UNIVERSIDADE FEDERAL DE JUIZ DE FORA
FACULDADE DE LETRAS
PROGRAMA DE PÓS-GRADUAÇÃO EM LINGUÍSTICA

TATIANE SILVA TAVARES

REQUISITOS PARA A MODELAGEM DE PADRÕES DE CUNHAGEM E CONSTRUÇÕES SEMI-PRODUTIVAS NO CONSTRUCTICON DA FRAMENET BRASIL COM FOCO NO FOMENTO AO DESENVOLVIMENTO DE TRADUTORES AUTOMÁTICOS

Juiz de Fora
2018

TATIANE SILVA TAVARES

REQUISITOS PARA A MODELAGEM DE PADRÕES DE CUNHAGEM E CONSTRUÇÕES SEMI-PRODUTIVAS NO CONSTRUCTICON DA FRAMENET BRASIL COM FOCO NO FOMENTO AO DESENVOLVIMENTO DE TRADUTORES AUTOMÁTICOS

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Linguística da Faculdade de Letras da Universidade Federal de Juiz de Fora, como requisito parcial para obtenção do título de Doutora em Linguística.

ORIENTADOR: Prof. Dr. Tiago Timponi Torrent

Juiz de Fora
2018

TATIANE SILVA TAVARES

REQUISITOS PARA A MODELAGEM DE PADRÕES DE CUNHAGEM E CONSTRUÇÕES SEMI-PRODUTIVAS NO CONSTRUCTICON DA FRAMENET BRASIL COM FOCO NO FOMENTO AO DESENVOLVIMENTO DE TRADUTORES AUTOMÁTICOS

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Linguística da Faculdade de Letras da Universidade Federal de Juiz de Fora, como requisito parcial para obtenção do título de Doutora em Linguística.

Presidente, Prof. Dr. Tiago Timponi Torrent – UFJF

Profa. Dra. Lilian Vieira Ferrari – UFRJ

Profa. Dra. Karen Sampaio Braga Alonso – UFRJ

Prof. Dr. Luiz Fernando Matos Rocha – UFJF

Prof. Dr. Ely Edison da Silva Matos – UFJF

Juiz de Fora

Setembro de 2018

AGRADECIMENTOS

Ao professor Tiago que a cada orientação ajudou a construir em mim a confiança necessária para seguir em frente. Agradeço por todo conhecimento compartilhado e por ter tornado tão leve todo o percurso até aqui.

À professora Thais Sampaio por ter me apresentado ao projeto, pelo qual desde o início tive muita simpatia.

Aos professores Luiz Fernando e Lilian Ferrari pelas preciosas contribuições na ocasião da minha Qualificação.

Ao meu orientador no Texas, Hans Boas, que me recebeu de maneira tão amistosa e que se preocupou em tornar meu período de doutorado sanduíche o mais proveitoso possível.

Aos colegas do projeto FrameNet pela ajuda mútua e pelo clima agradável e descontraído durante as reuniões do grupo.

À Capes pelo apoio concedido à realização do doutorado sanduíche.

Agradeço aos amigos que sempre me apoiaram nessa caminhada, especialmente a Milena Lepsch, parceira querida de estudos, de afinidades, de conversas sem fim.

Ao William pelo cuidado, por acreditar tanto em mim e nutrir meus sonhos.

A minha filha Sofia por me apoiar verdadeiramente em qualquer coisa que eu faça e por ser sempre tão generosa e compreensível.

RESUMO

Esta pesquisa tem por objetivo investigar as contribuições que a FrameNet e o Constructicon podem oferecer aos sistemas de Tradução por Máquina (TM) ao revisitar a literatura sobre Gramática das Construções e Padrões de Cunhagem (KAY, 2013). A hipótese é a de que utilização da base de dados de uma FrameNet, que oferece representações computacionais das estruturas cognitivas essenciais na construção do sentido (os *frames*), e de um Constructicon, o qual integra a informação sobre a gramática de uma língua, podem auxiliar o processamento de línguas naturais pelo computador e, consequentemente, auxiliar os sistemas de Tradução por Máquina. O segundo recorte deste trabalho refere-se à revisão da Gramática das Construções, especialmente em relação ao tratamento de padrões de cunhagem. Segundo a abordagem de Kay (2005, 2013), deve-se considerar como construção apenas a quantidade mínima de informação que o falante precisa ter para que seja capaz de entender e produzir sentenças da língua. Nesta perspectiva, assume-se que as construções de uma língua sejam estes padrões mais gerais e produtivos, os quais licenciam as mais diversas sentenças compreensíveis pelo falante. Por isso buscamos reanalisar a estrutura de quantificação indefinida: *i) mar de gente, ii) oceano de calúnias, iii) enxurrada de notícias,* já investigada em dissertação de mestrado (TAVARES, 2014), a fim de que se possa discutir a validade da abordagem de Kay para a modelagem computacional do padrão de quantificação. A análise dos dados é guiada pelos pressupostos da Semântica de Frames (FILLLMORE, 1982,1985; PETRUCK, 1996) e pela metodologia da FrameNet, a qual emprega a análise semântica e sintática dos objetos de investigação, pois, para as tarefas de Processamento de Línguas Naturais, é necessário que se leve em conta as regularidades da língua para que esta seja processada computacionalmente. A conclusão da análise aponta para o fato de que a modelagem da estrutura de quantificação indefinida como uma rede de padrões de cunhagem no Constructicon da FrameNet Brasil, definidos a partir de restrições *soft*, pode trazer um ganho de generalidade na análise e a possibilidade de que estruturas inovadoras, cunhadas por analogia, sejam igualmente reconhecíveis pelo modelo computacional resultante, o que representa um avanço em relação aos processos tradicionalmente empregados na hibridização de sistemas de tradução por máquina.

