estratos paralelos de segundo nível: o formato é importante para a amostragem da escrita publicada; o ambiente pode ser utilizado de maneira semelhante para fornecer amostras de escrita de textos não publicados, fala e roteiro de discurso. Diferenciam-se três tipos de ambiente aqui: públicos, institucionais e privados pessoais. Esses ambientes são menos

6 Este parâmetro não seria importante para muitas sociedades não ocidentais ou para certos tipos de corpora que representam diferentes períodos históricos; seriam necessárias, nesses casos, estratégias de amostragem muito diferentes.

adequados como base de amostragem do que catálogos de publicação – eles não fornecem limites bem definidos para textos escritos ou para fala não publicados nem fornecem uma lista exaustiva de textos dentro dessas categorias. O problema é que não existe uma base de amostragem direta de textos escritos ou falados não publicados. O ambiente, no entanto, pode ser usado indiretamente para retirar uma amostra dessas populações-alvo usando três subcategorias distintas: as instituições (escritórios, fábricas, empresas, escolas, igrejas, hospitais etc.), os ambientes privados (casas) e outros locais públicos (lojas, centros de recreação etc.). (Para roteiro de discurso, a categoria de "outros locais públicos" incluiria fala em vários meios de comunicação públicos, como noticiários, roteiro de discurso, diálogos em comédias e em séries na televisão). As bases de amostragem operacionais para cada um desses ambientes podem ser definidas a partir de vários registros governamentais e oficiais (por exemplo, registros censitários, declarações de impostos ou outros registros). O objetivo dessas bases de amostragem seria fornecer uma lista detalhada dos membros de cada tipo de ambiente, de modo que uma amostra aleatória de instituições, casas e outros locais públicos possam ser selecionados. (Esses três ambientes poderiam ser ainda mais estratificados em relação aos diversos tipos de instituições, os tipos de casa etc.). Para essa questão, propus as listas de amostragem indicadas na Tabela 2.

Antes de prosseguir, é necessário distinguir entre dois tipos de estratos de amostragem. O primeiro, conforme mencionado anteriormente, define, na verdade, uma base de amostragem, especificando os limites da população operacionalizada e fornecendo uma lista detalhada dos membros. O segundo, como nos demais parâmetros da Tabela 1, identifica as categorias que devem ser representadas em um *corpus*, mas não fornece bases de amostragem bem definidas. Por exemplo, o item Diversos – Destinatário (nº 4a: não enumerados/vários/individual/pessoal) não fornece as listagens dos textos com esses quatro tipos de destinatário; simplesmente especifica que os textos devem ser coletados até que essas quatro categorias estejam adequadamente representadas.

Além disso, os demais parâmetros da Tabela 1 não são igualmente relevantes para a grande maioria das listas de amostragem arroladas na Tabela 2. Consideremos, por exemplo, os parâmetros enumerados no item "Destinatário". Os textos escritos publicados têm sempre destinatários não enumerados[7], são sempre escritos para destinatários não presentes e são quase sempre não interativos (exceto as trocas de opiniões publicadas). Eles podem exigir tanto conhecimento de caráter geral quanto especializado (por exemplo, revistas populares *versus* revistas acadêmicas), mas raramente exigem conhecimentos prévios (embora sejam necessários para um pleno entendimento de memórias,

7 Coleções de cartas e diários publicados são casos especiais – originalmente eles têm destinatários individuais, mas eles geralmente são escritos com a esperança de eventual publicação e, portanto, com um público não enumerado em mente.

cartas publicadas, diários e mesmo alguns romances e contos). Textos escritos não publicados, por outro lado, podem se encaixar em todas essas categorias de destinatário. Os destinatários podem ser não enumerados (por exemplo, anúncios, catálogos de propagandas, formulários ou anúncios governamentais), diversos (circular ou memorando, relatórios técnicos ou comerciais), individuais (memorando para um indivíduo, cartas profissionais ou pessoais, mensagens eletrônicas) ou pessoal (diário, anotações, lista de compras). O destinatário de textos não publicados geralmente é ausente, exceto quando escreve para si mesmo. Os textos não publicados podem ser interativos (por exemplo, cartas) ou não. Finalmente, textos não publicados só podem exigir conhecimentos de caráter geral (por exemplo, algumas propagandas), conhecimentos especializados (por exemplo, relatórios técnicos) ou conhecimentos pessoais (por exemplo, cartas e diários).

A fala é normalmente dirigida a um destinatário individual ou plural que está presente. A fala dirigida a si mesmo é, em termos gerais, considerada estranha. A fala pode ser dirigida a destinatários não enumerados e ausentes através dos meios de comunicação em massa (por exemplo, uma entrevista televisiva). Os destinatários individuais ou de pequenos grupos também podem estar ausentes, como no caso de conversas telefônicas e "teleconferências". (Destinatários individuais podem até mesmo ser não interativos, no caso de falar com uma secretária eletrônica). Ambientes privados favorecem destinatários interativos (seja conversa individual ou com pequenos grupos), enquanto destinatários interativos e não interativos podem ser encontrados em ambientes institucionais (por exemplo, pensemos nos vários tipos de palestras, sermões e apresentações comerciais). Conhecimentos gerais podem ser necessários em todos os tipos de conversas; conhecimentos especializados prévios são mais necessários para destinatários em ambientes institucionais; conhecimento pessoal é mais necessário em ambientes privados.

Textos escritos (publicados). Livros/periódicos/etc. (com base em índices disponíveis)
Textos escritos (não publicados). Institucionais/públicos/privados
Fala. Institucional/pública/privada
Roteiro de discurso. Institucional/meios de comunicação públicos/outros

TABELA 2

O roteiro de discurso normalmente é dirigido para diversos destinatários (pequenos grupos em ambientes institucionais e públicos não enumerados nos meios de comunicação de massa). O diálogo nas peças de teatro e nas séries na televisão são exemplos de roteiro de discurso, que é direcionado para um indivíduo, mas ouvido por uma plateia não enumerada. Os destinatários estão tipicamente presentes no roteiro de discurso em ambientes institucionais, mas

não estão presentes (física ou temporalmente) nos roteiros projetados através dos meios de comunicação em massa. Aberta exceção para o palestrante, que permite perguntas durante um discurso escrito, o roteiro de discurso geralmente não é interativo. Por fim, o roteiro de discurso pode exigir tanto conhecimento de caráter geral quanto especializado por parte do destinatário, mas raramente requer conhecimentos pessoais prévios.

Os remetentes podem variar ao longo de um número de parâmetros demográficos (características do dialeto mencionadas acima), e as decisões devem ser tomadas com relação à representação desses parâmetros no *corpus*. (Coleções de textos de algumas categorias de remetentes serão difíceis para algumas bases de amostragem; por exemplo, há relativamente poucos textos escritos publicados por escritores da classe operária.) O segundo parâmetro, sendo o remetente reconhecido ou não, é relevante apenas para os textos escritos: alguns textos escritos não têm um autor específico mencionado (por exemplo, anúncios, catálogos, leis e tratados, formulários do governo, contratos comerciais), ao passo que categorias típicas de escrita possuem autor(es) específico(s).

O tipo de informação é semelhante às avaliações de conhecimento prévio, porque às vezes é difícil de medir de modo confiável. No entanto, este é um parâmetro importante a se distinguir entre os textos dentro da escrita e da fala. Em um polo, estão os relatórios e palestras científicos, que almejam ser factuais; no outro polo, estão os vários tipos de histórias imaginativas. Entre esses polos, há um *continuum* de textos que se apoiam de maneira diferente nos fatos, passando pela especulação, opinião, ficção histórica, fofoca etc.

O parâmetro do propósito requer mais pesquisas, tanto teóricas (como a base para o planejamento do *corpus*) quanto empíricas (utilizando os recursos de *corpora* grandes).

Incluo na Tabela 1 diversos propósitos que devem ser representados em um *corpus*, mas essa listagem não pretende ser exaustiva.

De modo semelhante, o parâmetro do tópico exige mais investigação teórica e empírica. Os sistemas de classificação de bibliotecas são bem desenvolvidos e fornecem estratos dos tópicos adequados para textos escritos publicados. Essas mesmas classificações também poderiam servir como estratos de textos escritos não publicados, mas eles precisariam ser testados empiricamente. Para os textos falados, principalmente em ambientes privados, é necessária uma investigação mais aprofundada sobre a gama de temas típicos.

O espírito da proposta esboçada nesta seção é mostrar como os parâmetros situacionais básicos podem ser usados como estratos de amostragem para oferecer um primeiro passo importante para alcançar representatividade. Os valores dos parâmetros particulares usados, no entanto, devem ser refinados, e a estrutura proposta aqui não é, obviamente, a palavra final em estratos de amostragem de *corpus*.

