PLANEJAMENTO, COMPILAÇÃO E ORGANIZAÇÃO DE CORPORA*

Simone Vieira Resende[1]
Rafael Maverick[2]

Resumo

É inquestionável o papel insigne da Linguística de Corpus (LC) na investigação da linguagem. São muitas as investigações que fazem uso do ferramental teórico metodológico da LC, como os trabalhos de Berber Sardinha e São Bento Ferreira (2014); Berber Sardinha, Kauffmann e Mayer Acunzo (2014); Biber, Conrad and Reppen (2009); Biber (1993, 2006) que asseguram à LC uma prestabilidade quase infinda. O mesmo acontece quando se trata de investigações sobre corpus design em língua inglesa, com publicações antigas como Biber et al. (1998); Crowdy (1993, 1994); Hunston (2002); McCarthy (1998). Poucos são os trabalhos em língua portuguesa sobre o planejamento e a organização de corpora. Esse artigo busca colaborar para o preenchimento dessa lacuna por meio da descrição da organização de dois corpora: o CCC**,** Corpus Comparável de Curriculum Vitae e o CARTS, Corpus of American Reality TV Shows. Os dois têm o objetivo de servirem como fontes para pesquisas linguísticas, terminológicas e material para ensino de tradução e língua. Resultados mostram que a pesquisa teórica deve ser anterior ao planejamento e compilação do corpus. A base teórica dá ao pesquisador a possibilidade de selecionar e compilar com mais segurança e propriedade, podendo justificar suas escolhas e posteriormente enriquecer suas análises. Estudos dessa natureza podem contribuir de forma satisfatória no planejamento de futuros corpora.

Palavras-chave: Linguística de Corpus; Planejamento de Corpus; Desenho de Corpus; CCC; CARTS;

Abstract

It is unquestionable the notable role which Corpus Linguistics plays in language investigation. There are many studies that make use of the theoretical and methodological tools of Corpus Linguistics, such as the works of Berber Sardinha and Ferreira (2014); Berber Sardinha, Kauffman and Mayer Acunzo (2014); Biber e Conrad (2009); Biber (2006) which assure to Corpus Linguistics an almost endless helpfulness. The same can be said when dealing with researches about Corpus Design in English; there are old publications such as Biber et al. (1998); Crowdy (1993, 1994); Hunston (2002); McCarthy (1998). There are very few works in Portuguese about corpora planning and the organization. This paper aims to fill this gap through the description of the organization of two corpora: CCC – Comparable Corpus of *Curriculum Vitae* – and CARTS - Corpus of American Reality TV Shows. Both corpora have the goal of being resources for linguistics and terminological researches, and teaching material to be used in translation and language teaching and learning. Investigations point out that the theoretical study must be before the corpus planning and organization. The theoretical base gives the researcher the possibility to select and compile with more safety and accuracy, being able to justify all choices, and posteriorly improve the analysis. Studies like this can contribute in a satisfactory way to the future corpora planning.

Key Words: Corpus Linguistics; Corpus Planning; Corpus Design; CCC, CARTS;

* Gostaríamos de agradecer à Capes (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) pelo financiamento dos anais da VII Escola Brasileira de Linguística Computacional e do XIII Encontro de Linguística de Corpus, processo nº 3472/2015-87.

[1]Simone V Resende é tradutora, mestre em Linguística e doutoranda do Programa de Pós-graduação em Linguística Aplicada e Estudos da Linguagem da PUC-SP. simonevieiraresende@gmail.com

[2] Rafael Maverick é professor de inglês e mestrando do Programa de Pós-graduação em Linguística Aplicada e Estudos da Linguagem da PUC-SP.

1. Introdução

O objetivo desse artigo é descrever as etapas de planejamento, compilação, organização e codificação de arquivos de texto na construção de dois *corpora*: **CCC** (Corpus de Comparável de *Curriculum Vitae*) e **CARTS** (*Corpus of American Reality TV Shows*), bem como algumas implicações relevantes e o percurso teórico-metodológico envolvidos neste processo. Cada pesquisador, a partir da busca e seleção de textos disponibilizados na internet, construiu seus *corpora* com características específicas para atender propósitos diferentes, entretanto, encontraram os mesmos desafios e percorreram caminhos semelhantes no processo de busca, seleção e criação.

A escolha de dois corpora de estudo tão distintos deve-se a necessidade de analisar a representatividade de forma mais abrangente, ou seja, incluindo usos variados da língua durante a compilação dos *corpora* para que as análises sejam enriquecidas, sem esquecer, porém, a predominância de um aspecto linguístico específico, como um tipo de discurso, nesse caso o discurso do Curriculum Vitae (modo escrito) e a inclusão de amostras do discurso oral (modo falado), como no caso das falas dos *Reality TV Shows.*

Segundo Miller e Biber (2015, p. 30) as considerações a respeito do design dos corpora geralmente focam no que eles chamam de representatividade externa, ou seja, aquela que representa o domínio discursivo alvo e muitas vezes deixam de lado as questões referentes à representatividade interna, ou seja, aquelas questões que levam em consideração se os corpora proporcionam uma descrição confiável das variações linguísticas.

O intuito de dissertar a respeito da compilação e organização de *corpora* pode ser justificado se considerarmos que à primeira vista, por pura intuição, pode-se acreditar que basta compilar um número colossal de textos para se obter um corpus eficiente, robusto e representativo, o que não é verdade. Segundo Biber (2012, p.11), não basta compilar os textos aleatoriamente, há necessidade de uma compilação cíclica, que segundo ele envolveria "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." Sendo assim, o pesquisador deveria, no processo de compilação e organização, rever seu planejamento inicial. Esse artigo relata a experiência dos dois pesquisadores ao colocar os ciclos sugeridos por Biber em prática.

É situação *sine qua non* que a compilação sistemática de *corpora* exige a organização e posteriormente análise empírica com o auxílio de ferramentas computadorizadas de grandes quantidades de dados linguísticos, conforme afirmou Biber (2012, p. 11). Contudo, os ciclos mencionados anteriormente não são os únicos passos da compilação de um *corpus*. Esse artigo descreve o planejamento, organização e codificação de dois *corpora*, o CCC e o CARTS a partir de normas internacionais já estabelecidas para a montagem de *corpora*. Algumas das premissas na construção de um banco de dados dizem respeito ao planejamento geral. Para Biber (2012, p. 12) o pesquisador deve se preocupar em "escolher os tipos de textos que vão compor o *corpus*, o número de textos, a seleção de determinados textos, a seleção de amostras retiradas de textos e o tamanho das amostras de texto." Além desses aspectos, é preciso considerar também a origem dos dados, eles devem ser autênticos e a composição do corpus, ou seja, o conteúdo que será escolhido, deve ter uma formatação legível pelo computador. Por último, a representatividade e a extensão.

Os critérios usados na compilação do CCC e do CARTS são os sugeridos por Biber (1990, 1993). São critérios propostos a partir de princípios de amostragem por meio de cálculos estatísticos que determinam o número total de textos, o número de textos por variedade e o tamanho dos textos. Escolhemos usar os critérios propostos por Biber porque as concepções propostas por ele tornam possível a comparação de textos com tamanhos diferentes, assim como conjuntos de textos de tamanhos diferentes, sendo esse o caso dos dois corpora descritos nesta pesquisa.

Esse artigo está organizado em três partes distintas depois da introdução. A primeira trata das linhas teóricas e metodológicas que dão base para a pesquisa, a segunda trata da descrição dos corpora compilados e a terceira e última parte apresenta a análise dos dados organizacionais dos dois corpora e traça um paralelo entre a compilação dos dois bancos de dados.