Palavras-chave: construções; padrões de cunhagem; modelagem construcional; tradução por máquina

ABSTRACT

This work aims to investigate the contributions that the FrameNet and the Constructicon could offer to Machine Translation (MT) systems by revisiting the literature about Construction Grammar and Coinage Patterns (KAY, 2013). The hypothesis is that by using the FrameNet database, which offers computational representations of essential cognitive structures formeaning construction (frames), and a Constructicon database, which integrates information about the grammar of a language, it is possible to support the natural language processing by the computer and, consequently, to assist Machine Translation systems. The second point of this work refers to the review of Construction Grammar, especially the treatment of coinage patterns. According to Kay's approach (2005, 2013), it must be considered a construction only the minimal quantity of information that a speaker needs to be able to understand and to produce sentences of the language. From this perspective, it is assumed that constructions of a language are these more general and productive patterns, which license many comprehensible sentences by the speaker. For this reason we seek to reanalyze the indefinite quantity structure: *i) mar de gente, ii) oceano de calúnias, iii) enxurrada de notícias,* already investigated in a dissertation (TAVARES, 2014), in order to discuss the validity of Kay's approach to the computational modeling of the quantification pattern. The analyses is guided by Frame Semantics principles (FILLLMORE, 1982,1985; PETRUCK, 1996) and by FrameNet methodology, which applies a semantic and syntactic analyses to the objects of this research, because dealing with Natural Language Processing demands an account for the language regularities so it can be computationally processed. The conclusion of the analysis points to the fact that the modeling of the indefinite quantification structure as a network of coinage patterns in the FrameNet Brasil Constructicon, defined from soft constraints, can bring about a gain of generality in the analysis and the possibility that innovative structures, coined by analogy, are equally recognizable by the resulting computational model, which represents an advance over the processes traditionally employed in the hybridization of machine translation systems.

Keywords: constructions; coinage patterns; constructional modeling; machine translation

SUMÁRIO

INTRODUÇÃO

As tarefas de Processamento de Línguas Naturais vêm buscando sanar os problemas que envolvem a aprendizagem de línguas por máquina e algum avanço foi alcançado no que se refere à tradução automática ou Tradução por Máquina (TM). Mas, o que temos ainda hoje neste quesito são resultados razoáveis. Resultados razoáveis podem ser úteis para alguns propósitos, há algumas aplicações para a TM que não exigem, de fato, traduções exatas. Mas tais resultados são ainda sim ineficientes, do ponto de vista das pesquisas que envolvem PLN.

Não é só de regras que se compõe todo o conhecimento sobre uma língua. Por que então esperar que um computador seja capaz de processar adequadamente uma língua se apenas o alimentamos com um amontoado de regras linguísticas? E se aliarmos a este modelo de regras um modelo estatístico? Neste caso, teríamos uma combinação interessante, pois além de operar com regras, a máquina seria capaz de indicar os padrões com ocorrência mais frequente na língua. Mas ainda sim faltaria um componente fundamental que dá suporte a toda a construção de sentido, quais sejam as estruturas cognitivas que emergem do contexto sociocultural dos falantes.

Estas estruturas fundamentais no processamento da língua e inerentemente humanas são possíveis de serem modeladas nas máquinas. A FrameNet é um recurso lexicográfico que busca representar de modo computacional os frames, para que se possa descrever os significados dos itens lexicais. E o Constructicon, por sua vez, busca representar computacionalmente as estruturas da língua que não podem ser tratadas de modo lexicográfico. Por que então não fazer com que as máquinas aprendam sobre frames, através da base de dados da FrameNet? E por que não oferecer modelos de língua baseado em regras mais sofisticados, que tomem as construções como ponto de partida, por meio da utilização de um Constructicon? Ainda, é preciso que se coloque em discussão uma questão teórica: o que deve ser de fato tomado como construção? E qual será a implicação disso para o tratamento computacional, que, por vezes, tem objetivos distintos de uma gramática das construções?

Os problemas de pesquisa levantados até aqui nos levam à proposição da seguinte hipótese:

Padrões de cunhagem e construções semi-produtivas são satisfatoriamente modelados linguístico-computacionalmente em recursos baseados em frames e construções.

Para testar a hipótese, remetemo-nos, primeiramente, à discussão dos limites do conceito de construção. Nesse contexto, a noção de Padrão de Cunhagem, apresentada por Kay (2005, 2013), vem confrontar a hipótese de que todo item linguístico seja igualmente uma construção da língua. Nosso propósito é verificar através de expressões bastante recorrentes do português, licenciadas pelo padrão Binominal de Quantificação Indefinida, os desdobramentos dessa discussão para a proposição de modelos computacionais que sejam capazes, ao mesmo tempo, de permitir o licenciamento de expressões de quantificação indefinida cristalizadas na língua – como *um caminhão de problemas* –, de indicar a estranheza daquelas que não são usuais – como *uma bicicleta de problemas* –, mas ainda de indicar a possibilidade de casos criativos – como *um navio de problemas*. Tais modelos poderiam então ser utilizados em substituição a listas fechadas de injeção terminológica, ou a regras fixas de conversão de expressões.

Para apresentar a tese de que modelos computacionais inspirados cognitivamente podem ser soluções melhores a listas de regras e termos, este texto se organiza da seguinte forma: os capítulos 1, 2 e 3 abordam os pressupostos teóricos que fundamentam as discussões acerca do Processamento de Línguas Naturais, em específico, da Tradução por Máquina, e da modelagem computacional no âmbito do Constructicon da FrameNet Brasil; o capítulo 4 expõe a metodologia de trabalho empregada; enquanto o capítulo 5 apresenta as análises referentes à remodelagem da Construção Binominal de Quantificação Indefinida; finalmente, o capítulo 6 trata das implicações desta remodelagem para os sistemas automáticos de tradução.

1 LINGUAGEM, TECNOLOGIA E TRADUÇÃO POR MÁQUINA

1.1 Computação e Linguística

Lidar com toda a complexidade da linguagem humana pode ser desafiador até mesmo para estudiosos da língua. O que dizer então sobre o processamento computacional da linguagem? Segundo Matos (2014), linguagem e computação parecem caminhar para lados opostos, tendo em vista que a primeira é vaga, ambígua e muitas vezes de difícil regularização, enquanto a segunda busca justamente por regras, limites e lógicas formais. Mas muito esforço foi empreendido no sentido de encontrar meios de processar a linguagem computacionalmente e, nesse contexto, a abordagem Gerativista serviu como uma teoria que viabilizava tratar a linguagem em termos formais (MATOS, 2014, p.19).