3. Outras questões de amostragem

3.1. Amostragem proporcional

Na maioria dos planejamentos de amostragem estratificada, a seleção das observações entre estratos deve ser proporcional a fim de ser considerada representativa (Williams, 1978; Henry, 1990), ou seja, o número de observações em cada estrato deve ser proporcional aos seus números na população maior. Por exemplo, um levantamento de cidadãos na Carolina do Norte (relatado em Henry, 1990, pp. 61-66) utilizou dois estratos, cada um baseado em uma lista de adultos realizada pelo governo: famílias que apresentaram declarações de imposto de renda em 1975 e famílias que eram selecionadas para a assistência *Medicaid*. Essas duas listas representam cerca de 96% da população. Na seleção das observações, no entanto, elas foram amostradas de modo proporcional – 89% da lista do imposto de renda e 11% da lista da *Medicaid* – para manter as proporções relativas desses dois estratos na população maior. Pode-se afirmar, portanto, que a amostra obtida representa a população adulta da Carolina do Norte. A representatividade, neste caso, significa fornecer a base para uma cuidadosa estatística descritiva de toda a população (por exemplo, a média salarial, educação etc.).

Amostras demográficas são representativas na medida em que refletem as proporções relativas dos estratos em uma população. Essa noção de representatividade foi desenvolvida no âmbito da investigação sociológica em que os pesquisadores têm por objetivo determinar as estatísticas descritivas que caracterizam a população em geral (como a média e desvio padrão da população). Toda e qualquer estatística que caracterize toda uma população depende fundamentalmente de uma amostra proporcional dos estratos na população – se for feita uma amostra expressiva de um estrato que compõe uma pequena proporção da população, ele terá pouco peso representativo no resumo da estatística descritiva.

Os *corpora* de idiomas exigem uma noção diferente de representatividade, fazendo com que a amostragem proporcional seja inadequada. Um *corpus* de idioma proporcional teria de ser demograficamente organizado (conforme será discutido no início da Seção 3.2), porque não temos um caminho *a priori* para determinar as proporções relativas dos diferentes registros em uma língua. Na verdade, uma simples amostra do uso do idioma baseada na demografia seria proporcional por definição – o *corpus* resultante contemplaria os registros que as pessoas costumam usar nas proporções reais em que eles são usados. É possível que um *corpus* com esse planejamento apresente cerca de 90% de conversa, 3% de cartas e bilhetes, e os 7% restantes estariam divididos entre os registros como reportagem de imprensa, revistas populares, jargão acadêmico, ficção, palestras, noticiários e textos não publicados. (Pouquíssimas pessoas produzem

textos escritos publicados ou não publicados ou textos orais para um grande público.) Tal *corpus* permitiria resumir estatísticas descritivas para todo o idioma representado pelo *corpus*. Esses tipos de generalizações, no entanto, não são tipicamente de interesse da pesquisa linguística. Em vez disso, os pesquisadores exigem amostras do idioma que são representativas no sentido de que incluem toda gama de variações linguísticas existentes em uma língua.

Em resumo, há dois problemas principais com *corpora* proporcionais de língua. Primeiro, as amostras proporcionais são representativas apenas na medida em que refletem exatamente as frequências numéricas relativas de registros em uma língua – eles não fornecem qualquer representação de importância relativa que não seja numérica. Os registros, tais como livros, jornais e noticiários são muito mais influentes do que as frequências relativas indicam. Em segundo lugar, os *corpora* proporcionais não fornecem uma base adequada para análises linguísticas, em que a gama de características linguísticas encontradas em diferentes tipos de texto é de interesse primário. Por exemplo, não é necessário ter um *corpus* para descobrir que 90% dos textos em uma língua/linguagem são linguisticamente semelhantes (porque todos são conversações); em vez disso, queremos analisar as características linguísticas dos outros 10% dos textos, uma vez que representam a grande maioria dos tipos de registros e distribuições linguísticas de uma língua[8].

3.2. Tamanho da amostra

Há muitas equações para determinar o tamanho da amostra baseadas nas propriedades da distribuição normal e da distribuição por amostragem da média (ou a distribuição de amostragem do desvio padrão). Uma das equações mais importantes afirma que o erro padrão da média para algumas variáveis ($S\bar{x}$) é igual ao desvio padrão da variável (S) dividido pela raiz quadrada do tamanho da amostra ($n^{1/2}$), por exemplo:

$$S\bar{x} = s/n^{½}$$

O erro padrão da média indica o quão longe pode estar a média da amostra da verdadeira média da população. Se o tamanho da amostra é maior que 30, a distribuição das médias da amostra tem distribuição aproximadamente normal, tal que 95% das amostras tiradas de uma população terão médias que cairão no intervalo de 1,96 vez, a mais ou a menos, o erro padrão. Quanto menor for o

8 Um *corpus* proporcional seria útil para as avaliações de que uma palavra ou uma construção sintática é "comum" ou "rara" (como em aplicações lexicográficas). Infelizmente, a maioria das palavras raras não apareceria em um *corpus* proporcional (isto é, principalmente em conversa), tornando o banco de dados inadequado para a pesquisa lexicográfica.

intervalo, maior a confiança que um pesquisador terá de que ele está representando a média da população com precisão. Como mostra a equação para o erro padrão, esse intervalo de confiança depende da variação natural da população (estimados pelo desvio padrão da amostra) e do tamanho da amostra (*n*). A influência do tamanho da amostra nessa equação é constante, independente do valor do desvio padrão (ou seja, o erro padrão é uma função de um dividido pela raiz quadrada de *n*). Para reduzir o erro padrão (e, portanto, reduzir o intervalo de confiança) pela metade, é necessário aumentar o tamanho da amostra em quatro vezes.

Por exemplo, se o desvio padrão da amostra para o número de substantivos em um texto foi de 30, a pontuação média da amostra foi de 100, e o tamanho da amostra é de nove textos, logo o erro padrão será igual a 10:

$$\text{Erro padrão} = 30/\sqrt{(9)} = 30/3 = 10$$

Esse valor indica que há uma probabilidade de 95% de que a verdadeira média da população para o número de substantivos por texto esteja dentro do intervalo de 80,4 a 119,6 (ou seja, a média da amostra de 100 ± 1,96 vez o erro padrão de 10). Para reduzir esse intervalo de confiança por meio da divisão do erro padrão pela metade, o tamanho da amostra deve ser aumentado quatro vezes, resultando em 36 textos, ou seja:

$$\text{Erro padrão} = 30/\sqrt{(36)} = 30/6 = 5$$

Do mesmo modo, se a amostra inicial é de 25 textos, é necessário aumentar a amostra para 100 textos, a fim de reduzir o erro padrão pela metade, ou seja:

$$\text{Erro padrão} = 30/\sqrt{(25)} = 30/5 = 6$$

$$\text{Erro padrão} = 30/\sqrt{(100)} = 30/10 = 3$$

Infelizmente, existem algumas dificuldades em usar a equação do erro padrão para determinar o tamanho da amostra necessária de um *corpus*. Em especial, é necessário abordar três problemas:

(1) O tamanho da amostra (*n*) depende de uma determinação prévia do intervalo de confiança tolerável exigido para o *corpus*; ou seja, é necessário que haja uma estimativa *a priori* do grau de incerteza que pode ser tolerado em análises típicas baseadas no *corpus*.

(2) A equação depende do desvio padrão da amostra, mas este é o desvio padrão para uma variável específica. Diferentes variáveis podem ter diferentes desvios padrão, resultando em diferentes estimativas do tamanho da amostra necessária.

(3) A equação deve ser usada de maneira circular; ou seja, é necessário ter selecionado uma amostra e ter calculado o desvio padrão da amostra antes que a equação possa ser usada (e isso é baseado no pressuposto de que a amostra-piloto possui pelo menos certa representatividade) – mas o propósito da equação é determinar o tamanho necessário da amostra.

Na Seção 4, avalio a distribuição de várias características linguísticas e abordo esses três problemas, elaborando propostas preliminares a respeito do tamanho da amostra.

3.3. Uma observação sobre amostragem em "textos"

Até o momento, ainda não abordei a questão do tamanho que as amostras de texto devem ter. Irei abordar essa questão mais detalhadamente na Seção 4, discutindo a distribuição de várias características linguísticas em textos. Aqui, porém, quero salientar que a preferência por amostragem estratificada aplica-se a amostragem em textos, bem como entre textos. Os compiladores de *corpus* normalmente têm tentado atingir uma melhor representação dos textos simplesmente coletando mais palavras dos textos. No entanto, essas palavras certamente não são selecionadas aleatoriamente (ou seja, são sequenciais), e a adequação da representação, portanto, depende do tamanho da amostra em relação ao tamanho total do texto. Em vez disso, é possível usar uma abordagem estratificada para a seleção das amostras de textos a partir de textos; isto é, especialmente no caso de textos escritos e textos falados planejados, a seleção de amostras de textos pode usar os subcomponentes típicos de textos nesse registro como estratos de amostragem (por exemplo, capítulos, seções, possivelmente pontos principais em palestra ou sermão). Essa abordagem resultará em uma melhor representação do texto como um todo, independentemente do número total de palavras selecionadas de cada texto.