2. Revisão da literatura

Para iniciar um estudo que trata da compilação e organização de *corpora* é necessária uma compreensão dos princípios básicos do que é um *corpus* e do que a Linguística de Corpus trata. A Linguística de Corpus (LC) é a área de estudos linguísticos que se interessa pela exploração de *corpora*. Para McEnery (2001), uma forma simplificada de definir o que é um corpus seria dizer que se trata de uma coletânea gigante de palavras e textos armazenados em um computador e que nos permite fazer consultas de forma rápida e precisa. Para Berber Sardinha (2004, p. 3), a LC é definida como a "área de estudo que se ocupa da coleta e da exploração de corpora, ou conjunto de dados linguísticos textuais coletados criteriosamente, com o propósito de servirem para a pesquisa de língua ou variedade linguística", ou seja, dedica-se à criação e análise de corpora, que são conjuntos de textos ou transcrições de fala criteriosamente coletados e armazenados em arquivos de computador e que possibilitam a investigação de ocorrências e coocorrências de padrões léxico-gramaticais frequentes de certas variedades textuais. Para Rajagopalan (2006, p. 161) a LC é considerada "um empreendimento empírico por excelência" e nos últimos anos tem trazido contribuições relevantes para os estudos da linguagem, atraindo cada vez mais pesquisadores e praticantes.

Evidentemente, a LC ganhou força e combustível a partir do surgimento do microcomputador pessoal nos anos oitenta, facilitando a atividade de coleta, seleção e investigação linguística que incrivelmente era realizada manualmente antes do advento da informática. A tecnologia facilitou a compilação e o crescimento das pesquisas com *corpora*. A internet é fonte abundante de informação, ela ajuda e facilita a busca e seleção de textos. Mas, por que compilar e usar um corpus? Para que serve esse tipo de dado?

O pesquisador monta e analisa um *corpus* por vários motivos. Uma grande quantidade de dados pode nos dizer muito sobre as tendências da língua ou variedade linguística. Um banco de dados grande pode revelar muito sobre o que seria comum e o que seria típico dentro de um contexto real de uso da língua. Os corpora também revelam exemplos de casos raros e excepcionais de uso, que seriam praticamente imperceptíveis se fossem analisados em contextos individuais ou mesmo apenas por meio do recurso humano e não o computacional. Diferente dos computadores que são rápidos e precisos, o pesquisador humano é lento e ainda comete erros.

Tanto no que tange a pesquisa qualitativa ou quantitativa é fundamental que os conteúdos dos corpora compilados sejam autênticos e não sofram nenhum tipo de manipulação. Conforme Resende (2011, p. 71), apesar do advento da internet é fundamental ressaltar a importância do planejamento na busca e o estabelecimento de critérios bem definidos na compilação dos corpora que vão possibilitar uma pesquisa mais direcionada a determinados registros, atendendo propósitos específicos traçados pelas questões e objetivos das pesquisas.

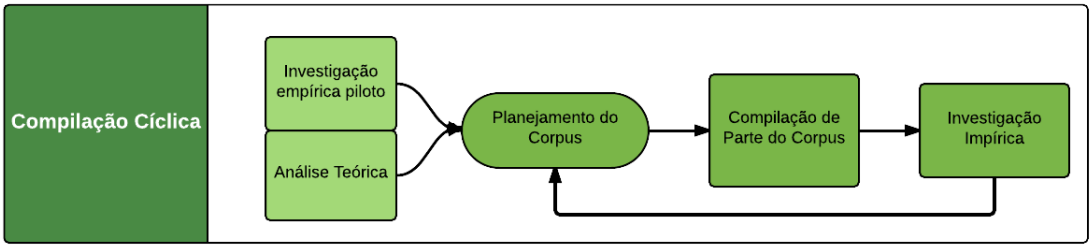
Para entender melhor a necessidade do planejamento e principalmente do planejamento cíclico é fundamental entender a forma pela qual Biber divide as duas grandes áreas dos estudos linguísticos: o estudo da estrutura e o estudo do uso. Os estudos da estrutura da língua são mais comuns, mas tão importante quanto os estudos do uso. Os corpora organizados aqui permitem a realização dos dois tipos de investigação e ambas permitem uma visão da dinamicidade dos corpora. Podemos dizer que essas duas perspectivas são complementares, porque para notarmos os traços do uso da linguagem, é preciso saber em que estrutura o uso está inserido e para se notarmos essa estrutura, é preciso entender o contexto de uso em que ela aparece.

Grande parte dos *corpora* compilados foram para atender a questões e objetivos de pesquisas em relação à investigação da frequência de padrões léxico-gramaticais e seus padrões de coocorrência, analisados por meio da utilização de ferramentas de concordância, listadores e etiquetadores, de forte tradição europeia, tendo como maior expoente John Sinclair. Encontramos no contexto norte-americano, uma outra vertente da LC é a da Análise Multidimensional (AMD) desenvolvida por Douglas Biber. Essa abordagem metodológica tem como objetivo principal identificar e descrever a variação de aspectos linguísticos e suas funções comunicativas dentro dos gêneros textuais ou dos gêneros do discurso, chamados por Biber de registro. Acreditamos que tal metodologia é a mais adequada para os estudos que vamos propor futuramente na análise dos corpora e por isso escolhemos seus princípios para iniciar o planejamento, a compilação e mais tarde a interpretação dos dados gerados a partir do mapeamento dos corpora. Tal adequação se deve ao seu caráter essencialmente computacional e comparativo, prestando-se a descrição, comparação e contraste de conjuntos de textos e não apenas de textos individuais. Berber Sardinha (2004, p. 300) define a AMD como sendo uma "abordagem para análise de corpus que usa procedimentos estatísticos (principalmente a análise fatorial) visando ao mapeamento das associações entre um conjunto variado de características linguísticas dentro de um corpus de estudo." Como esse artigo não pretende relatar os desdobramentos da AMD e sua aplicação nos corpora de estudo, não vamos prolongar a descrição de seus aspectos, contudo achamos importante pontuar algumas noções essenciais para embasar o início da compilação dos corpora que serão analisados de acordo com essa metodologia no futuro.

A AMD tem características sociolinguísticas, seus conceitos nucleares podem ser listados por meio dos termos que representam sua especialidade, como a variedade, a dimensão, o tipo de texto, o fator, os traços, as características e os registros. Para entender um pouco seus princípios, podemos dizer que ela investiga a linguagem usada em diferentes tipos de texto, tanto os falados quanto os escritos, uma vez que cada tipo de texto tem suas características bem definidas. Biber (1993, p. 5) usa o termo *variety*, que vamos traduzir como **variedade**, para referir-se à categoria dos textos que compartilham de uma mesma característica social ou situacional. Os **traços** (*features*) seriam os elementos nucleares para as análises, ou seja, os componentes linguísticos que podem ser quantificados nos textos. As características podem ser situacionais ou linguísticas. As características situacionais referem-se ao uso de uma variedade, e as linguísticas seriam os traços que escolhemos para quantificar. Estudando os escritos de Biber, podemos perceber que ele não faz distinção entre **gênero** e **registro**, ambos são definidos como sendo uma variedade definida por variáveis situacionais que podem ser identificadas no cotidiano de uso, como por exemplo, uma conversa natural, o editorial de uma revista, um texto acadêmico, uma conversa ao telefone, etc. Um grupo de textos formado com os mesmos critérios linguísticos é chamado por Biber de **tipo de texto**. O **fator**, assim como o tipo de texto, também é um grupo, porém um grupo de variáveis que coocorrem de forma significativa e que pode ser mensurado por meio de cálculos estatísticos. Assim que um fator é interpretado a partir da função comunicativa que ele exerce, ele passa a se chamar **dimensão** (BERBER SARDINHA, 2004, p. 304). Entendidos os pressupostos da AMD, podemos retomar os critérios elaborados para a compilação e o design de corpora usados no planejamento do CCC e do CARTS.