Jones (2007) afirma que, de modo geral, a Linguística Computacional teve seu desenvolvimento independente das principais correntes linguísticas, como podemos observar pelo breve panorama histórico apresentado pela autora. No início dos anos 50, desenvolveram-se os primeiros trabalhos em Linguística Computacional, mais especificamente em Tradução por Máquina, os quais tiveram como principais desafios o desenvolvimento de algoritmos para o *parsing* sintático e a formulação de estruturas de dados apropriadas para dicionários e gramáticas (JONES, 2007, p.438). Nos anos 60, os estudos nessa área avançaram significativamente e foi também nessa época que linguistas computacionais e linguistas gerativistas compartilhavam a crença na importância de um aparato formal capaz de parear sequências de palavras a descrições de estrutura gramatical. Já a partir dos anos 70, a Linguística Computacional teve seu crescimento de modo independente de correntes linguísticas. Isso se tornou mais evidente nos anos 90, com o surgimento expressivo de pesquisas baseadas em *corpus* e com o desenvolvimento de aprendizagem por máquinas. Nesse ponto, a Linguística Computacional e, mais especificamente, as ideias probabilísticas apresentadas pelas pesquisas baseadas em *corpus*, ganham relevo, oferecendo contribuições consideráveis aos modelos de língua.

De fato, muitas são as contribuições dos modelos probabilísticos atuais e não se pode negar o desenvolvimento considerável do Processamento de Língua Natural (PLN) e suas aplicações. Mas uma reaproximação com a Linguística ofereceria uma fonte muito maior de recursos que preencheria diversas lacunas no processamento das línguas. Muito se desenvolveu em Linguística, a Semântica ganhou espaço e teorias linguísticas desenvolveram aplicações computacionais, como a FrameNet, capazes de modelar estruturas de conhecimento altamente

complexas (os *frame*s). Por isso, não se justifica tal distanciamento e muito menos a utilização de recursos linguísticos tão limitados, focados apenas na sintaxe.

As próximas seções deste capítulo se ocupam, portanto, de apresentar questões relativas ao PLN e aos modelos de língua estatísticos, para que, no próximo capítulo, discutamos de que modo as teorias linguísticas de viés cognitivista podem contribuir para a Linguística Computacional.

1.2 Processamento de Línguas Naturais

O objetivo maior do PLN é o de viabilizar a compreensão e a produção de línguas naturais por uma máquina, tornando mais harmoniosa a relação entre usuário e sistemas computacionais.

Embora em suas raízes haja alguma herança gerativista, esta subárea da Inteligência Artificial ampara-se em modelos de língua estatísticos. Segundo Manning & Schutze (1999), o posicionamento deste modelo é diferente daquele que assume uma abordagem chomskyana, na qual se busca descrever a gramática subjacente à língua e que considera gramaticais apenas as sentenças licenciadas por essa gramática inata. Os pesquisadores ponderam que não é relevante que se faça distinção entre o que é considerado gramatical e agramatical, pois seu principal propósito é descobrir quais são os padrões mais comuns na língua, usando para isso a estatística e assumindo, portanto, uma abordagem empirista para a linguagem. Além disso, afirmam que sentenças podem ser sintaticamente gramaticais, mas anômalas do ponto de vista da convencionalidade, ou seja, um falante não-nativo de inglês pode produzir sentenças gramaticais, mas inadequadas por serem pouco usuais na língua, e isto teria a ver com a frequência de uso de tais sentenças e não com gramaticalidade.

Ainda de acordo com Manning & Schutze (1999, p. 15), há um forte argumento a favor de se tratar a probabilidade como parte de um entendimento científico da linguagem: o de que a cognição humana é probabilística e que, portanto, a linguagem, como parte integrante da cognição, também o seria. Além disso, não é possível que se tenha acesso a todos os dados e que se construa um conjunto completo de regras linguísticas, daí a importância da probabilidade. Com relação à cognição, os autores ilustram alguns exemplos em que o indivíduo utiliza a probabilidade para interagir com mundo, já que este apresenta muitas vezes informações incompletas. Para atravessar um rio, por exemplo, o indivíduo precisa fazer uma série de cálculos para que sua travessia seja efetuada com sucesso, ou seja, ele precisa integrar todas as informações que possui, seja a partir da visualização do rio, do conhecimento sobre a existência

de animais perigosos ou de uma forte corrente, dentre outras informações. De acordo com os autores, tais cálculos também são feitos quando se envolve a linguagem. Se, em vez de visualizar o rio, alguém lhe der alguma informação sobre o mesmo, como: “a água vai até o joelho se você andar em direção àquela árvore grande logo ali”, o indivíduo processaria tal informação levando em conta as palavras que compõe a sentença e o seu sentido geral e avaliaria se sua travessia seria segura. É nesse sentido que consideram a relação entre cognição, linguagem e probabilidade, e que afirmam que uma teoria probabilística possa ser parte central de uma teoria de linguagem.

Os pesquisadores também afirmam que os atuais modelos probabilísticos de língua – diferentes dos primeiros modelos dos anos 50, considerados bastante simplistas – são capazes de lidar com a complexidade das línguas humanas. Além disso, eles estariam aptos para explicar fenômenos que envolvam incerteza e incompletude, fenômenos próprios da cognição e da própria linguagem. Contudo, veremos nos próximos capítulos que nem todos os aspectos das línguas naturais podem ser recobertos apenas por métodos probabilísticos.

O tratamento da semântica por abordagens probabilísticas sempre foi visto com desconfiança, pois pouco se acreditava que a estatística seria capaz de lidar com um aspecto tão complexo como o significado. Segundo Manning & Schutze (1999), lidar com o significado depende primeiramente de como este será definido. Se for tomado como expressões simbólicas de uma língua, é possível que se façam traduções automáticas a partir de sistemas de PLN estatísticos. Contudo, consideram que uma maneira mais natural de se pensar o significado das expressões com as quais lidamos seria defini-lo a partir de sua distribuição na língua, nos contextos nos quais são usados. Nesse sentido, os pesquisadores assumem uma abordagem do significado a partir do uso.