4. As distribuições das características linguísticas: recomendações preliminares relativas ao tamanho da amostra

4.1 As distribuições nos textos: tamanho das amostras de texto

Nesta seção, considero primeiramente a distribuição das características linguísticas nos textos como base para abordar a questão do tamanho ideal de texto. A teoria da amostragem tradicional é menos útil aqui do que para os outros aspectos do planejamento do *corpus*, porque as palavras individuais não podem ser tratadas como observações à parte nas análises linguísticas;

ou seja, uma vez que as características linguísticas normalmente englobam mais de uma palavra, qualquer seleção aleatória de palavras de um texto não conseguiria representar diversas características e arruinaria a estrutura geral do texto. A questão principal aqui é, desse modo, o número de palavras contíguas exigido em amostras de texto. A presente seção apresenta como esse problema pode ser resolvido através de investigações empíricas da distribuição das características linguísticas em textos.

Em Biber (1990), abordo esse problema através da comparação de pares de amostras de 1.000 palavras retiradas de textos isolados dos *corpora* LOB e London-Lund. (As amostras de texto são de 2.000 palavras do *corpus* LOB e de 5.000 palavras do *corpus* London-Lund.) Se encontramos diferenças grandes entre duas amostras de 1.000 palavras, então podemos concluir que esse tamanho de amostra não representa adequadamente as características linguísticas gerais de um texto e que talvez sejam necessárias amostras muito maiores. Se, por outro lado, as duas amostras de texto de 1.000 palavras são semelhantes linguisticamente, então podemos concluir que as amostras de textos relativamente pequenas representam adequadamente suas características linguísticas.

No caso de textos escritos (a partir do *corpus* LOB), dividi cada texto original pela metade e comparei as duas partes. No caso de textos falados (a partir do *corpus* London-Lund), quatro amostras de 1.000 palavras foram extraídas de cada texto original e então comparadas em pares.

Para fornecer uma base de dados relativamente ampla, foram analisadas 10 características linguísticas comumente utilizadas em estudos de variação. Essas características foram escolhidas a partir de diferentes classes funcionais e gramaticais, já que cada classe representa potencialmente uma distribuição estatística diferente entre as categorias de texto (ver Biber, 1988). As características são: pronomes em primeira pessoa, pronomes de terceira pessoa, contrações, verbos no passado, verbos no presente, preposições, construções de voz passiva (combinando passivas com e sem agente), orações relativas QU- e orações subordinadas condicionais. Os pronomes e as contrações são relativamente interativos e coloquiais na função comunicativa; os substantivos e as preposições são usados para integrar informações em textos; orações relativas e subordinadas condicionais representam tipos de elaboração estrutural; e a voz passiva é característica de estilos científicos ou técnicos. Essas características também foram escolhidas para representar uma grande variedade de distribuições de frequência em textos, como mostra a Tabela 3, que apresenta suas frequências (a cada 1.000 palavras) em um *corpus* de 481 textos escritos e falados (tirado de Biber, 1988, p. 77-78). As 10 características diferem consideravelmente em sua frequência média geral e em seus intervalos de variação. Os substantivos e as preposições são extremamente comuns; os marcadores do tempo presente são bastante comuns; pretérito, pronomes em primeira pessoa e pronomes de terceira pessoa

são todos relativamente comuns; as contrações e voz passiva são relativamente raras; e orações relativas QU- e subordinadas condicionais são bastante raras. (Além disso, essas características são distribuídas de maneira diferenciada entre os diferentes tipos de textos; ver Biber, 1988, pp. 246-269.) A comparação dessas 10 características entre os pares de textos de 1.000 palavras, portanto, representa vários dos tipos de padrões de distribuição encontrados em inglês.

Característica linguística	*Média*	*Mín.*	*Máx.*	*Faixa*
Substantivos	181	84	298	214
Preposições	111	50	209	159
Verbos no presente	78	12	182	170
Verbos no pretérito	40	0	119	119
Pronomes de 3ª pessoa	30	0	124	124
Pronomes de 1ª pessoa	27	0	122	122
Contrações	14	0	89	89
Voz passiva	10	0	44	44
Orações relativas QU-	3,5	0	26	26
Subordinação condicional	2,5	0	13	13

TABELA 3

As distribuições dessas características linguísticas foram analisadas em 110 amostras de texto de 1.000 palavras (ou seja, cinquenta e cinco pares de amostras), extraídos de sete categorias de texto: conversas, programas de rádio/TV, discursos, documentos oficiais, discurso acadêmico, ficção em geral e ficção de romance. Essas categorias representam uma gama de situações comunicativas em inglês, diferenciando-se no propósito, tópico, foco informacional, modo, interatividade, formalidade e circunstâncias de produção; em suma, o objetivo era representar uma ampla gama das distribuições de frequência.

Os coeficientes de confiabilidade foram calculados para avaliar a estabilidade da contagem da frequência de todas as amostras de 1.000 palavras. No caso do *corpus* London-Lund (os textos falados), quatro amostras de 1.000 palavras de cada texto foram analisadas, e para o *corpus* LOB (os textos escritos), duas subamostras de 1.000 palavras de cada texto foram analisadas.

O coeficiente de confiabilidade para cada característica representa a correlação média entre as contagens da frequência dessa característica (ou seja, uma contagem para cada uma das subamostras). Para as amostras de fala, todos os coeficientes foram altos. As menores confiabilidades foram para a voz passiva (0,74) e para subordinação condicional (0,79), ao passo que todas as outras tiveram coeficientes de confiabilidade acima de 0,88. Os coeficientes foram um pouco menores para as amostras de texto escrito, em parte porque são baseados em duas, em vez de quatro, subamostras. A subordinação condicional nos textos escritos tinha um baixo coeficiente de confiabilidade (0,31), enquanto as orações relativas e os verbos no presente nos textos escritos apresentaram coeficientes de confiabilidade relativamente baixos (0,58 e 0,61, respectivamente); todas as

outras características tiveram coeficientes de confiabilidade acima de 0,80. Acima de tudo, essa análise indica que as contagens da frequência para as características linguísticas comuns são relativamente estáveis em amostra de 1.000 palavras, enquanto a contagem da frequência para características raras (como subordinação condicional e orações relativas QU-; ver Tabela 3) são menos estáveis e exigem amostras de textos maiores para serem representadas de maneira confiável[9].

Essas análises anteriores podem ser complementadas pelo rastreamento da distribuição de várias características linguísticas em segmentos de textos de 200 palavras. Por exemplo, a Figura 1 mostra a distribuição das locuções prepositivas ao longo de cinco textos acadêmicos das Humanas – a figura dispõe o número acumulado de locuções prepositivas, em cada intervalo de 200 palavras nesses textos. Como pode ser visto na figura, locuções prepositivas são distribuídas de maneira linear nesses textos, isto é, há aproximadamente o mesmo número de locuções prepositivas que ocorrem em cada segmento de 200 palavras (cerca de 30 por segmento em três dos textos, e 25 por segmento nos outros dois textos).

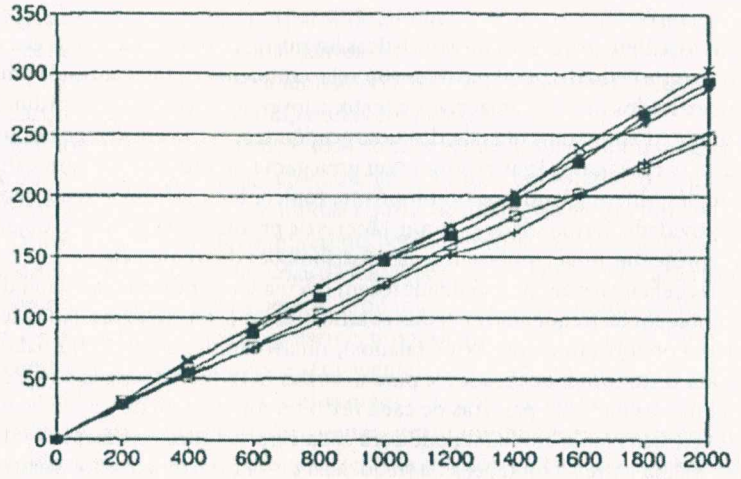


Figura 1: Preposições X Número de palavras – Distribuição das preposições em cinco textos das Humanas

(A natureza linear dessas distribuições pode ser confirmada ao colocar uma régua próximo à representação de cada texto.) A Figura 1 indica que uma característica comum, como locuções prepositivas, é extremamente estável

9 Essas são, principalmente, locuções preposicionais funcionando como modificadores de substantivos, ao contrário de locuções prepositivas funcionando como advérbios.

em sua distribuição dentro de textos (pelo menos em textos acadêmicos das Humanas) – que, mesmo em segmentos de 200 palavras, todos os segmentos apresentarão aproximadamente o mesmo número de locuções prepositivas.