O planejamento, a compilação e o design de um corpus já foram temas de outros artigos acadêmicos. Orenha (2009) relatou o processo de organização e compilação de corpora comparáveis na área de negócios e sua contribuição para a tradução e terminologia. Oliveira e Dias (2009) relataram o processo de compilação de um corpus do português brasileiro e dão ênfase à compilação e a questão da representatividade. Miller e Biber (2015) descreveram a compilação de um corpus usado para estudos da variação do léxico, segundo eles, uma tarefa muito desafiadora devido à dificuldade de se conseguir uma estabilidade e resultados confiáveis a partir das análises. Eles analisaram um corpus representativo de um discurso bem específico, textos sobre introdução à psicologia extraídos de livros didáticos universitários. Eles descobriram que o design, a compilação e o planejamento têm mais influência na variação lexical do que se pensava e destacaram o quão importante é a avaliação interna da representatividade de um corpus, principalmente quando o objetivo é usar o corpus para uma investigação quantitativa. Brum de Paula e Espinar (2002) relatam a coleta e a transcrição de produções orais. Esses relatos contribuem de forma sistemática para uma orientação piloto do planejamento de novos corpora. No primeiro seminário sobre Estudos de Corpus realizado na USP em 1999, Berber Sardinha (1999) já traçava as principais noções de compilação de corpus em sua apresentação. Para condensar essas leituras e filtrar os aspectos mais nucleares desses processos selecionamos alguns passos principais para o planejamento e a compilação de um corpus, além dos sugeridos por Biber (1993) quando se refere à representatividade em um *corpus*. Ele sugere que um *corpus* seja representativo quanto ao idioma, a população, texto e o volume da amostra. É preciso definir a população alvo e os métodos de amostragem, bem como a representatividade. O planejamento do trabalho inclui a definição do tipo de corpus, a escolha dos registros, a escolha das fontes, a organização dos textos e posteriormente a anotação do corpus.

Quadro 1 – Compilação cíclica de corpora - (adaptado de Biber, 2012, p. 43)



A figura 1 apresenta os pontos principais para o planejamento de uma compilação cíclica de um corpus. Para Biber (2012, p. 43) os parâmetros de um corpus plenamente representativo não podem ser definidos logo no início da compilação. Em vez disso, o trabalho funciona de maneira cíclica, como mostra o gráfico acima. Em outras palavras, Biber acredita que um corpus piloto dever ser compilado primeiro, representando uma gama de variação e o aprofundamento no registro escolhido. Logo após deve-se fazer uma etiquetagem gramatical desses textos, em seguida entra em cena a pesquisa empírica. Ela vai ajudar a testar os parâmetros do planejamento e o ciclo vai se fechar novamente quando o planejamento for revisto, tudo isso pode acontecer de forma praticamente contínua, uma vez que novos textos podem ser inseridos durante o processo de compilação e de teste com os corpora pilotos. Contudo, a compilação precisa ter um fim. No momento em que a representatividade almejada é alcançada a compilação cessa.

No decorrer da história da organização de corpora, os pesquisadores geralmente se concentram em dois fatores primordiais: os tipos de discurso e o tamanho do corpus (BIBER, 2015, p. 33) e a questão da representatividade costuma ficar de lado. Porém, como os dois corpora descritos nessa pesquisa têm o objetivo de lançar mão de uma sofisticada técnica de análise estatística, como a AMD, a questão da representatividade passa a ser fundamental, afinal, fica quase impossível não se questionar se esses corpora realmente representam a parcela do discurso que se comprometem a descrever.

De acordo com Biber (2015, p. 33) há dois tipos principais de corpus representativos: os situacionais e os linguísticos. Da perspectiva do design de um corpus, isso significa que o corpus pode ser considerado a partir de uma abordagem situacional e a partir de uma perspectiva linguística e assim, o corpus será avaliado de acordo com a extensão que ele representa. Uma outra forma de explicar esses dois tipos de representatividade é relatada por McEnery et al. (2006, p. 14), ele descreve esses dois tipos de abordagens como sendo um de critérios externos e outro de critérios internos. Se pensarmos que a representatividade se refere ao quanto uma amostra inclui de toda a gama de variabilidade de uma população, vamos entender que durante o planejamento dos corpora, a variabilidade pode ser considerada a partir dessas duas perspectivas: situacional (externa) e linguística (interna). Assim, a variedade dos tipos de textos em uma língua (população) vai se referir às considerações externas e a variedade de distribuição linguística em uma língua (população) vai se referir às considerações internas. Segundo Biber (2015, p. 34), não é possível avaliar a representatividade externa se considerarmos apenas o *corpus* por si só. Em vez disso, deve-se considerar uma descrição detalhada do que Biber chama de mundo externo e a partir disso avaliar até que ponto os textos incluídos nos corpora são amostras de um espectro completo de textos da população-alvo. Ambos CARTS e CCC são exemplos de corpora que possuem essa representatividade externa, uma vez que já foram desenhados para representar essa fatia situacional externa. Contudo, as considerações internas não foram esquecidas durante o planejamento, uma vez que ambos proporcionam ao pesquisador a oportunidade de investigar características linguística e fazer descobertas estáveis e confiáveis.

Tendo em vista que uma descrição linguística probabilística, como a que se pretende fazer posteriormente com o CARTS e com o CCC, envolve a questão da representatividade, vale ressaltar que segundo Berber Sardinha (2004), essa visão probabilística da linguagem defende que a representatividade pode ser assegurada por meio das respostas às seguintes perguntas: representativo de quê? Para quem? Para quê? Em outras palavras, a composição dos dois corpora leva em conta o conceito de representatividade do *corpus* (SINCLAIR, 2005), segundo o qual se deve levar em consideração quais sujeitos de pesquisa se quer representar (população), o mesmo ponto da pergunta levantada por Berber Sardinha. Em seguida, em qual situação se deseja representá-los (registro), quais tipos de documentos estão disponíveis para que sejam representados, bem como se tais documentos dão conta de representar a linguagem dos sujeitos de pesquisa naquela determinada situação. Tanto o CARTS, quanto o CCC preenchem os requisitos descritos e respondem as questões levantadas logo acima e que são descritas logo a seguir. Acreditamos ter contemplado o conceito de representatividade. A próxima seção trata da descrição dos passos seguidos para a compilação cíclica dos corpora.

3. Metodologia e Descrição dos *Corpora*

A metodologia de compilação de *corpora* será apresentada em cinco passos distintos, o primeiro trata da definição do tipo de corpus, a segunda trata da escolha dos textos e dos registros, a terceira lida com a escolha das fontes e a quarta trata da organização dos textos. A quinta e última etapa abrange a anotação do corpus.

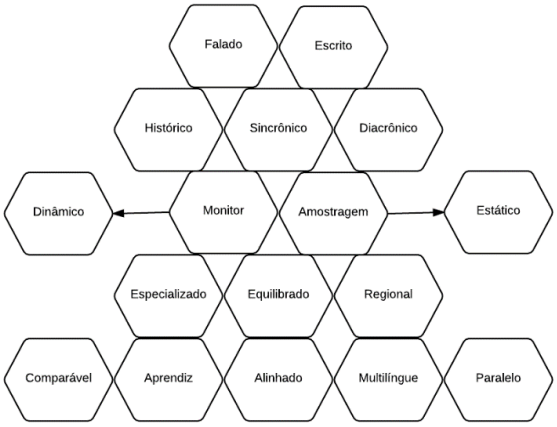
Segundo Biber (2012, p. 43) "a pesquisa teórica deve preceder o planejamento inicial do *corpus* e a compilação de textos em si." Para ele, determinados tipos de pesquisa podem ter bons desenvolvimentos antes de qualquer investigação empírica, ele sugere identificar os parâmetros situacionais que distinguem os textos em uma comunidade discursiva e ainda identificar a constelação de características linguísticas importantes que serão analisadas no corpus. Ele lembra ainda que algumas outras questões relacionadas ao planejamento vão depender de um *corpus* piloto que servirá para análises preliminares. Vale lembrar que os dois corpora descritos nesse artigo procuram seguir esses parâmetros e os passos seguidos são relatados aqui com o intuito de corroborar outros pesquisadores que desejam planejar e compilar corpora.