Ainda com relação ao significado, nosso posicionamento é um pouco diferente. A Semântica de *Frames* traz respostas ao estudo da significação, tanto no âmbito teórico quanto lexicográfico. Sua principal premissa é a de que os significados são relativizados às cenas (FILLMORE, 1977) e, por isso, uma aproximação desta teoria com a abordagem probabilística pode ser interessante para resolver alguns problemas relacionados ao tratamento da semântica em tarefas de PLN. Isto, pois a FrameNet (FILLMORE ET AL., 2003), aplicação computacional da Semântica de *Frames*, também se baseia em sentenças extraídas do uso para realizar suas análises, de modo que o sentido das Unidades Lexicais é determinado a partir do *frame* que evocam. No contexto da tradução automática, por exemplo, uma teoria puramente estatística não nos parece capaz de atuar sozinha na desambiguação de sentido, ao menos não satisfatoriamente. Entretanto, associada a uma base de conhecimento, como uma FrameNet, ela poderá oferecer

equivalentes de tradução e tornar os sistemas automáticos mais eficientes e confiáveis para seus usuários.

Outro desafio apontado pelos autores, relacionado a um aspecto próprio da linguagem, refere-se ao fenômeno da ambiguidade. Segundo eles, um sistema de PLN precisa determinar algo sobre a estrutura do texto, geralmente o necessário para que se possa responder à pergunta *Quem fez o quê para quem?*. O problema, no entanto, é que "os sistemas convencionais de *parsing* buscam responder a esta pergunta apenas em termos de possíveis estruturas tidas como gramaticais a partir de escolhas de palavras de determinada categoria" (MANNING & SCHUTZE, 1999, p. 96). A abordagem estatística de PLN, de acordo com os autores, busca oferecer soluções a este impasse através da aprendizagem lexical automática e de preferências estruturais abstraídas de *corpora*. A grande mudança nesse sentido envolve não só o reconhecimento de categorias sintáticas, como classes gramaticais, oferecidas por um *parser,* mas também o reconhecimento de que existe uma quantidade de informação considerável na relação que as palavras estabelecem entre si – as *collocations*– e a informação semântica que estas carregam[1].

Em relação às *collocations,* estas seriam peças fundamentais para o PLN estatístico, pois são expressões que se tornam fixas na língua pela frequente repetição no uso. Uma *collocation* seria qualquer combinação sintática na qual, de algum modo, percebe-se que o todo tem uma existência independente da soma de suas partes (MANNING & SCHUTZE, 1999, p. 29). Segundo esta definição, o reconhecimento de tais expressões é relevante para a tradução por máquina, no que se refere à desambiguação de sentido, uma vez que as palavras são traduzidas de acordo com a *collocation* na qual se inserem. Considerem-se as expressões de (1) a (4):

(1) Heavy rain
(2) Strong rain
(3) Chuva forte
(4) Chuva pesada

O uso da *collocation* inglesa *heavy rain* é aceito de modo mais amplo do que *strong rain*, embora este último uso esteja gramaticalmente correto. De modo oposto, no português, dizemos *chuva forte* preferencialmente à *chuva pesada* e, embora esta última expressão seja

1 Cumpre ressaltar que o conceito de semântica em uma teoria probabilística é equivalente à noção rasa de semântica, ou seja, refere-se a combinações superficiais ao nível da forma que podem ser substituídas por outras combinações superficiais comparáveis estatisticamente. Não se trata aqui de semântica nos termos da semântica cognitiva, como se discutirá mais adiante.

compreensível, ela não se iguala à primeira em termos de convencionalidade, ou seja, para um falante nativo, a expressão *chuva pesada* não soará tão natural quanto *chuva forte*.[2] A resposta para tais preferências de combinação pode estar no próprio uso, expressões reiteradas no uso linguístico tornam-se preferenciais a outras expressões e se convencionalizam. Essa é uma questão fundamental para a Tradução por Máquina, pois sabemos que, em muitos casos, não é possível traduzir *collocations* palavra-por-palavra, como no caso de *chuva forte* e *heavy rain,* pois existem certas preferências de combinação que serão reguladas pelo uso. Há também casos em que não se pode interpretar uma sequência de palavras de modo composicional. Considere agora uma expressão idiomática inglesa em (5) e uma expressão correspondente no português em (6):

(5) It's a piece of cake!

(6) Isso é mamão com açúcar!

As expressões em (5) e (6) são *collocations* cristalizadas nas línguas e que não podem ser interpretadas a partir das palavras que as compõem. Por esse motivo, uma tradução palavra-por-palavra seria ainda mais ineficiente, tendo em vista que dizer "isso é um pedaço de bolo!" não sugere que algo seja muito fácil no português. Desse modo, tradutores automáticos precisam reconhecer sequências de palavras como possuidoras de sentido, bem como estruturas mais gerais da língua.

A vantagem da TM estatística, que será apresentada na seção 1.2.3.1, é a de considerar a frequência do *input* a ser traduzido. Se um tradutor utilizar apenas regras linguísticas para realizar uma tradução, seria muito provável que *chuva forte* fosse traduzido como *strong rain*. Como vimos anteriormente, tal expressão seria aceitável do ponto de vista gramatical, porém não usual. O papel de um tradutor nesse contexto não é o de apenas oferecer combinações gramaticalmente corretas, mas sim de oferecer ao usuário combinações que de fato fazem parte do repertório da língua. Por isso, a tradução estatística ofereceria uma solução a este impasse, ao considerar as combinações mais frequentes nos *corpora,* sendo possível reconhecer mais facilmente as *collocations* da língua.

2 Uma busca no corpus Portuguese Web 2011, composto por 4 milhões de palavras e disponível na ferramenta Sketch Engine, retorna 5.064 ocorrências do sintagma chuva forte, contra apenas 143 de chuva pesada.

1.2.1 Teoria da Probabilidade

Ao abordar o PLN estatístico, é imprescindível que se trate de alguns conceitos essenciais da teoria da probabilidade, ao menos de maneira introdutória. De modo geral, a teoria da probabilidade lida com a previsão da possibilidade de algo ocorrer. Para ilustrar esta propriedade, Manning & Schutze (1999) recorrem a moedas e dados – elementos mais simples que a linguagem –, através de experimentos que facilitam o entendimento da teoria.