A Figura 2 ilustra uma distribuição curvilínea, neste caso os tipos de palavras cumulativas (ou seja, o número de palavras diferentes) em cinco textos das Humanas. Em geral, as contagens da frequência de uma característica linguística serão distribuídas linearmente (embora essa distribuição seja mais ou menos estável em um texto, veja adiante), enquanto as frequências de diferentes tipos de características linguísticas (lexicais ou gramaticais) serão distribuídas curvilineamente, ou seja, como muitos tipos são repetidos em todos os segmentos de texto, cada segmento subsequente contribui com menos tipos novos que o segmento anterior. Na Figura 2, a linha reta marcada por triângulos mostra o limite de 50% dos tipos de palavras (o valor quando 50% das palavras em um texto são tipos de palavras diferentes). Nos cinco textos, pelo menos 50% das palavras são tipos diferentes no primeiro segmento de 200 palavras (ou seja, pelo menos a metade das palavras não é repetida), e dois dos textos apresentam mais de 50% de tipos diferentes nos três primeiros segmentos (até 600 palavras). No entanto, todos os textos mostram um declínio gradual no número de tipos de palavras. O texto mais diversificado cai para cerca de 780 tipos de palavras, 2.000 palavras (39%), enquanto o texto menos diversificado cai para cerca de 480 tipos de palavras, 2.000 palavras (apenas 24%). Essas tendências continuarão em textos maiores, com cada segmento subsequente contribuindo com menos tipos novos.

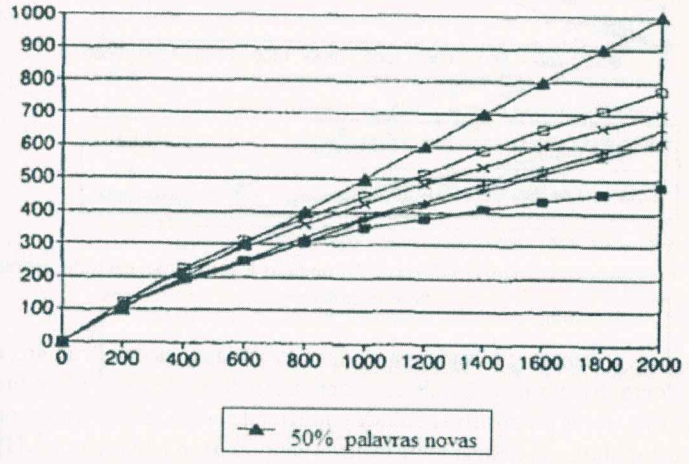


Figura 2: Tipos de palavras X Números de palavras - Distribuição de tipos de palavras em cinco textos das Humanas

Esses dois tipos de distribuições devem ser tratados de modo diferente. Nas Figuras 3-9, organizo as distribuições das sete características linguísticas nos textos que representam três registros. Três das características são contagens de frequência cumulativa: a Figura 3 apresenta as frequências das locuções prepositivas, uma característica gramatical comum; a Figura 4 apresenta as frequências das orações relativas, uma característica gramatical relativamente rara; e a Figura 5 apresenta as frequências da sequência substantivo preposicionado, uma sequência gramatical relativamente comum. As outras quatro figuras apresentam as distribuições de tipos em textos. As Figuras 6 e 7 apresentam a distribuição dos tipos lexicais: tipos de palavra (o número de palavras diferentes) na Figura 6 e *hapax legomena* (palavras que ocorrem uma vez) na Figura 7. As Figuras 8 e 9 apresentam a distribuição dos tipos gramaticais: diferentes categorias gramaticais ou "etiquetas" na Figura 8, diferentes sequências de etiquetas gramaticais na Figura 9. As figuras, então, ilustram as características lexicais e gramaticais, com frequências gerais raras e comuns, apresentando distribuições lineares e curvilíneas.

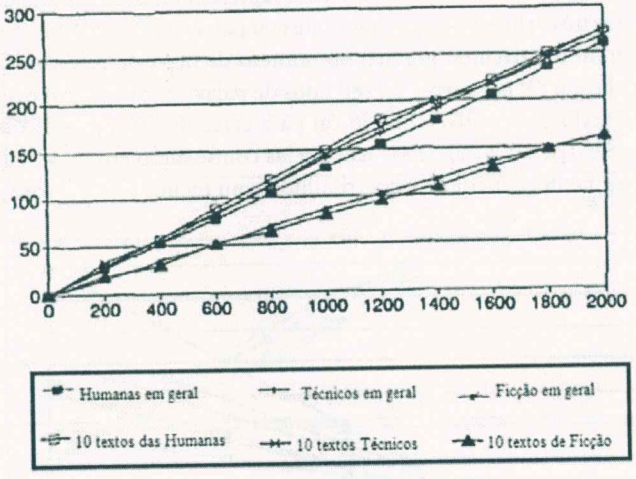


Figura 3: Preposições X Número de palavras – Distribuição de preposições em textos retirados de três registros

As figuras podem ser utilizadas para lidar com várias questões. Primeiro, elas apresentam as distribuições gerais dessas características. A distribuição linear estável das locuções prepositivas é ainda confirmada pela Figura 3. Em contrapartida, a distribuição relativamente instável das orações relativas, indicada por um coeficiente de confiabilidade relativamente baixo, é ainda confirmada pelas rupturas frequentes da linearidade na Figura 4. Isto é, uma vez que as orações relativas são relativamente raras em geral, até mesmo duas ou três orações relativas em um segmento de 200 palavras resultariam em uma irregularidade. A Figura 5 mostra que a distribuição das sequências substantivo preposicionado é semelhante à das locuções prepositivas pelo fato de serem lineares e bastante estáveis (embora menos frequentes no geral).[10]

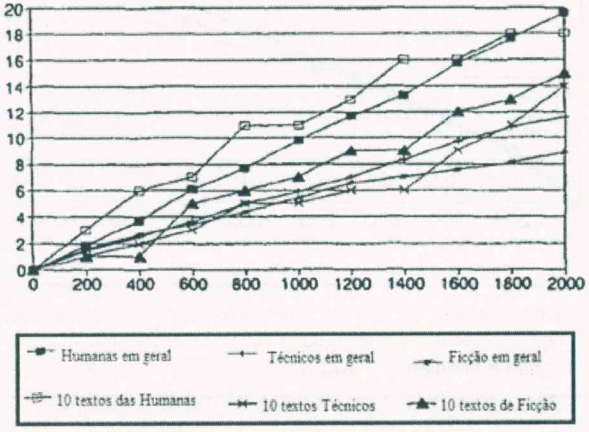


Figura 4: Orações relativas X Número de palavras – Distribuição das orações relativas retiradas de textos de três registros

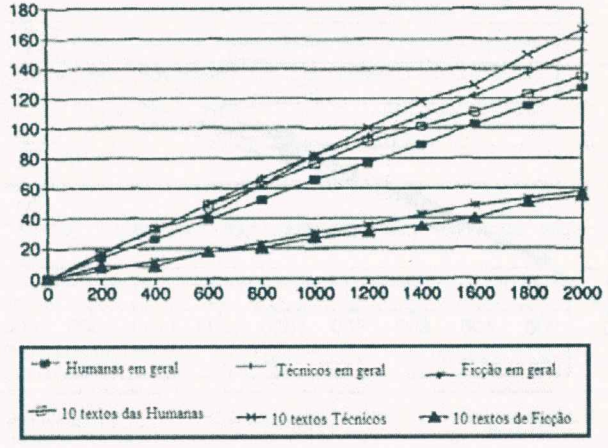


Figura 5: Substantivos preposicionados X Número de palavras - Distribuição de substantivos preposicionados retirados de três registros de texto

10 Na verdade, essa última questão foi abordada ao calcular a diferença dos valores, para a média, para o desvio padrão e para a abrangência entre as amostras dos 10 textos.

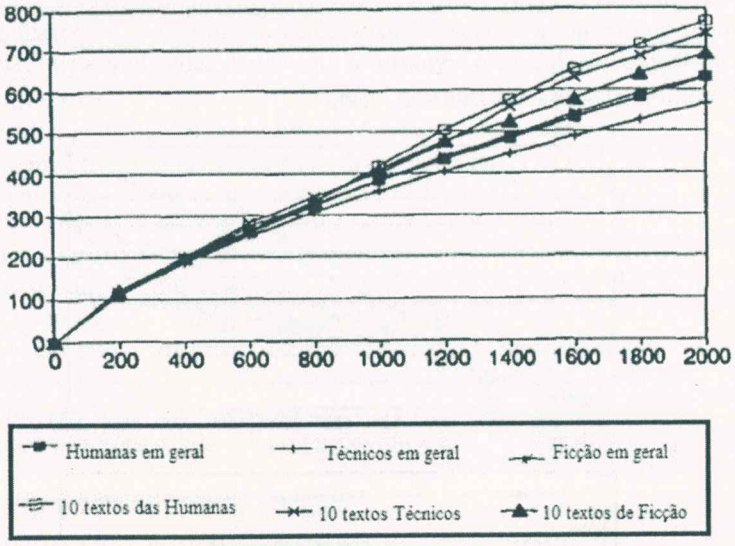


Figura 6: Tipos de palavras X Número de palavras – Distribuição de tipos de palavras em textos retirados de três registros