Em relação à relevância e representatividade do *corpus* é possível afirmar que em essência, um *corpus* deve ser representativo de uma língua ou variedade linguística, pois é uma amostragem de certa variedade textual contendo características linguísticas articuladas ao contexto de uso e atendendo à funções comunicativas específicas, no entanto, para que se possa atender ao aspecto de representatividade, um *corpus* precisa se também maior possível para dar conta de que o maior número possível das características possam o ocorrer, uma vez que um *corpus* pequeno pode ressaltar características léxico-gramaticais que não refletem a realidade. É uma tarefa árdua estabelecer com precisão o tamanho de um *corpus*, no entanto, certamente quanto maior seu tamanho, mais representativo será. Contudo, como descrito nas considerações teóricas, o tamanho não é tudo. Os corpora devem ser representativos externa e internamente. Segundo Biber (2015, p. 34) uma maneira simples de se avaliar a representatividade interna de um corpus e dividindo-o em subcorpora, principalmente durante os testes com os corpora-pilotos. Caso o pesquisador obtenha as mesmas descobertas com as investigações feitas nos subcorpora e no corpus como um todo, há, portanto, marcas que asseguram a representatividade do corpus, mostrando que ele realmente representa uma parcela significativa dos padrões linguísticos de variação. Segundo Biber (2015, p. 34), a representatividade interna do corpus, ou seja, as perspectivas linguísticas, devem sempre ser avaliadas porque são elas que nos "permitem determinar se o corpus possibilita uma análise linguística quantitativa confiável." Segundo ele, esse é um pré-requisito para realmente documentar e representar as distribuições linguísticas.

A relevância na construção de *corpora* tem relação estreita com o conceito de relevância de *corpus*, ou seja, o corpus deve ser representativo (uma boa amostragem) de uma língua ou variedade dela, na medida que pode, também, responder questões referentes aos objetivos e questões de uma pesquisa. “Além de representativo, o corpus deve ser adequado aos interesses do pesquisador, que deve ter uma questão a investigar para a qual necessite de um corpus específico. ” (BERBER SARDINHA, 2004, p. 29).

No caso dos corpora tratados nesse artigo, O CCC e o CARTS, o planejamento de compilação e design dos dois foi elaborado em conjunto e seguiu cinco princípios básicos apresentados no início dessa seção: (i) a definição do tipo de corpus (quadro 2), (ii) a escolha de textos e registros, (iii) a escolha das fontes, (iv) a organização dos textos e por último (v) a anotação do corpus. Todos os passos são apresentados e exemplificados a seguir. Conforme mencionado na seção anterior, o primeiro passo para a compilação de um *corpus* seria a definição do **tipo de *corpus*** que se pretender compilar. A lista do quadro abaixo mostra uma compilação dos tipos de *corpora*.

Quadro 2 – Tipos de Corpus



A lista do quadro 2 não é hierárquica. Ela apenas elenca os tipos de *corpus* mais conhecidos. Um *corpus* pode ser **falado** ou **escrito**. O *corpus* falado é aquele composto por porções de falas que foram transcritas e preparadas para serem armazenas em um banco de dados, em nosso caso, as transcrições de falas das legendas do Reality TV shows que compõem o CARTS. O *Corpus* escrito é composto por textos escritos, que podem ser impressos ou não, mas devem ser digitalizados e como todos os *corpora*, armazenados em um banco de dados, o mesmo realizado com os currículos que compõem o CCC. Além de serem falados ou escritos, os corpora também podem ser considerados sincrônicos, diacrônicos ou históricos. Os corpora **sincrônicos** são formados por textos com uma linguagem contemporânea, os corpora **diacrônicos** retratam diferentes períodos de tempo e os corpora **históricos** retratam um tempo passado. Há ainda os corpora de amostragem e os monitores, os de **amostragem** são compostos por porções de textos ou variedades textuais finitas, eles são planejados para serem uma amostra finita da linguagem como um todo. Os corpora de amostragem são sempre **estáticos**, ou seja, depois de compilados, não é mais possível acrescentar e nem diminuir dados. Os opostos dos corpora de amostragem são chamados de corpora **monitores**, eles são compostos por dados constantemente reciclados de forma que possam refletir o estado atual de uma língua e por essa característica são também chamados de corpora **dinâmicos** ou corpora orgânicos.

Quando os corpora são compostos por um número semelhante de textos ou de dados, ou seja, se possuem uma quantidade de dados equilibrada, como por exemplo o mesmo número de textos ou o mesmo número de gêneros ou registros, eles recebem o nome de corpora **equilibrados**. Os corpora podem ser **especializados**, isto é, específicos de uma área ou um registro. Os corpora podem ser **regionais** ou dialetais, isso significa que seus textos são provenientes de uma ou mais variedades sociolinguísticas específicas. Os corpora **multilíngues** possuem textos de idiomas diferentes. Os corpora de **aprendiz** são formados por textos produzidos por estudantes ou não nativos. Os corpora **comparáveis** possuem textos do mesmo registro em dois idiomas diferentes, porém um texto não é a tradução do outro. Os corpora **paralelos** são também compostos por textos em dois idiomas diferentes, mas, diferente dos comparáveis, os textos são formados por traduções e podem ou não ser alinhadas. Quando os corpora são paralelos e alinhados isto quer dizer que uma linha do texto original fica alinhada com a linha do texto traduzido e pode-se ler o texto fonte e o texto traduzido um ao lado do outro ou um acima do outro. Por último temos os corpora de **treinamento** que também são chamados de corpora de testes, eles são desenhados para permitir o desenvolvimento de aplicações e ferramentas de análise linguística.

Uma vez definido o tipo de *corpus*, o planejamento cíclico nos leva para a escolha dos textos e registros. Essa parte da metodologia prevê a quantidade de textos que vamos compilar, ela leva em consideração a representatividade, a quantidade de registros que serão compilados. Em relação a quantidade de registros, o pesquisador pode compilar corpora gerais, que podem ser uma representação de alguma tendência histórica por meio de critérios baseados na produção de textos e não na recepção deles. Em outras palavras, os corpora podem ser compilados de forma oportuna ou de forma estatística. A oportunista é baseada na quantidade de textos disponíveis sobre o assunto que se quer representar e a compilação estatística é baseada na variação interna dos traços presentes nos textos. Ambos corpora descritos nesse artigo lançam mão da compilação estatística para a organização do banco de dados. Essa metodologia é proposta por Biber (1995) e foi escolhida pelos pesquisadores pelo desejo de construir *corpora* representativos, com o foco em aspectos de diversidade e balanceamento (Biber, Conrad and Reppen, 2009) e para dar seguimento às pesquisas usando os pressupostos da AMD.

Uma pesquisa que utiliza a AMD, segundo Berber Sardinha (2004, p. 306), envolve duas etapas distintas que ele chama de análises macroscópicas, ou seja, a computação dos fatores e as análises microscópicas, a interpretação desses fatores de modo funcional.