Formaliza-se a noção de probabilidade a partir do conceito de experimento: "o processo pelo qual uma observação é feita" (MANNING & SCHUTZE, 1999, p. 40). Sendo assim, arremessar três moedas constitui um experimento. A partir daí, é preciso definir o conjunto de todos os resultados possíveis do experimento, o que em teoria da probabilidade é denominado espaço amostral **Ω**. Espaços amostrais podem ser discretos, quando envolvem um número finito de resultados possíveis, ou contínuos, ao possuir um número incontável de resultados. De acordo com os pesquisadores, no que se refere à linguagem, é necessário lidar apenas com espaços de amostragem discretos.

Voltando ao experimento da moeda, um resultado básico como Cara/Coroa/Coroa constitui um evento A. Então, temos que A é um subconjunto de **Ω**. Cabe ainda ressaltar que probabilidades são números entre 0 e 1, onde 0 indica impossibilidade e 1, certeza. Sendo assim, ao lançarmos a moeda três vezes, qual é a chance de obtermos 2 caras? O espaço amostral é representado em (7), no qual Cara é representado por H (head) e Coroa por T (tail):

(7) **Ω** = [HHH, HHT, HTH, HTT, THH, THT, TTH, TTT]

Neste experimento, cada um dos resultados de **Ω** é igualmente provável, tendo por isso probabilidade 1/8. Este tipo de situação é denominada distribuição uniforme. Considerando que se trata de um espaço amostral finito com resultados equiprováveis, temos P(A) = **{A}** / **{Ω}** e o evento que se espera obter é A = {HHT, HTH, THH}. Portanto, temos a fórmula em (8):

(8) $$P(A) = A \vee \frac{\Omega}{\Omega \vee = \frac{3}{8}}$$

No entanto, se considerássemos que a primeira moeda lançada fosse Cara, teríamos agora quatro resultados possíveis, sendo que em dois pode-se obter Cara novamente. Dessa forma, a chance de se ter duas Caras lançando a moeda três vezes agora é de ½. Este tipo de probabilidade, em que se tem algum conhecimento prévio sobre o resultado de um experimento,

é chamada probabilidade condicional. A fórmula em (9) ilustra que o conhecimento prévio de *B* influencia na probabilidade de *A*. Neste tipo de situação, em que o evento *B* já ocorreu, a probabilidade de *A* pode ser representada por:

(9) $$P(A \vee B) = \frac{P(A \cap B)}{P(B)}$$

Outro conceito fundamental no contexto da teoria da probabilidade é de variável aleatória. Segundo Manning & Schutze (1999), em vez de se trabalhar com uma série de eventos irregulares num espaço amostral, é possível tratar a probabilidade em termos de valores numéricos. Ainda segundo os autores, uma variável discreta é uma função X: $\Omega \rightarrow S$, em que *S* é um subconjunto contável do conjunto de todos os números reais R^n, de modo que X é denominado indicador variável aleatório.

Ao lidar com línguas naturais, como se observa o que foi dito até agora com relação à função de probabilidade *P*? O que dizer sobre a probabilidade de uma sentença como "*A vaca mastigou sua rumina"*? Como bem colocam os pesquisadores, eventos linguísticos possuem *P* desconhecido, o que nos leva à tarefa de estimar *P.* E é através de amostras de dados que se poderá observar a probabilidade de determinado evento. A proporção de vezes que determinado resultado ocorre é chamado de frequência relativa do resultado. Se *C* (*u*) é o número de vezes que um resultado *u* ocorre em N tentativas, então *C*(*u*)/*N* é a frequência relativa de *u.*

A distribuição binominal é uma função muito utilizada em tarefas de PLN estatístico em geral. Segundo os autores, quando se tem uma série de tentativas e apenas dois resultados possíveis e independentes, temos a distribuição binominal, como o exemplo clássico da moeda. Quando se trata de probabilidade na linguagem, embora não se possa considerar total independência entre os itens linguísticos num *corpus*, é possível abordar a probabilidade a partir da distribuição binominal para determinados propósitos. Exemplo disso seria procurar estimar em um *corpus* a frequência em que determinado verbo tem seu uso transitivo. Neste caso, estaríamos lidando com um experimento com dois resultados possíveis: ser transitivo ou não.

Quando se fala em probabilidade e, especialmente quando a relacionamos à linguagem, é importante que se considere a noção de entropia, que corresponde à quantidade de incerteza que se tem sobre uma variável aleatória. De acordo com Manning & Schutze (1999, p. 63)*, "entropia pode ser interpretada como a medida do tamanho do 'campo de busca', consistindo de possíveis valores de uma variável aleatória e suas probabilidades associadas".* Além disso, quanto maior a entropia de um experimento, maior a incerteza sobre o resultado do mesmo. Por isso, os autores também afirmam que o conceito de entropia pode ser empregado para medir a qualidade

de modelos probabilísticos, ou seja, estes são considerados bons modelos quando são capazes de diminuir a entropia de seus experimentos.

No âmbito da tradução automática, a entropia está diretamente relacionada com o processo de desambiguação de sentido. Ou seja, um tradutor que possui uma série de possibilidades de tradução para determinado *input* e que, por outro lado, não possui boas ferramentas de desambiguação de sentido, será pouco eficiente na tarefa de oferecer um *output* equivalente. Neste caso, a entropia do tradutor será consideravelmente grande e a probabilidade de se ter resultados confiáveis será bem menor.

Dentre os possíveis sistemas de desambiguação de sentido, interessam-nos aqueles que envolvem modelos de língua baseados em regras e modelos de língua estatísticos, os quais serão apresentados em mais detalhes na seção sobre tradução automática. Tradutores que baseiam suas traduções num modelo de língua, por exemplo, estão utilizando um modelo que envolve conhecimento sintático, semântico e morfológico, de modo a diminuir o nível de incerteza e oferecer os equivalentes de tradução mais prováveis. O mesmo é válido para aqueles que utilizam um modelo de língua estatístico como ferramenta desambiguadora de sentido. O *Google Translate*, por exemplo, utiliza ambos os modelos e garante, com isso, uma entropia consideravelmente menor em seus resultados.