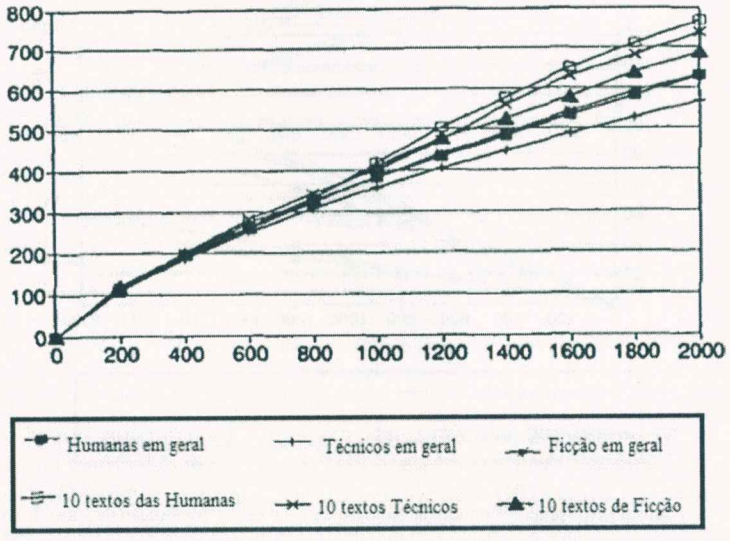


Figura 7: *Hapax legomena* X Número de palavras – Distribuição de *Hapax legomena* em textos retirados de três registros

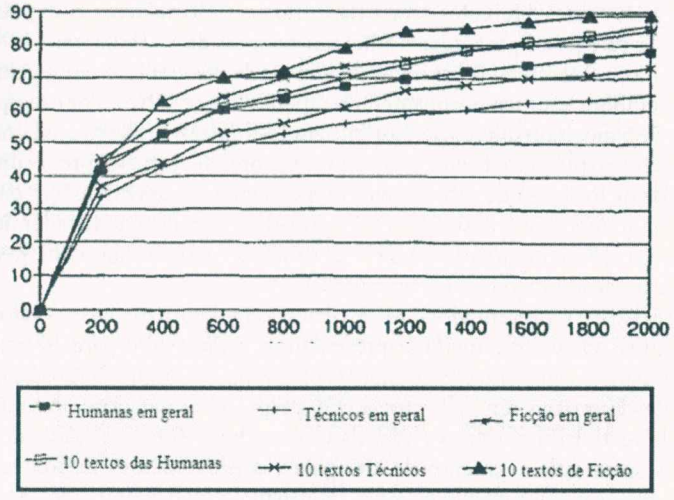


Figura 8: Tipos de etiquetas gramaticais X Número de palavras – Distribuição de etiquetas gramaticais retiradas de três registros de texto

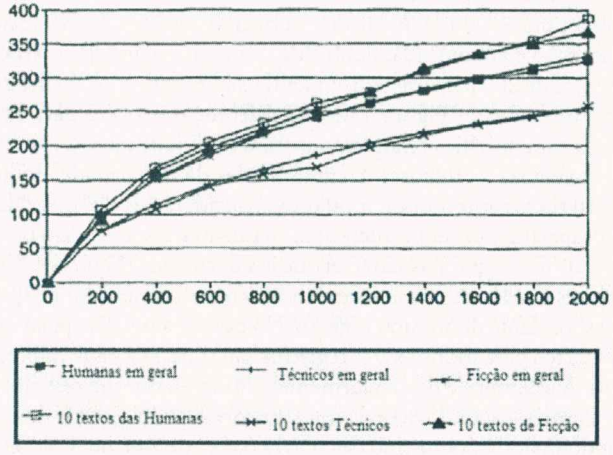


Figura 9: Sequências de etiquetas gramaticais X Número de palavras – Distribuição de sequências de etiquetas gramaticais em textos retirados de cinco registros

As Figuras 6-9 mostram diferentes graus de curvatura, com os tipos gramaticais e sintáticos mostrando quedas mais acentuadas do que os tipos lexicais. Os tipos de etiquetas gramaticais registram o maior declínio: a maioria das

categorias gramaticais diferentes ocorre nas primeiras 200 palavras, com pouco acréscimo de categorias gramaticais adicionais depois de 600 palavras.

As Figuras 3-9 também ilustram as diferenças de distribuição em registros diferentes, embora aqui sejam considerados apenas três registros. Por exemplo, as Figuras 3 e 5 mostram diferenças bastante grandes entre o discurso acadêmico e a ficção, tendo o primeiro frequências muito maiores de locuções prepositivas e sequências de locuções de substantivos preposicionados. As diferenças entre os registros são menos claras na Figura 4, mas o discurso de textos acadêmicos das Humanas apresenta uma frequência de orações relativas mais consistente do que o texto acadêmico técnico ou a ficção.

Cada registro é representado duas vezes nessas figuras: a média e os valores dos "10 textos". Os valores médios apresentam o valor médio para 10 textos do registro para o segmento em questão. (Por exemplo, a Figura 3 mostra que os textos das Humanas têm, em média, 130 preposições nas primeiras 1.000 palavras de texto.) Em contrapartida, os valores dos "10 textos" são valores compostos, com cada segmento de 200 palavras provenientes de um texto diferente. Assim, o valor de 400 palavras representa o total acumulado para as primeiras 200 palavras de dois textos, o valor de 600 palavras é a somatória das primeiras 200 palavras totais de três textos etc.

No caso das distribuições estáveis e lineares, há muito pouca diferença entre os valores da média e os valores dos "10 textos". Na verdade, as Figuras 3 e 5 mostram uma notável coincidência dos valores da média e dos 10 textos; uma única distribuição é encontrada, dentro de um registro, independente do fato de os segmentos de 200 palavras subsequentes serem retirados do mesmo texto ou de textos diferentes. A Figura 4 mostra diferenças maiores para as orações relativas (uma característica relativamente rara e menos estável). Aqui, a média dos 10 textos suaviza as maiores irregularidades da linearidade, enquanto os valores dos 10 textos apresentam mudanças consideráveis da linearidade.

Em contrapartida, existem diferenças marcantes entre as distribuições da média e dos 10 textos para as características curvilíneas (Figuras 6-9). Nesses casos, os valores dos 10 textos são sistematicamente mais elevados do que o valor médio correspondente do mesmo registro. No caso dos tipos de palavras, os valores dos 10 textos para todos os três registros são superiores aos valores médios dos registros. A diferença é particularmente notória no que diz respeito ao texto acadêmico técnico – após 2.000 palavras de texto, o texto técnico dos 10 textos tem o segundo maior valor de tipo de palavra (cerca de 740, ou 37%), enquanto, em média, textos do tipo técnico têm o menor valor em tipo de palavra (cerca de 570, ou 28%). Isso mostra que existe um elevado grau de repetição lexical em textos técnicos, mas há um alto grau de diversidade lexical quando se tomam textos técnicos diferentes. A distribuição de *hapax legomena*, conforme mostra a Figura 7, apresenta o mesmo traçado da distribuição dos tipos de palavras – da mesma forma, todos os três valores dos 10 textos são superiores aos escores médios; a amostra de 10 textos das Ciências Humanas apresenta o valor maior, e o

valor médio dos textos técnicos é notoriamente o mais baixo. Essas distribuições refletem a grande diversidade lexical encontrada entre os textos das Ciências Humanas e relativamente baixa dentro de cada texto técnico.

Há mais semelhanças entre os valores dos 10 textos e as médias no que diz respeito à distribuição dos tipos gramaticais (Figuras 8 e 9), embora para cada registro o escore dos 10 textos seja maior que o valor médio correspondente. Curiosamente, esses números mostram que o discurso técnico possui a menor diversidade gramatical, bem como a menor diversidade lexical.
Em resumo, as análises apresentadas nesta seção indicam o seguinte:

(1) As características linguísticas lineares comuns são distribuídas de maneira bastante estável dentro dos textos e, portanto, podem ser confiavelmente representadas por segmentos de texto relativamente curtos.
(2) As características linguísticas raras mostram uma variação de distribuição muito maior nos textos e, portanto, exigem amostras de texto mais longas para a representação confiável.
(3) As características distribuídas em uma forma curvilínea, isto é, os tipos de características diferentes, são relativamente estáveis entre os segmentos de texto subsequentes, mas as ocorrências de novos tipos diminuem ao longo de um texto. A frequência de novos tipos é consistentemente superior entre amostras de textos diferentes do que em amostras de texto único. Foi demonstrado que esse padrão se mantém para os segmentos de textos relativamente curtos (total de 2.000 palavras) e de amostras entre textos retirados de um registro único; os padrões devem ser ainda mais acentuados para os segmentos de textos mais longos e para amostras entre registros. Esses resultados sustentam a preferência por amostragem estratificada – uma maior diversidade entre os textos incluídos em um *corpus* significará uma representação mais ampla dos tipos de características linguísticas.

No que diz respeito à questão da extensão do texto, o sentido da presente seção é simplesmente que as amostras de texto devem ser longas o suficiente para representar de maneira confiável as distribuições das características linguísticas. Para características distribuídas de modo linear, a extensão necessária depende da estabilidade geral da característica. Para as características curvilíneas, um corte arbitrário deve ser especificado marcando uma representação "adequada", por exemplo, quando os segmentos dos textos subsequentes contribuem com menos de 10% para novos tipos adicionais. Dado um esforço finito investido no desenvolvimento de um *corpus*, uma representação linguística mais ampla pode ser alcançada, concentrando-se na diversidade de textos e tipos de textos diferentes em vez de concentrar-se em amostras mais longas dentro de textos.