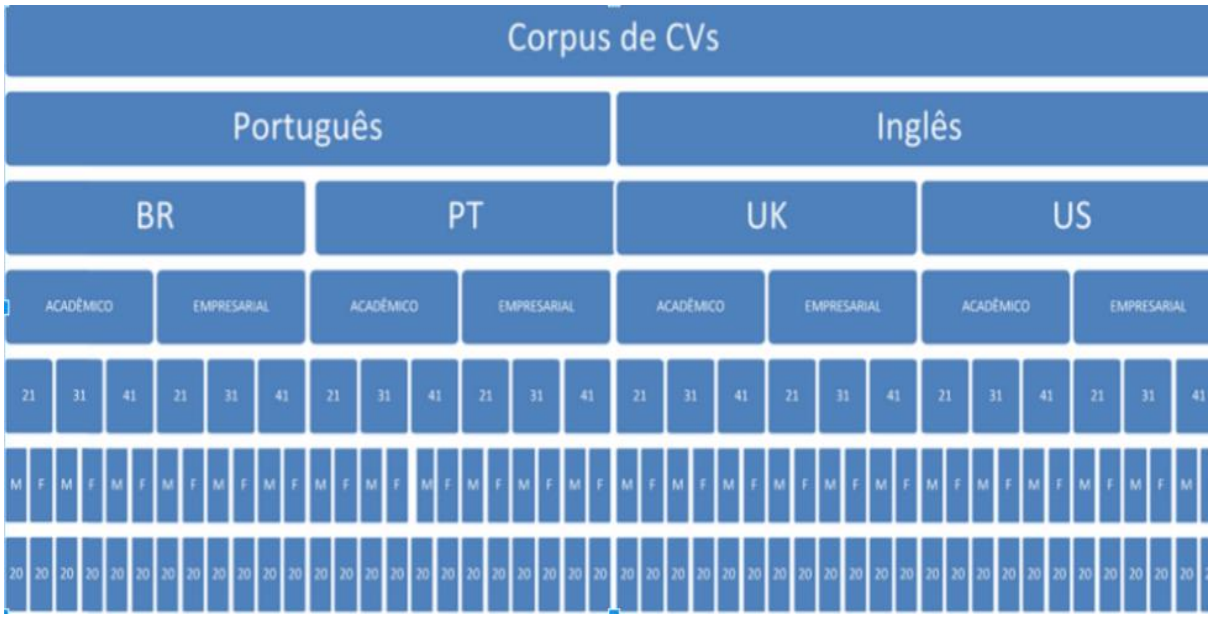
Segundo Biber (1993, p. 63), quando se trata da compilação de *corpora*, pode-se afirmar que o ponto nuclear de todo o processo é a compilação de um *corpus* que seja representativo e para isso, é preciso haver uma definição completa da população-alvo, ou seja, é preciso determinar o tipo de registro que pretendemos investigar e que sejam mais adequados para atender nossas questões de pesquisa e que seja também abrangente e representativo. É possível entender que a representatividade se refere ao quanto uma amostra representa as variações de uma população, ou seja, o objeto do estudo realizado, a língua de uma forma geral. Biber sugere algumas fórmulas e cálculos que podem ajudar a garantir essa tão almejada representatividade. Nesse trabalho procuramos seguir as sugestões do autor no que tange a compilação de um número mínimo de textos. Sendo assim, o para o CARTS foram coletados dois mil textos e o CCC possui dois mil CVs. O pesquisador que compila os corpora chega a representatividade usando uma formula que consegue determinar o tamanho mínimo exigido para que uma amostra possa representar uma população. O processo acontece a partir da determinação de uma variedade X de tipos de texto para uma determinada população e daí o pesquisador vai aproximando essa quantidade até o total estabelecimento da representatividade. Segundo Biber (2012, p. 13) a questão da definição da população é a primeira consideração no planejamento de um corpus. Ele recomenda ainda que o pesquisador escolha textos com boas bases de amostragem para que ela possa ser tanto probabilística, quanto aleatória. "Ao escolher e avaliar uma base de amostragem, aspectos de eficiência e de relação custo/benefício devem ser comparados com maiores graus de representatividade.

A metodologia proposta por Biber (1993) prevê que os textos com mais variação têm de ser mais numerosos. Para dar início a compilação, o pesquisador escolhe a quantidade de textos que deseja ter no corpus, por exemplo, 200 textos e o número de registros que se deseja incluir, por exemplo 3 (conversação, ficção e acadêmico). É preciso eleger a quantidade mínima inicial de textos igual para todos os registros, como por exemplo 20. Logo em seguida aplica-se a fórmula usando a média de variação por registro, que indica quantos dos textos restantes devem ir para cada registro. No caso, a questão é dividir os 140 textos restantes, ou seja, 200 menos os 60 já alocados, entre os 3 registros. No caso do exemplo usado por Biber, 61 textos foram alocados como conversação, 64 de ficção e 75 textos acadêmicos fecham o total de 200 textos. Os textos acadêmicos têm uma proporção maior do que os de ficção e conversação simplesmente porque exibem uma variação média maior. Caso a compilação de textos seja feita apenas de forma proporcional, cada um dos registros teria um terço do total do corpus, os seja 66 ou 67 textos.

No caso dos corpora CCC e CARTS, a compilação inicial foi feita de forma aleatória e proporcional, porque só havia um tipo de registro sendo compilado. No caso do CCC, o registro é o escrito, CVs e no caso do CARTS, o registro é *reality TV shows*. A compilação estatística acontece em uma fase secundária do planejamento cíclico, depois que o corpus piloto foi montado e a média de variação entre os textos foi feita. De uma forma geral o CCC possui um total de dois mil textos e CARTS possui um total de 780 textos que correspondem a 6.4 milhões de palavras (tokens) e 53.520 types.

O segundo passo do planejamento prevê a escolha dos textos e dos registros. No caso do CCC, a compilação inicial selecionou CVs em português do Brasil e em português Europeu, sendo a metade Cvs acadêmicos e a outra metade CVs das áreas empresariais, agrupados em três diferentes faixas etárias, até 21 anos, até 31 anos e acima de 41 anos, e também nos gêneros femininos e masculinos, totalizando 1.960 CVs, como mostra a tabela a seguir (quadro 3). De acordo com Sinclair (1991) as decisões que são feitas sobre o que vai estar presente no *corpus* e como a seleção será organizada, controla quase tudo o que acontece posteriormente. Portanto, é extremamente importante, que o pesquisador planeje a compilação de seu *corpus* com critérios claros. O CCC foi planejado para representar uma população de Curriculum Vitae produzidos originalmente em português e em inglês. A amostra inclui currículos acadêmicos armazenados na plataforma Lattes e também currículos empresariais armazenados na página do programa LinkedIn, uma das plataformas mais populares para o compartilhamento de CVs no mundo, além desse programa, a página da internet chamada Indeed também foi utilizada para a compilação de CVs autênticos em língua inglesa. Essa plataforma é internacional e abriga CVs dos mais variados tipos.