Um modelo bastante recorrente em PLN estatístico, que também auxilia a tarefa de tradução automática, é o modelo n-*gram*[3]. A capacidade de prever a próxima palavra a partir de outra palavra já dada é uma tarefa clássica da modelagem de línguas e tal tarefa é viável pois, tendo como base uma grande quantidade de textos, é possível que se saiba quais palavras acompanham outras palavras. O histórico destas palavras então pode ser agrupado pelo método de Markov, que leva em conta apenas o contexto local – as últimas palavras –, já que estas influenciariam a probabilidade da palavra que viria a seguir. A utilização de n-*grams* nesta tarefa pode ser observada a partir do exemplo (10):

(10) Sue swallowed the large green ___.

Para que se possa prever a última palavra em (10) é preciso que se considere o histórico de combinações das palavras anteriores. Se utilizássemos um 2-*gram*, estaríamos levando em conta a combinação de duas propriedades representadas pelos adjetivos *grande* e *verde*, de modo que palavras como *árvore* e *montanha* seriam candidatas muito prováveis de continuar esta

3 Dentre alguns dos maiores corpora disponíveis online, está o N-gram Corpus, lançado pela Google, com 1 trilhão de palavras. O corpus apresenta as combinações N-gram e suas frequências, além de permitir que se faça a busca por unigrams ou por collocations de até cinco palavras.

sentença. Contudo isso não é possível, pois há um predicador – swallowed 'engolir' – que influenciará a probabilidade da última palavra. Considerando então o contexto mais amplo da sentença e, neste caso utilizando um 5-*gram*, temos que *pílula*, por exemplo, seria uma palavra bastante provável de ocorrer.

A possibilidade de prever a próxima palavra em função da palavra anterior, como proposto pelo modelo n-*gram*, é um tipo de aplicação muito útil aos sistemas de tradução automática, uma vez que a consulta do histórico de combinações do *input* com outras palavras pode auxiliar o tradutor no reconhecimento de quais combinações são mais prováveis. Tal método é utilizado por alguns recursos da *web*, como aqueles que sugerem as próximas palavras numa busca textual – vide Figura 1 –, além de também ser utilizado em desambiguação do discurso – quando um falante produz uma sentença não muito comum na língua, um modelo n-*gram* pode ajudar a reconhecer o problema e encontrar as palavras que o falante provavelmente quis dizer.

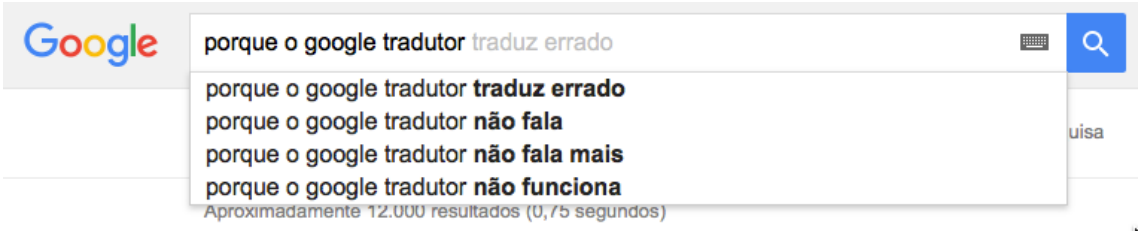


Figura 1: Exemplo de previsão de buscas do Google.

Contudo, veremos mais adiante que, mesmo com todo este amparo, a qualidade da tradução de um sistema automático ainda não se aproxima daquela oferecida por um tradutor humano. Isso nos leva a pensar o que seria preciso implementar nas máquinas, além das ferramentas de desambiguação já empregadas, para que elas desempenhem traduções mais adequadas, por meio de análises semânticas mais refinadas, tema este que será discutido no capítulo de análise.

Por ora, a subseção a seguir trará questões relativas ao trabalho com *corpus* e suas implicações para o PLN estatístico.

1.2.2 Corpus

Para se desenvolver um trabalho com *corpus* no âmbito do PLN, é preciso dispor basicamente de computadores, *corpora* e um *software*. Os trabalhos iniciais com *corpora,* como a construção do *Brown corpus* em 1960, não eram tão simples, devido à pouca memória dos

recursos computacionais disponíveis na época, mas, atualmente, é possível trabalhar com grande quantidade de textos e processá-los rapidamente.

O uso cada vez mais frequente da Internet pelas pessoas fez com que crescesse exponencialmente a quantidade de dados disponíveis para análise. No que se refere à abordagem estatística, isso é ainda mais relevante, pois, nos anos 50, quando tal método tornou-se popular, a quantidade insuficiente de dados não permitia que se fizessem generalizações estatísticas interessantes sobre a linguagem. Mas não só a quantidade de dados é interessante para o PLN, a diversidade também é um fator a se considerar. A noção de que um *corpus* precisa ser balanceado está relacionada à premissa de que ele deve ser um elemento representativo da língua, de modo que possa englobar fontes textuais diversas. O *corpus Penn Treebank* é bem conhecido por isso, tendo como fonte: Wall Street Journal, Brown Corpus, ATIS e Switchboard *Corpus*. O British National Corpus (BNC), por sua vez, também possui grande variedade de gêneros, domínios e mídias.

Os avanços na Linguística de *Corpus* viabilizaram muitas tarefas de PLN, especialmente no que se refere à possibilidade de lidar com grande quantidade de dados. No entanto, para que uma máquina seja capaz de aprender a processar uma língua, não basta oferecermos a ela apenas uma grande quantidade de dados. O processamento de textos não é uma tarefa simples, por isso é preciso prepará-los para que o computador possa encontrar mais facilmente os padrões que precisa aprender e fazer as inferências necessárias.

Manning & Schutze (1999) destacam que há uma série de propriedades subjacentes ao texto que podem torná-lo difícil de ser processado automaticamente. Os autores chamam de Formatação de Baixo Nível o tipo de tratamento inicial dado ao texto, antes mesmo que se comece a trabalhar com ele em determinada pesquisa. Isto é necessário pois há casos em que o texto eletrônico apresenta formatos e conteúdos de difícil processamento, os quais precisam ser filtrados, tais como: figuras, quadros, diagramas, dentre outros. Outro tipo de problema que precisa ser identificado inicialmente é a diferenciação entre letras maiúsculas e minúsculas. Ao encontrar uma mesma palavra escrita com letras maiúsculas e com letras minúsculas, esta será processada como sendo um único *token* ou teremos dois tokens diferentes? A resposta a esta questão pode depender do propósito do estudo, visto que seria interessante, por exemplo, reconhecer a distinção entre o nome próprio *Richard Brown* e a palavra *brown* por se tratarem de *tokens* diferentes.