As propostas específicas do tamanho do texto requerem investigações mais detalhadas do que foi apresentado aqui, incidindo especialmente sobre a distribuição das características menos estáveis, para determinar o tamanho do texto necessário

para estabilidade, e sobre a distribuição de outros tipos de características (por exemplo, as características discursivas e de empacotamento das informações).

4.2. As distribuições entre textos diversos: número de textos

Uma segunda questão estatística importante na construção de um *corpus* de textos refere-se à amostragem de textos: como as características linguísticas são distribuídas entre os textos e registros, e quantos textos devem ser coletados para o *corpus* total e para cada registro para representar as distribuições?

4.2.1. Pesquisas anteriores sobre a variação linguística dentro de e entre registros

Embora os registros sejam definidos com referência às características situacionais, eles podem ser analisados linguisticamente, e há diferenças linguísticas importantes entre eles; ao mesmo tempo, alguns registros também têm uma gama relativamente grande de variação linguística interna (ver Biber, 1988, Capítulos 7 e 8). Por essa razão, a caracterização linguística de um registro deve incluir sua tendência principal e sua gama de variação. Na verdade, alguns registros são semelhantes em suas tendências principais, mas diferem significativamente em sua gama de variação (por exemplo, ficção científica *versus* ficção em geral, e documentos oficiais *versus* discurso acadêmico, em que o primeiro registro de cada par tem uma gama de variação mais restrita). Em Biber (1988, pp. 170-198), descrevo a variação linguística dentro dos registros, incluindo as relações linguísticas entre diversos sub-registros.

O número de textos necessários em um *corpus* para representar determinados registros está diretamente relacionado com o tamanho da variação interna. Em Biber (1990), analiso a estabilidade das contagens das características entre textos de um registro, comparando as contagens da frequência média das subamostras dos 10 textos tirados de determinados registros. Cinco registros foram analisados: conversas, discursos públicos, reportagem de imprensa, discurso acadêmico e ficção em geral. Três amostras dos 10 textos foram extraídas de cada um desses registros, e as contagens da frequência média de seis características linguísticas foram comparadas entre as amostras (pronomes de primeira e de terceira pessoa, verbos no pretérito, substantivos, preposições, voz passiva). A análise de confiabilidade dessas frequências médias entre as três amostras dos 10 textos mostraram um altíssimo grau de estabilidade para todas as seis características linguísticas (todos os coeficientes foram superiores a 0,95). Esses coeficientes mostram que as médias das amostras dos 10 textos são altamente correlacionadas; ou seja, as tendências linguísticas principais desses registros com relação a essas características linguísticas são bastante estáveis, mesmo quando medidas pelas amostras de 10 textos. No entanto, há duas questões importantes que não estão sendo

consideradas nesta análise. Em primeiro lugar, as seis características linguísticas consideradas foram todas relativamente comuns; características raras, tais como orações relativas QU- ou subordinação condicional, podem mostrar um grau de confiabilidade muito menor. Em segundo lugar, esta análise abordou o número de textos que foram necessários para representar confiavelmente os valores médios, mas não abordou a representação da diversidade linguística nos registros[11].

4.2.2. Variação linguística total em um corpus*: tamanho da amostra total para um* corpus

Na Seção 3, discuti como o tamanho necessário da amostra está relacionado com o erro padrão (Sx^-) pela equação:

$$Sx^- = S/n^{1/2} \quad (4.1)$$

O cálculo real do tamanho da amostra depende de uma especificação do erro tolerável (*et*):

$$et = t * Sx^- \quad (4.2)$$

A equação 4.2 indica que o erro tolerável é igual ao erro padrão multiplicado pelo valor *t*. Dada uma amostra de tamanho superior a 30 (que permite a suposição de uma distribuição normal), um pesquisador pode saber com 95% de certeza que o valor médio de uma amostra vai cair no intervalo da verdadeira média da população mais ou menos o erro tolerável.

A equação 4.2 pode ser manipulada para fornecer uma segunda equação para calcular o erro padrão, ou seja, $Sx^- = et/t$. Se a razão et/t for substituída por Sx^- na Equação 4.1, e a equação é então resolvida para *n*, obtemos um cálculo direto do tamanho da amostra necessária para um *corpus*:

$$n = S^2/(et/t)^2 \quad (4.3)$$

onde *n* é o tamanho calculado da amostra, *S* é o desvio padrão estimado para a população, *et* é o erro tolerável (igual a 1/2 do intervalo de confiança desejado) e *t* é valor *t* para o nível de probabilidade desejado.

Na Seção 3, observo que existem problemas na aplicação da Equação 4.3. Em certo sentido, a equação simplesmente desloca o peso de responsabilida-

11 Por exemplo, um dos tipos de texto mais marcantes identificados em Biber (1989) consiste de textos em que o remetente está produzindo uma reportagem *on-line* de eventos em andamento. Linguisticamente, esse tipo de texto é marcado por ser extremamente situado em referências (muitos advérbios de tempo e de lugar, e uma orientação de um tempo presente). Infelizmente, há apenas sete textos dessa natureza nos *corpora* London-Lund e LOB, indicando que esse tipo de texto é sub-representado e precisa ser objeto de análise no desenvolvimento de um *corpus* futuramente.

de – da estimativa da quantidade desconhecida para o tamanho necessário da amostra para a estimativa das quantidades desconhecidas para o erro tolerável e o desvio padrão da população –, isto é, a fim de utilizar a equação, é necessário que haja uma estimativa prévia do erro tolerável ou do intervalo de confiança permitido no *corpus* e uma estimativa prévia do desvio padrão das variáveis na população como um todo.

O erro tolerável depende da precisão exigida das estimativas da população com base na amostra do *corpus*. Por exemplo, digamos que quiséssemos saber quantos substantivos ocorrem, em média, em textos de conversa. O intervalo de confiança é a margem na qual podemos ter 95% de certeza de que a verdadeira média da população se encontra. Por exemplo, se a média da amostra para substantivos em conversas foi de 120, e precisássemos estimar a verdadeira média da população de substantivos com uma precisão de ± 2, então o intervalo de confiança seria 4, indo de 118 a 122. O erro tolerável é simplesmente um lado (ou uma metade) do intervalo de confiança. O problema aqui é que é difícil fornecer uma estimativa *a priori* da precisão exigida da análise que será baseada em um *corpus*.

Problemas semelhantes surgem com as estimativas dos desvios padrão. Nesse caso, não é possível estimar o desvio padrão de uma variável em um *corpus* sem já se ter uma amostra representativa de textos. Aqui, assim como em muitos aspectos do planejamento do *corpus*, o trabalho deve ser feito de maneira circular, com investigações empíricas com base em *corpora*-piloto informando o processo do planejamento. O problema para o planejamento inicial do *corpus*, porém, consiste em fornecer uma estimativa inicial do desvio padrão.

Um último problema é que os desvios padrão devem ser estimados quanto a variáveis específicas, mas, no caso da linguística de *corpus*, há inúmeras variáveis linguísticas de interesse. A escolha de variáveis diferentes, com desvios padrão diferentes, resultará em diferentes estimativas do tamanho da amostra necessária.

Nesta seção, utilizo as análises em Biber (1988, pp.77-78, 246-269) para abordar os dois primeiros desses problemas. Esse estudo é baseado em um *corpus* de textos em inglês relativamente grande e abrangente: 481 textos retirados de 23 registros falados e escritos. As análises estatísticas desse *corpus* podem assim ser usadas para fornecer estimativas iniciais para o erro tolerável e para o desvio padrão da população.

No planejamento de um *corpus* de texto, o erro tolerável não pode ser afirmado em termos absolutos, porque a magnitude das contagens de frequência varia consideravelmente entre as características (como foi mostrado na Seção 3). Por exemplo, um erro tolerável de ± 5 pode funcionar bem para características comuns, tais como substantivos, que têm uma média geral de 180,5 em 1.000 palavras no *corpus*-piloto, mas não seria aceitável para características raras, como as orações condicionais subordinadas, que têm uma média geral de apenas 2,5 no *corpus* (de modo que um erro tolerável de 5 se traduziria em um intervalo de confiança de -2,5 a 7,5, e um texto poderia ter três vezes o número médio de orações condicionais e ainda estar dentro do intervalo de confiança). Em vez disso, proponho

aqui um cálculo da estimativa do erro tolerável distinto para cada característica linguística com base na magnitude do valor médio para a característica; para ilustrar, especificarei o erro tolerável de ± 5% do valor médio (para um intervalo de confiança total de 10% do valor médio). A Tabela 4 apresenta o valor médio e o desvio padrão de sete características linguísticas no *corpus*-piloto, juntamente com o erro tolerável calculado para cada característica. Pode-se observar que o erro tolerável varia de 9,03 para substantivos (que têm uma média de 180,5) até 0,13 para as orações condicionais (que têm uma média de apenas 2,5).