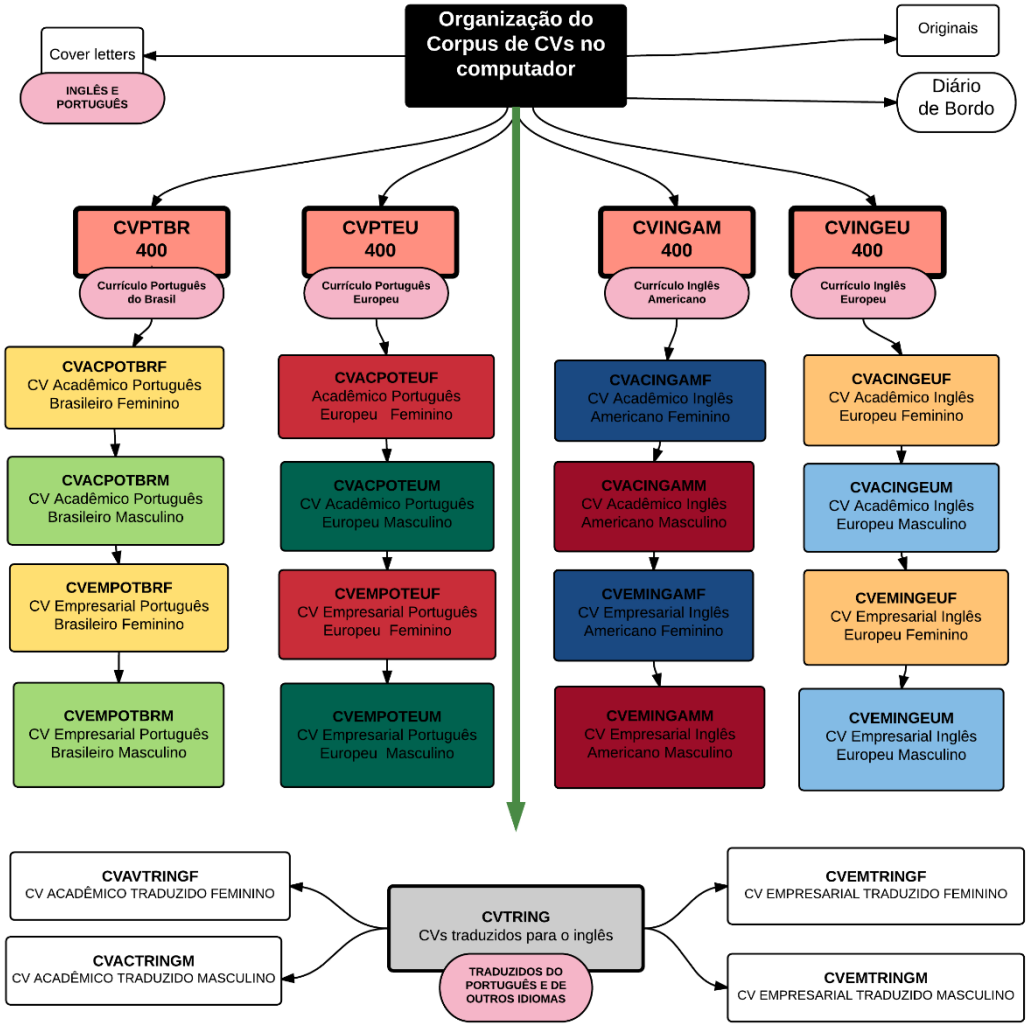
Quadro 3 – O primeiro desenho do CCC



Posteriormente, o planejamento apresentado no quadro 1 precisou sofrer alterações, porque durante a compilação percebeu-se que os autores dos CVs não incluem a idade nos textos e nem a data de nascimento, o que tornou impossível o uso da variável faixa etária. Percebeu-se também que era necessário registrar o processo de limpeza e compilação em um diário de bordo, pois as percepções e observações do momento de compilação e limpeza poderiam vir a ser importante depois de algum tempo. A ideia de registrar o processo de compilação em um diário de bordo mostrou-se muito relevante a partir do momento que no momento da limpeza, ou seja, no momento em que o nome, telefone e contato do dono do CV era retirado do texto, era possível fazer pequenas anotações sobre a origem dos CVs, como a área de formação, o gênero do autor e outros tipos de informação. Foi exatamente no momento dessas anotações no diário e durante a compilação e a limpeza que a variável faixa etária foi considerada como irrelevante e impossível de se concretizar. Outro ponto relevante refere-se à origem dos CVs. No caso dos CVs de língua portuguesa eles são compilados tanto em português do Brasil, quanto em português europeu e no caso dos CVs em língua Inglesa, a compilação foi feita a partir de CVs europeus e americanos. No entanto, com o passar do tempo e de acordo com as buscas e anotações no diário de bordo, percebeu-se que os CVs australianos também representam uma grande fatia das amostras dos CVs em língua inglesa e assim que iam sendo identificados eles foram agrupados em um único arquivo. Nessa fase da compilação e anotação do processo de compilação, ainda não é sabido se todos os CVs compilados serão utilizados na pesquisa, porém são compilados do mesmo jeito e fazem parte do corpus piloto. O mesmo aconteceu com os CVs traduzidos e com as cartas de apresentação (*cover letters*), mesmo não aparecendo no desenho do corpus piloto, os CVs traduzidos para o inglês foram armazenados em uma pasta (quadro 4) e poderão ser futuramente considerados como dados para a pesquisa e inicialmente fazem apenas parte do material piloto.

A limpeza dos CVs é primordial, uma vez que os nomes, e-mails e telefones não precisam aparecer nos textos e podem até comprometer a pesquisa, já que a identificação compromete a disponibilidade dos CVs por parte dos RHs das empresas cedentes.

Quadro 4 – O segundo desenho do CCC



O quadro 4 apresenta a organização do CCC dentro do computador. Cada um dos retângulos representa uma pasta do computador e representa os CVs que estão armazenados dentro das pastas.

No caso do CARTS, inicialmente também foi realizado a coleta de arquivos com transcrição das legendas de programas do tipo *reality TV shows* indicados e premiados na televisão americana, a grande maioria eram do tipo competição. Nas discussões junto ao GELC (Grupo de Estudos em Linguística de Corpus) na PUC-SP, percebeu-se que os programas não possuíam uma diversidade temática para que o *corpus* fosse reconhecido como representativo do gênero de televisão e, portanto, havia a necessidade de repensar seu design.

Quadro 5 – O CARTS em detalhes

CARTS - CORPUS OF AMERICAN REALITY TV SHOWS			
CATEGORIES	SUBCATEGORIES	TV SHOWS	TEXTS
1. COMPETITION	1.1 CELEBRITY COMPETITION	1.1.1 CELEBRITY APPRENTICE	30
		1.1.2 DANCING WITH THE STARS	30
	1.2 DATING	1.2.1 THE BACHELOR	12
		1.2.2 THE BACHELOR PAD	12
		1.2.3 THE BACHELORETTE	18
		1.2.4 THE MILLIONAIRE MATCHMAKER	18
	1.3 GAMEDOC	1.3.1 BIG BOTHER US	20
		1.3.2 SURVIVOR	20
		1.3.3 THE AMAZING RACE	20
2. DOCUMENTARY	2.1 DOCUSOAP/DOCUDRAMA	2.1.1 REAL L WORLD OF LOS ANGELES	20
		2.1.2 THE REAL HOUSEWIVES OF ATLANTA	14
		2.1.3 THE REAL HOUSEWIVES OF NYC	12
		2.1.4 THE REAL HOUSEWIVES OF OC	14
	2.2 HIDDEN & LIVE-IN CAMERA	2.2.1 BETTY WHITE'S OFF THE ROCKERS	20
		2.2.1 COPS	20
		2.2.3 DEALIEST CATCH	20
	2.3 REALITY SITCOM	2.3.1 JERSEY SHORE	30
		2.3.2 KEEPIN UP WITH THE KARDASHIANS	30
3. TALENT E PROFESSION	3.1 BUSINESS	3.1.1 MILLION DOLLAR LISTING LOS ANGELES	10
		3.1.2 MILLION DOLLAR LISTING NEW YORK CITY	10
		3.1.3 THE APPRENTICE	20
		3.1.4 SHARK TANK	20
	3.2 COOKING	3.2.1 CAKE BOSS	10
		3.2.2 HELL'S KITCHEN US	25
		3.2.3 MASTER CHEF US	25
	3.3 FASHION	3.3.1 AMERICA'S NEXT TOP MODEL	30
		3.3.2 PROJECT RUNAWAY	30
	3.4 PERFOMING	3.4.1 AMERICAN IDOL	10
		3.4.2 AMERICA'S GOT TALENT	15
		3.4.3 SO YOU THINK YOU CAN DANCE	15
		3.4.4 THE VOICE US	10
		3.4.5 THE X-FACTOR	10
4. TRANSFORMATION	4.1 EXPERT GUIDANCE	4.1.1 SAY YES TO THE DRESS	20
		4.1.2 SUPERNANNY US	20
		4.1.3 CATFISH.TV.SHOW	20
	4.2 SELF-IMPORVEMENT AND RENOVATION	4.2.1 EXTREME MAKEOVER: HOME EDITION	30
		4.2.2 THE BIGGEST LOSER	30
	4.3 SOCIAL EXPERIMENT	4.3.1 NAKED AND AFRAID	30
		4.3.2 UNDERCOVER BOSS	30
		TOTAL	780