A palavra *tokenização* surge a partir da necessidade de reconhecimento do que seja uma palavra. Isso, pois um passo primário do PLN é dividir o texto em unidades menores, chamadas *tokens,* os quais podem ser uma palavra, um número ou até mesmo um sinal de pontuação. Mas

afinal, o que conta como palavra? Embora possa se considerar os espaços entre as palavras como elementos limitadores, nem sempre textos eletrônicos comportam-se de maneira tão previsível, como é o caso de *smiles* feitos basicamente por sinais de pontuação. Outros problemas frequentes apontados pelos autores são:

(i) **Período:** um dos fatores contrários à definição de palavra exposta acima é o de que nem sempre palavras são cercadas por um espaço em branco, muitas das vezes elas são acompanhadas de um sinal de pontuação. Aqui surge outro problema relacionado aos sinais de pontuação, um ponto final pode indicar o final de um período, mas pode também estar indicando uma palavra abreviada, como *etc.*.

(ii) **Hifenização:** a hifenização também pode apresentar alguma dificuldade no processamento do texto, já que sequências de palavras unidas por um hífen podem contar como uma ou duas palavras. Há também a distinção entre hifens que unem palavras e aqueles que apenas indicam quebra de linha. Voltando ao problema inicial, o uso de fontes diversas de textos pode apresentar palavras ora grafadas com hífen ora sem hífen, como é o caso das palavras inglesas *data-base* e *database* (que pode, inclusive, ser representada do seguinte modo: *data base).* De todo modo, tais representações não indicam que se trate de palavras distintas, por isso pode-se pensá-las como um único lexema (uma única entrada de dicionário com um significado determinado). Sequências de palavras como *in order to* e *because of,* por exemplo, também deveriam ser tratadas como um único lexema, embora, na maioria das vezes, não o sejam.

(iii) **Homônimos:** já no caso de palavras homônimas, como *saw* (ferramenta) e *saw* (tempo pretérito do verbo *see*), é preciso que se considere a existência de dois lexemas, tendo em vista que tais palavras apresentam significados distintos;

(iv) ***Corpora* falados:** o trabalho com *corpora* falados apresenta desafios adicionais, visto que a fala é composta por contrações, diferentes representações fonéticas, variantes de pronúncia e muitas sentenças fragmentadas.

No campo da morfologia, os pesquisadores pontuam que numa língua morfologicamente simples, como o inglês, palavras como *sit, sits* e *sat* possam ser unidas num mesmo grupo e tratadas em termos de lexemas, ou seja, pela sua base comum, sem afixos ou flexão (o verbo *sit).* Contudo, cabe questionarmos se tal proposta seria viável em outras línguas, cujo sistema morfológico é muito mais rico, como o português. Considerem-se, a título de exemplo, as

palavras no diminutivo: elas deveriam ser tratadas como flexões dos nomes, ou como lexemas autônomos?

Apesar de tais dificuldades, há recursos que tornam o trabalho com *corpora* textuais mais satisfatórios. Trata-se de explorar a estrutura dos textos, através de um tipo de etiquetagem dos dados de um *corpus,* seja por um tratamento manual, automático ou até mesmo pela junção destes dois métodos[4]. De acordo com Manning & Schutze (1999), há níveis distintos de etiquetagem que podem dar conta desde os limites de sentenças e parágrafos, até a marcação de informações mais complexas que envolvem a estrutura sintática dos textos – como ocorre no *corpus TreeBank.* A etiquetagem mais comum, segundo os pesquisadores, é a relacionada às classes de palavras, a qual é feita para cada palavra de um *corpus* por meio de códigos correspondentes às classes de palavras. A Figura 2 ilustra a etiquetagem de uma sentença "She was told that the journey might kill her"[5] em alguns *corpora* do inglês (MANNING & SCHUTZE, 1999, p. 140):

Sentence	CLAWS c5	Brown	Penn Treebank	ICE
she	PNP	PPS	PRP	PRON(pers,sing)
was	VBD	BEDZ	VBD	AUX(pass,past)
told	W N	VBN	VBN	V(ditr,edp)
that	CJT	c s	IN	CONJUNC(subord)
the	AT0	AT	DT	ART(def)
journey	NN1	NN	NN	N(com,sing)
might	VM0	MD	MD	AUX(modal,past)
kill	W I	VB	VB	V(montr,infin)
her	PNP	PPO	PRP	PRON(poss,sing)
	PUN			PUNC(per)

Figura 2: Uma sentença etiquetada a partir de vários conjuntos de etiquetas

Esse tipo de etiquetagem é relevante para diversas áreas do PLN, como a Tradução por Máquina, por exemplo, que precisa reconhecer as classes gramaticais das palavras do texto-fonte e do texto-alvo. Além disso, quando a TM é baseada em regras, ou seja, quando utiliza um modelo de língua, a identificação de tais categorias é um passo fundamental no reconhecimento de unidades estruturais maiores, tais como sintagmas (PUSTEJOVSKY & STUBBS, 2012, p. 12).

Há também um tipo de anotação na qual sequências específicas de palavras são identificadas, através de estruturas que organizam palavras em sintagmas coerentes. É o que

4 Ressalta-se que mesmo o processo de taggeamento é probabilístico, assim o sistema de desambiguação também pode comprometer o resultado da etiquetagem.
5 "Disseram a ela que a viagem poderia matá-la"

ocorre no *corpus Penn Treebank,* em que se apresenta explicitamente a quebra do sintagma, introduzindo entre as palavras as relações de ordem e hierarquia.

Além de anotação sintática, é preciso levar em conta a etiquetagem semântica no *corpus.* Segundo Pustejovsky & Stubbs (2012), este tipo de anotação pode ser feita de duas formas: pela marcação do tipo semântico (o que é X) e pela marcação do papel semântico (qual o papel de X na sentença). O primeiro tipo de marcação envolve a identificação de uma palavra ou sintagma a partir de um vocabulário ou ontologia, demonstrando o que ela denota. Até mesmo uma estrutura ontológica simples, encobrindo aspectos como Pessoa, Lugar e Tempo, pode revelar fatos interessantes sobre a linguagem e facilitar o reconhecimento de padrões pelo computador. Considere o exemplo (11):

(11) [Ms. Ramirez]Person of [QBC Productions]Organization visited [Boston]Place on [Saurday]Time, where she had lunch with [Mr. Harris]Person of [STU Enterprises]Organization at [1:15 pm]Time.