Dados os erros toleráveis e os desvios padrão estimados listados na Tabela 4, o tamanho necessário da amostra (isto é, o número total de textos a serem incluídos no *corpus*) pode ser calculado diretamente usando a Equação 4.3. A Tabela 4 mostra diferenças muito grandes no tamanho da amostra necessária entre essas características linguísticas. Essas diferenças são uma função do tamanho do desvio padrão em relação à média para uma determinada característica. Se o desvio padrão é muitas vezes menor do que a média, como no caso das características comuns, como os substantivos e as preposições, o tamanho necessário da amostra é muito pequeno. Se, por outro lado, o desvio padrão se aproxima da média em magnitude, como no caso das características raras, tais como as orações relativas QU- e as orações condicionais, o tamanho necessário da amostra se torna muito grande. Os marcadores do tempo pretérito são interessantes na medida em que eles são relativamente comuns (média de 40,1), mas com um desvio padrão relativamente elevado (30,4). Portanto, requerem uma amostra relativamente grande de textos para representação (883). Em geral, a abordagem mais conservadora no planejamento de um *corpus* seria usar a característica mais variável (proporcional à sua média, nesse caso as orações condicionais) para definir o tamanho total da amostra.

	Valor médio no *corpus*-piloto	Desvio padrão no *corpus*-piloto	Erro tolerável	*N* necessário
Substantivos	180,5	35,6	9,03	59,8
Preposições	110,5	25,4	5,53	81,2
Verbos no presente	77,7	34,3	3,89	299,4
Verbos no pretérito	40,1	30,4	2,01	883,1
Voz passiva	9,6	6,6	0,48	726,3
Orações relativas QU-	3,5	1,9	0,18	452,8
Orações condicionais	2,5	2,2	0,13	1.190,0

Tabela 4

4.2.3. Variação linguística dentro dos registros: números de textos necessários para representar registros

A questão que permanece diz respeito ao tamanho necessário da amostra para cada registro. Embora a maioria dos livros sobre planejamento de amostra simplesmente recomende a amostragem proporcional para o planejamento de amostras estratificadas (ver Seção 3), poucos livros discutem a necessidade

de utilizar amostragem estratificada não proporcional em certos casos; no entanto, esses livros diferem quanto ao método para determinar o tamanho recomendado da amostra para subgrupos. Por exemplo, Sudman (1976, pp. 110-111) afirma que uma amostragem estratificada não proporcional deve ser usada quando os próprios subgrupos são de interesse primário (como no caso de um *corpus* de texto), e que os tamanhos das amostras dos subgrupos deveriam ser iguais nesse caso (para minimizar o erro padrão da diferença). Esse procedimento é apropriado quando as variâncias dos subgrupos são aproximadamente as mesmas. Em contrapartida, Kalton (1983, pp. 24-25) recomenda o uso do desvio padrão de subgrupo para determinar os tamanhos relativos das amostras. Esse procedimento é mais adequado para o planejamento do *corpus*, uma vez que os desvios padrão das características linguísticas variam consideravelmente de um registro para outro.

Embora não faça recomendações específicas para o tamanho da amostra do registro, ilustro essa abordagem na Tabela 5, considerando as variâncias relativas de sete características linguísticas (as mesmas da Tabela 4) entre três registros: conversas, ficção em geral e discurso acadêmico. Como acima, os dados são retirados de Biber (1988, pp. 246-269).

Conversas. Média do desvio normalizado = 0,37

	Valor médio no *corpus*-piloto	Desvio padrão no *corpus*-piloto	Razão do desvio padrão/média (desvio normalizado)
Substantivos	137,4	15,6	0,11
Preposições	85,0	12,4	0,15
Verbos no presente	128,4	22,2	0,17
Verbos no pretérito	37,4	17,3	0,46
Voz passiva	4,2	2,1	0,50
Orações relativas QU-	1,4	0,9	0,64
Orações condicionais	3,9	2,1	0,54

Ficção em geral. Média do desvio normalizado = 0,39

	Valor médio no *corpus*-piloto	Desvio padrão no *corpus*-piloto	Razão do desvio padrão/média (desvio normalizado)
Substantivos	160,7	25,7	0,16
Preposições	92,8	15,8	0,17
Verbos no presente	53,4	18,8	0,35
Verbos no pretérito	85,6	15,7	0,18
Voz passiva	5,7	3,2	0,56
Orações relativas QU-	1,9	1,1	0,58
Orações condicionais	2,6	1,9	0,73

Discurso acadêmico. Média do desvio normalizado = 0,49

	Valor médio no *corpus*-piloto	Desvio padrão no *corpus*-piloto	Razão do desvio padrão/média (desvio normalizado)
Substantivos	188,1	24,0	0,13
Preposições	139,5	16,7	0,12
Verbos no presente	63,7	23,1	0,36
Verbos no pretérito	21,9	21,1	0,96
Voz passiva	17,0	7,4	0,44
Orações relativas QU-	4,6	1,9	0,41
Orações condicionais	2,1	2,1	1,00

Tabela5

A Tabela 5 apresenta o valor médio, o desvio padrão e a razão entre o desvio padrão e o valor médio para sete características linguísticas nos três registros. A razão representa o desvio normalizado de cada uma dessas características dentro de cada registro – o tamanho da variação interna em relação à magnitude do valor médio. O desvio padrão bruto não é adequado aqui (semelhante à Tabela 4), pois os valores médios das características variam de maneira muito ampla.

A Tabela 5 mostra que o desvio padrão normalizado varia consideravelmente entre as características dentro de um registro. Por exemplo, em conversas, as contagens de substantivos, preposições e verbos no presente mostram os desvios normalizados relativamente pequenos, enquanto a voz passiva, as orações relativas QU- e as orações condicionais mostram os desvios normalizados iguais ou superiores a 50%. Conforme foi apresentado anteriormente, as características com menor frequência geral tendem a apresentar desvios normalizados bem maiores.

Há também grandes diferenças entre os diversos registros. Por exemplo, os verbos no pretérito têm uma variação normalizada de 46% em conversas e apenas 18% na ficção em geral, mas apresentam uma variação normalizada de 96% no discurso acadêmico. A subordinação condicional também mostra grandes diferenças entre esses três registros: apresenta uma variação normalizada de 54% nas conversas, de 73% na ficção em geral e de 100% no discurso acadêmico.

Para determinar o tamanho da amostra para cada registro, é necessário calcular uma única medida da variação dentro de cada registro. Essa medida é então utilizada para alocar uma amostra proporcionalmente maior para registros com desvios maiores. (Isto não deve ser confundido com uma representação proporcional dos registros.) Um número mínimo de textos deve ser alocado para cada registro (por exemplo, pelo menos 20 textos por registro), e em seguida os textos restantes no *corpus* podem ser divididos proporcionalmente em função da variação relativa dentro dos registros.

Para ilustrar, analisemos a Tabela 5 novamente. Essa tabela apresenta um desvio médio normalizado para cada registro, o que representa um valor do desvio total calculado pelo estabelecimento da média dos desvios padrão normalizados das sete características linguísticas. As conversas e a ficção em geral apresentam desvios gerais relativamente semelhantes (37% e 39%, respectivamente), enquanto o discurso acadêmico tem um desvio geral ligeiramente superior (49%). Para seguir com este exemplo, suponhamos que houvesse um total de 200 textos em um *corpus*, tomados a partir desses três registros. Cada registro teria o mínimo de 20 textos, deixando 140 textos para serem divididos proporcionalmente entre os três registros. Para determinar o tamanho da amostra relativa dos registros, a seguinte equação teria que ser resolvida baseada nos desvios gerais relativos.

$$0{,}37x + 0{,}39x + 0{,}49x = 140$$
$$1{,}25x = 140$$
$$x = 112$$

e, assim, os tamanhos das amostras seriam:

Conversa 0,37 * 112 = 41
Ficção em geral 0,39 * 112 = 44
Discurso acadêmico 0,49 * 112 = 55
Total de textos atribuídos = (41 + 20 para conversa) + (44 + 20 para ficção em geral) + (55 + 20 para discurso acadêmico) = 200 textos

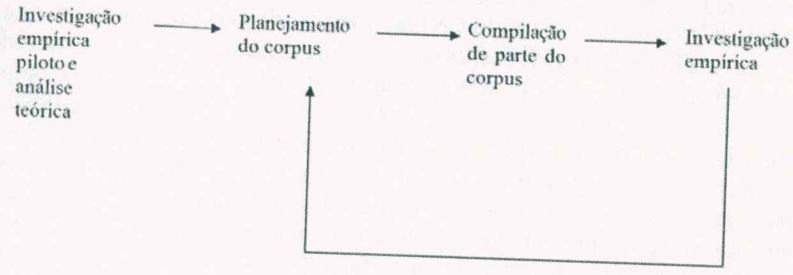
Para calcular os valores reais para o tamanho das amostras dos registros, é necessário analisar toda a gama de características linguísticas em todos os registros, calculando um único valor do desvio médio para cada registro. Isso poderia ser feito calculando-se a média entre as variantes normalizadas de todas as características linguísticas, conforme ilustrado aqui. Uma abordagem alternativa seria usar as variantes normalizadas de todas as dimensões linguísticas identificadas em Biber (1988). Essa última abordagem teria uma base teórica mais sólida, na medida em que as dimensões representam os parâmetros básicos da variação entre os registros, sendo cada uma baseada em um padrão importante de coocorrência entre as características linguísticas. Em contrapartida, a abordagem ilustrada nesta seção depende da influência conjunta das características linguísticas de modo isolado, e assim as distribuições, relativamente anormais, podem ter uma influência relativamente forte sobre o resultado final. Além disso, o uso das dimensões permite uma análise das distribuições com relação a determinados parâmetros funcionais, de modo que algumas dimensões possam ter mais peso do que outras. Por outro lado, não há uma forma motivada para distinguir entre a gama de características individuais por razões funcionais.