O segundo passo do planejamento se deu por meio da definição de um desenho de *corpus* que refletisse uma diversidade maior deste dos Reality TV shows, diante da dificuldade de reconhecer uma taxonomia única para a categorização destes programas, uma vez que não há uma concordância nas formas de categorizá-los nas emissoras de televisão, em websites sobre notícias para fãs e em premiações realizada por instituições renomadas do gênero televisivo, bem como o *Emmy Awards* da *Academy of Television Arts & Science* e *People´s Choice Awards* produzido pela CBS. Houve, então, a necessidade de estudar a história deste gênero da televisão e entender melhor os temas e formatos mais recorrentes nos dias atuais para a elaboração de uma taxonomia própria para o estudo e a construção de um desenho de corpus mais coerente, diversificado e equilibrado.

Composição do CARTS

Quadro 6 – A organização do CARTS

Categorias e Subcategorias	Programas	Textos
1 Competição		
1.1 Celebridades	2	60
1.2 Namoro	4	60

1.3 Jogo Documentário	3	60
2 Documentário		
2.1 Novela Documentário	4	60
2.2 Câmeras Escondidas	3	60
2.3 Reality Sitcom	2	60
3 Talento e Profissão		
3.1 Business	4	60
3.2 Cozinha	4	60
3.3 Moda	2	60
3.4 Performance de Palco	4	60
4 Transformação		
4.1 Orientação de Especialistas	3	60
4.2 Melhoria Pessoal	2	60
4.3 Experiência Social	2	60
Total	39	780

O terceiro passo da compilação se preocupa com a escolha das fontes dos textos. Ambos CCC e CARTS foram compilados a partir de textos disponíveis na Internet. As vantagens da compilação com fonte na internet é que os textos já estão digitalizados O CCC usou sites como o LinkedIn e o CARTS utilizou três websites: springfieldsprindfield.co.uk, tv.ark.com e tvsubs.net.

O quarto estágio da compilação trata da organização dos corpora no computador. Ela lida com a organização dos arquivos e dos diretórios no computador. Quando os textos são compilados e armazenados de forma fragmentada, chamamos a organização de fragmentada e quando os textos são compilados e armazenados em um único arquivo dizemos que a organização é unitária. No caso de ambos corpora, o CCC e o CARTS, a organização inicial dos arquivos foi fragmentada, ou seja, cada arquivo de texto foi nomeado e agrupado de acordo com sua subcategoria e inserido em uma pasta, que por sua vez estava inserida em outras pastas das categorias principais dentro da pasta do *corpus*.

Os diretórios podem ser classificados como tendo uma estrutura plana ou uma estrutura hierárquica. A estrutura plana tem uma visualização mais simples e direta, porém fica mais complicado incluir ou excluir um novo arquivo, uma vez que todos estão juntos. No caso da estrutura hierárquica dos arquivos, há uma maior flexibilidade para a inclusão ou exclusão de arquivos, contudo, a manutenção exige muito mais consistência.

O quinto e último estágio do planejamento e da compilação dos *corpora* trata da anotação dos *corpora*. A anotação pode ser feita como uma marcação, muitas vezes aparece na forma de cabeçalho e também na forma de etiquetagem, como no caso das classes gramaticais que aparecem como etiquetas junto às palavras do corpus.

A marcação pode ser feita como um cabeçalho, no caso do CCC, o cabeçalho que introduz o corpus é formado pela autoria, a fonte e as datas em que os CVs foram compilados. No caso do CARTS, o cabeçalho apresenta os mesmos itens extralinguísticos listados no CCC e na segunda fase, após a seleção final dos arquivos, a organização deixou de ser hierárquica para tornar-se plano, ou seja, todos os arquivos ficaram armazenados em uma única pasta nomeada *carta_full*, com o intuito de facilitar o uso das ferramentas de análise linguísticas, bem como os etiquetadores, concordanciadores e listadores.

Para Berber Sardinha (2004), o pesquisador pode lançar mão de quatro tipos diferentes de anotação linguística, a primeira é a morfossintática, ou seja, aquela que cuida da marcação das partes do discurso. A segunda é a marcação sintática (*parsing*), a terceira é a marcação semântica e a quarta e última é a marcação discursiva. Essas marcações são realizadas por programas chamados etiquetadores (*taggers*). Eles adicionam uma etiqueta, ou seja, um código a cada uma das palavras dos corpora, esse código indica qual e a parte do discurso que a palavra etiquetada indica. Sendo assim, a palavra currículo que é parte do CCC recebe a etiqueta de substantivo.

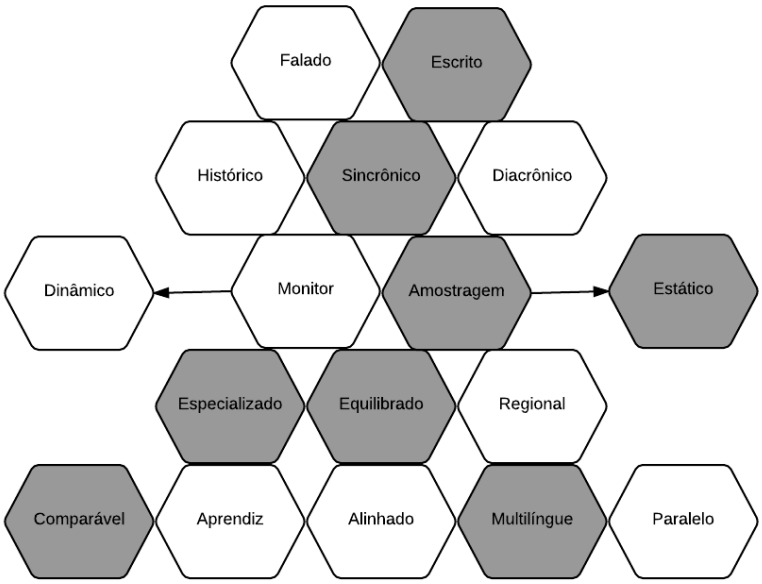
No caso da etiquetagem, ela pode ser de vários tipos, o mais comum lida com as classes gramaticais, são as marcações morfossintáticas, como substantivos, verbos, adjetivos e etc. de cada uma das palavras dos corpora. Contudo, a etiquetagem ainda pode ser sintática, semântica, discursiva, lematizada ou de tradução, como já mencionado anteriormente. No caso da marcação sintática, denominamos o programa de *parsers*. Eles adicionam códigos, ou seja, etiquetas que contém informações que identificam estruturas sintáticas, como os sintagmas verbais e nominas, por exemplo. Ambas ferramentas podem tanto usar princípios baseados em regras quanto em probabilidade. No caso das regras, as ferramentas inserem etiquetas de acordo com as regras gramaticais (etiquetadores) e de acordo com as concepções de língua (*parsers*). Já no caso da probabilidade, as etiquetas são escolhidas de acordo com a probabilidade estatística que uma etiqueta tem de ocorrer em um determinado contexto (ZUPPARDO, 2013, p. 34)

Ambos CCC e CARTS usam dois programas de etiquetam diferentes para fazer as marcações dos corpora. O primeiro programa utilizado é o *Biber Tagger* e o segundo é o USAS, desenvolvido na universidade de Lancaster. *O Biber Tagger* funciona em dois níveis, o primeiro cuida da identificação inicial da categoria gramatical de cada palavra e o segundo cuida da solução de ambiguidades nos casos em que o dicionário apresenta mais de uma possibilidade de categoria gramatical para uma palavra (ZUPPARDO, 2013, p. 35). O *Biber Tagger* foi escolhido como etiquetador dos dois corpora, CCC e CARTS por dois motivos: o primeiro baseia-se na eficácia da ferramenta, que possuiu uma margem de erro inferior à 10% e o segundo deve-se ao fato dos dois corpora serem usados para investigações posteriores usando AMD. O USAS funciona como um etiquetador semântico e funciona em vários idiomas. Ele é gratuito e está disponível na página http://ucrel.lancs.ac.uk/usas/. A marcação dos corpora é a última etapa do planejamento e design da montagem de um corpus, essa etapa encerra a compilação cíclica, marca o fim da compilação e design e o início da etapa de investigação e análise.