Está além do escopo deste trabalho ilustrar o uso dos valores das dimensões para a caracterização linguística dos registros (uma vez que eu precisaria antes explicar as bases teóricas e metodológicas das dimensões). Entretanto, a mesma abordagem básica, ilustrada nesta seção, seria usada. A principal diferença envolve a análise de desvio ao longo das dimensões básicas da variação linguística, ao invés de considerar as diversas características linguísticas de maneira isolada.

5. Conclusão: início

Tentei desenvolver aqui um conjunto de princípios para obter "representatividade" no planejamento de *corpus*. Ofereci recomendações específicas em relação a alguns aspectos do planejamento de *corpus* e remeti a exemplos em outras referências (a respeito de questões importantes sobre as recomendações

finais, que não poderiam ser desenvolvidas em um trabalho deste porte). O ponto principal no planejamento do *corpus*, porém, é que os parâmetros de um *corpus* plenamente representativo não podem ser determinados logo no início. Em vez disso, o trabalho do *corpus* funciona de maneira cíclica, que pode ser representado esquematicamente da seguinte forma:



A pesquisa teórica deve sempre preceder o planejamento inicial do *corpus* e a compilação de textos em si. Determinados tipos de pesquisa podem ter bons desenvolvimentos antes de qualquer investigação empírica, tais como identificar os parâmetros situacionais que distinguem entre os textos em uma comunidade discursiva e identificar a gama de características linguísticas importantes que serão analisadas no *corpus*. Outras questões relacionadas ao planejamento, no entanto, dependem de um *corpus*-piloto de textos para as investigações preliminares. Os pesquisadores contemporâneos de *corpora* de língua inglesa são extremamente afortunados, pois eles têm *corpora* como Brown, LOB e London-Lund para as pesquisas-piloto, proporcionando uma base empírica sólida para o planejamento inicial do *corpus*. Os compiladores desses *corpora* não tinham qualquer *corpus*-piloto para orientar seus planejamentos. Há situações semelhantes para projetos atuais de planejamento de *corpora* na representação de línguas não ocidentais. Por exemplo, um *corpus* recentemente concluído de somali exigiu um trabalho prático muito grande para orientar o planejamento inicial (ver Biber e Hared, 1992). Assim, o planejamento inicial de um *corpus* será mais ou menos avançado, dependendo da disponibilidade de pesquisas e *corpora* anteriores.

Independentemente do planejamento inicial, a compilação de um *corpus* representativo deve ser executada de maneira cíclica: um *corpus*-piloto deve ser compilado primeiro, representando uma gama relativamente ampla de variação, mas também representando um aprofundamento em alguns registros e textos. A etiquetagem gramatical deve ser feita nesses textos, servindo de base para investigações empíricas. Em seguida, a pesquisa empírica deve ser realizada nesse *corpus*-piloto para confirmar ou alterar os diversos parâmetros do planejamento. Algumas partes desse ciclo poderiam ser realizadas de maneira

quase contínua, com textos novos a serem analisados na medida em que eles se tornam disponíveis, mas também deve haver fases discretas de investigação empírica ampla e revisão do planejamento do *corpus*.

Por fim, deve-se notar que várias técnicas multivariadas podem ser aproveitadas para essas investigações empíricas. Neste artigo, limitei-me a técnicas univariadas e a simples estatística descritiva. Outro estudo, porém, sugere a utilidade de duas técnicas multivariadas para a análise da variação linguística em *corpora* informatizados: análise fatorial e análise de *cluster*. A análise fatorial pode ser utilizada tanto de modo exploratório (por exemplo, Biber, 1988) como em análises de "confirmação" baseadas em teoria (por exemplo, Biber, 1992). As duas seriam adequadas para o trabalho de planejamento de *corpus*, especialmente para a análise da gama e dos tipos de variação dentro de um *corpus* e nos registros. Tais análises indicariam se os diferentes parâmetros de variação foram igualmente bem representados e forneceriam uma base para decisões sobre o tamanho da amostra. A análise de *cluster* foi utilizada para identificar os "tipos de texto" em inglês – categorias de textos definidas em termos estritamente linguísticos (Biber, 1989). Os tipos de texto não podem ser identificados com base em fundamentos *a priori*; em vez disso, eles representam os agrupamentos de textos em um *corpus* que são semelhantes em suas caracterizações linguísticas, independente de suas categorias de registro. Em condições ideais, um *corpus* representaria a gama de registros e a gama de diferentes tipos de texto em uma língua e, por isso, é necessária a pesquisa sobre a variação tanto dentro dos dois tipos de categorias de texto[12] como entre eles.

Em suma, o planejamento de um *corpus* representativo não é finalizado de maneira absoluta até que o *corpus* esteja concluído, e as análises dos parâmetros de variação são necessárias durante todo o processo de desenvolvimento do *corpus* a fim de aumentar a representatividade da coleção resultante de textos.

Agradecimentos

Gostaria de agradecer a Edward Finegan por seus diversos comentários úteis sobre uma versão preliminar deste trabalho. Uma versão modificada deste artigo foi distribuída no *Pisa Workshop on Textual Corpora*, realizado na Universidade de Pisa (janeiro/1992), e os debates com vários dos participantes do *workshop* também foram úteis na revisão do artigo.

12 Por exemplo, um dos tipos de texto mais marcantes identificados em Biber (1989) consiste de textos em que o remetente está produzindo uma reportagem *on-line* de eventos em andamento. Linguisticamente, esse tipo de texto é marcado por ser extremamente situado em referências (muitos advérbios de tempo e de lugar, e uma orientação de um tempo presente). Infelizmente, há apenas sete textos dessa natureza nos corpora London-Lund e LOB, indicando que esse tipo de texto é sub-representado e precisa ser objeto de análise no desenvolvimento de um corpus futuramente.

Referências:

Biber, D. Spoken and Written Textual Dimensions in English: Resolving the Contradictory Findings. *Language*, 62: 384-414, 1986.

___ *Variation across Speech and Writing.* Cambridge University Press: Cambridge. (1989). A Typology of English Texts, *Linguistics*, 27:3-43, 1988.

___ Methodological Issues Regarding *Corpus*-basedAnalyses of Linguistic Variation. *Literary and Linguistic Computing*, 5: 257-69, 1990.

___ On the Complexity of Discourse Complexity: AMultidimensional Analysis. *Discourse Processes.* 15: 133—163, 1992.

___ An Analytical Framework for Register Studies. *ives on Register.* Oxford University Press, New York. Inpress, 1993a.

___ Register Variation and *Corpus* Design, *ComputationalLinguistics.* In press, 1993b.

___; Hared, M. Dimensions of Register Varia-In D. Biber and E. Finegan (eds), *Sociolinguistic Perspect*tion in Somali, *Language Variation and Change*, 4: 41-75, 1992.

Brown, P.; Fraser, C. Speech as a Marker of Situation. In K. R. Scherer and H. Giles (eds), *SocialMarkers in Speech.* Cambridge University Press, Cambridge,pp.33-62, 1979.

Duranti, A. Sociocultural Dimensions of Discourse. In T. van Dijk (ed.), *Handbook of Discourse Analysis,*Vol. 1. Academic Press: New York: pp.193-230, 1985.

Francis, W. N.; Kucera, H. *Manual of Information to Accompany a Standard Corpus of Present-Day Edited American English, for use with Digital Computers.* Department of Linguistics, Brown University, 1964/1979.

Halliday, M. A. K.; Hasan, R. *Language, Context,and Text: Aspects of Language in a Social-semiotic Perspective.* Oxford University Press, Oxford, 1989.

Henry, G. T. *Practical Sampling.* Sage, Newbury Park, CA, 1990.

Hymes, D. H. *Foundations in Sociolinguistics.* University of Pennsylvania Press, Philadelphia, 1974.

Johansson, S.; Leech, G. N.; Goodluck, H. Manual of information to accompany the Lancaster-Oslo/Bergen *Corpus* of British English, for use with digital computers. Department of English, University of Oslo, 1978.

Kalton, G. *Introduction to survey sampling.* Sage, Newbury Park, CA, 1983.

Sudman, S. *Applied Sampling.* Academic Press, New York, 1976.

Svartvik, J. and Quirk, R. (eds) *A Corpus of English Conversation.* C. W. K. Gleerup, Lund, 1980.

Williams, B. *A Sampler on Sampling.* John Wiley and Sons, New York, 1978.