4. Considerações Finais

Os dois corpora descritos nesse artigo possuem tipos diferentes de dados. Se observarmos o quadro 3, veremos que o CCC é um corpus formado por textos escritos, no caso CVs, originais e autênticos compilados de fontes diversas e disponibilizados por empresas e indivíduos. Todos os CVs são amostras autênticas e não sofreram nenhum tipo de manipulação. Não há nenhum tipo de julgamento de valor a respeito da qualidade ou não dos textos produzidos nos CVs, não houve escolha e nem critérios que levasse o conteúdo intelectual dos textos e seus produtores em consideração, ou seja, foram considerados CVs curtos e longos, aprovados ou não pelas empresas, porém, todos autênticos. Quando classificamos o corpus como sendo falado ou escrito, queremos apenas dizer que um corpus falado é composto por porções de transcrições de fala, uma vez que estes textos foram transcritos a partir de falas autênticas (discursos espontâneos, entrevistas, etc.) ou foram criados para serem falados (scripts, roteiros, discursos planejados, etc.). Em contrapartida, um corpus escrito, nesse caso o CCC, é composto por porções autênticas de textos escritos elaborados para serem lidos.

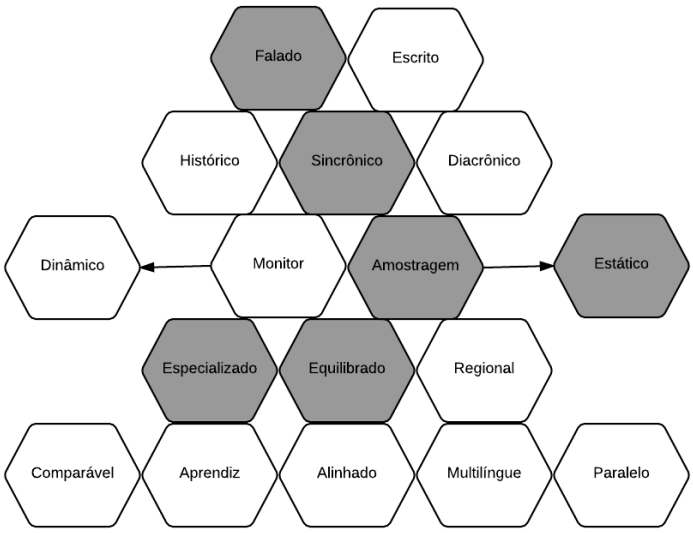
Quadro 7 – Características do CCC



O CARTS é um corpus é composto pelas legendas dos Reality TV Shows, sendo assim é um corpus falado e sincrônico, uma vez que as legendas foram extraídas de programas contemporâneos e representam apenas uma amostragem estática e especializada e equilibrado de um único tipo de registro de linguagem de televisão norte-americana.

Ao relatar um pouco da experiência adquirida pelos dois pesquisadores nas tarefas de planejamento, compilação, organização, replanejamento e reorganização seus *corpora* na tentativa de construir um desenho de *corpus* coerente e adequado para atender os objetivos e questões de suas pesquisas, este trabalho pode contribuir para que futuros pesquisadores conheçam aos elementos e etapas essenciais no percurso teórico-metodológico da construção de corpora e promover a reflexão sobre a importância de uma base teórica consistente para evitar problemas que muitas vezes emergem apenas no momento da análise de dados.

Quadro 8 – Características do CARTS



Ao relatar um pouco da experiência adquirida pelos dois pesquisadores nas tarefas de planejamento, compilação, organização, replanejamento e reorganização seus *corpora* na tentativa de construir um desenho de *corpus* coerente e adequado para atender os objetivos e questões de suas pesquisas, este trabalho pode contribuir para que futuros pesquisadores conheçam aos elementos e etapas essenciais no percurso teórico-metodológico da construção de corpora e promover a reflexão sobre a importância de uma base teórica consistente para evitar problemas que muitas vezes emergem apenas no momento da análise de dados.

Referências

BERBER SARDINHA, Tony e Telma SÃO BENTO FERREIRA. *Working with Portuguese Corpora*. London: Bloomsbury Academic, 2014.

BERBER SARDINHA, Tony. *Linguística de Corpus*. Sao Paulo: Manole, 2004.

BERBER SARDINHA, Tony, Carlos KAUFFMAN e Cristina MAYER-ACUNZO. "A Multi-dimensional analysis of register variation in Brazilian Portuguese." *Corpora* (2014): 239-271.

BIBER, Douglas. "Methodological issues regarding corpus-based analyses of linguistic variation." *Literary and Lingusitc Computing*. Vol. 5. 1990. 257-269.

BIBER, Douglas. "Representativeness in corpus design." *Literacy and Linguistics Computing 8*. 1993. 243-257.

—. "Representatividade em Planejamento de Corpus." *Cadernos de Tradução* JAN-JUN de 2012: 11-46.

—. *Variation Across Speech and Writing*. Cambridge: Cambridge University Press, 1988.

BIBER, Douglas, Susan CONRAD e Randi REPPEN. *Corpus Linguistics: Invstigating Language Structure and Use*. 2nd . Cambridge: Cambridge University Press, 2000.

BRUM-DE-PAULA, M. R. e G. S. ESPINAR. "Coleta, descrição e análise de produções orais." *Letras, Santa Maria* 2002: 69-84.

CROWDY, S. "Spoken Corpus Design." 1993: 259-265.

CROWDY, S. "Spoken Corpus Transcription." *Literary and Linguistic Computing* 1994: 25-28.

HUNSTON, Susan. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press, 2002.

McCarthy, M.J. *Spoken Language and Applied Linguistics*. Cambridge: Cambridge University Press, 1998.

McENERY, T. e A. W. *Corpus Linguistics*. Edinburgh: Edinburgh University Press, 2001.

MILLER, Donald e Douglas BIBER. "Evaluating reliability in quantitative vocabulary studies: The influence of corpus design and composition." *International Journal of Corpus Linguistics* 2015: 30-53.

RAJAGOPALAN, Kanavillil. "Repensar o Papel da Lingüística Aplicada." MOITA LOPES, Luiz Paulo. *Por uma Lingüística Aplicada Indisciplinar*. São Paulo: Parabola, 2006.

RESENDE, Simone Vieira. "Normas tradutórias: o caso dos artigos científicos e suas condicionantes culturais." *Dissertação (Mestrado em Linguística)*. Rio de Janeiro: Universidade do Estado do Rio de Janeiro - UERJ, 2011.

SINCLAIR, John. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press, 1991.

ZUPPARDO, Maria Carolina. " A linguagem da aviação: um estudo de manuais aeronáuticos baseado na Análise Multidimensional." *Revista Virtual de Estudos da Linguagem ReVEL* 11 (21) (2013).