Embora a categoria Tempo esteja sendo representada por elementos distintos, como dia da semana (Saturday) e hora (1:15pm), o reconhecimento de tais categorias em um *corpus* relativamente extenso pode trazer inferências importantes a respeito de como estas categorias interagem e sobre o funcionamento da língua.

Em relação ao segundo tipo de marcação semântica, consideram-se as respostas às seguintes palavras interrogativas: Quem, Qual, Onde e Quando. Ou seja, cabe aqui identificar quais são os papéis semânticos atribuídos pelo verbo, como Agente, Tema, Experienciador, Fonte, Objetivo, Paciente, dentre outros.

A FrameNet é um bom exemplo de recurso computacional que trata o *corpus* de modo a torná-lo legível pelas máquinas. O processo envolve camadas semânticas e sintáticas de anotação, de modo que tais informações fiquem integradas. Os papéis semânticos, denominados Elementos de *Frame*, são atribuídos aos itens lexicais a partir da cena conceptual (*frame*) de que participam. Na verdade, tal recurso opera com papéis microtemáticos, ou seja, funções mais específicas, já que cada *frame* possui participantes específicos – no *frame* de Comércio_comprar, por exemplo, o que seria tradicionalmente denominado Agente, é tomado como Comprador; já no *frame* de Comércio_vender, para o que seria Agente, temos o termo Vendedor. Já a anotação morfossintática envolve duas camadas: a anotação do tipo sintagmático e a camada de função gramatical dos itens lexicais. Tal processo será apresentado em mais detalhes no capítulo dedicado à Semântica de *Frames* e à FrameNet.

Na próxima seção apresentamos uma das aplicações do PLN estatístico, a Tradução por Máquina, a qual é escopo deste trabalho.

1.2.3 Tradução por Máquina

O interesse pelo desenvolvimento de sistemas de tradução por máquina se deu a partir do momento em que computadores britânicos foram capazes de decodificar enigmas alemães durante a II Guerra Mundial. Este feito motivou pesquisadores a desenvolverem a capacidade de ferramentas computacionais para decodificarem línguas estrangeiras, oferecendo traduções mais rápidas e rompendo com as barreiras linguísticas entre os países, ainda que por motivos econômicos e políticos. A ideia então vigente era a de que traduzir de uma língua para outra era uma tarefa semelhante à de descriptografar mensagens. Diante disso, métodos variados foram criados, desde tradução direta, com utilização de regras básicas, até métodos mais sofisticados baseados em análise sintática e morfológica (KOEHN, 2010, p.15).

Em linhas gerais, a Tradução por Máquina consiste em realizar a tradução de uma língua-fonte (a ser traduzida) para a língua-alvo (a tradução de fato), utilizando para isso determinado *software* e levando em consideração a estrutura gramatical de cada língua, o que permite ao tradutor automático selecionar uma tradução dentre outras possíveis. A tradução pode envolver não só textos, como também um discurso, áudio/vídeos, páginas da *web* e outras fontes. O objetivo final da TM, segundo Sawaf et al. (2010), também poderia ser o de dar suporte aos analistas de inteligência artificial e linguistas em suas tarefas de tradução humana.

Todo este processo de tradução não é tão simples. Um tradutor humano precisa decodificar o significado do texto-fonte e codificá-lo na língua-alvo, porém, há outras tarefas mais complexas subjacentes a esta decodificação. O tradutor também precisa interpretar e analisar as propriedades do texto-fonte a partir de seu conhecimento linguístico, o que envolve o conhecimento semântico, sintático, idiomático, bem como conhecimento cultural dos falantes da língua. E todo este conhecimento também é necessário com relação à língua-alvo (SAWAF ET AL., 2010, p.2).

Desenvolver tradutores automáticos capazes de oferecer traduções tão boas quanto um tradutor humano oferece tem sido o objetivo e também um grande desafio para muitos pesquisadores da área. Porém, cabe a pergunta apresentada por Sawaf et al (2010, p.1): *"Como você programaria um computador para 'entender' um texto da mesma forma que uma pessoa entende, e ainda como criaria um novo texto na língua-alvo que 'soe' como se este tivesse sido*

escrito por uma pessoa?”.[6] Se mesmo para um tradutor humano tal tarefa pode ser um tanto complexa, para uma máquina isso representa um grande desafio. Estudos recentes de TM estão trabalhando nestas questões, através da utilização de abordagens baseadas em regras e em estatística.

Nas palavras de Kohen (2010, p.20): “a tradução por máquina não precisa ser perfeita para ser útil, uma tradução por máquina de má qualidade também tem suas aplicações”.[7] Para se traduzir páginas da internet, por exemplo, não é preciso que se tenha um sistema de tradução automática de alta qualidade. O autor reconhece, então, três das principais aplicações de um tradutor automático, as quais requerem velocidade e qualidade diferenciadas: i) assimilação, quando a tradução oferece o entendimento genérico de determinado conteúdo; ii) disseminação, que, como o próprio nome sugere, contribui para a disseminação do texto traduzido para outras línguas; e iii) comunicação, que envolve a tradução de e-mails, conversas de salas de bate-papo, etc.

A seguir serão apresentados dois modelos de tradução amplamente utilizados nos sistemas de tradução automática atuais.

1.2.3.1 Tradução Estatística

A evolução da Tradução por Máquina se deu, principalmente, quando se descobriu que não bastava ensinar a uma máquina as regras de formação das línguas, tendo em vista a complexidade destas. Um caminho apontado para esta questão foi a de se considerar grandes *corpora* de textos traduzidos, pareando *input* a *output* e oferecendo traduções baseadas em análise estatística.

O primeiro modelo desenvolvido nesse sentido, o Modelo Baseado em Palavras, teve sua importância no desenvolvimento da Tradução por Máquina Estatística (TME) e no estabelecimento de alguns princípios que regem os modelos atuais. Esse modelo consistia em oferecer uma tradução palavra-por-palavra, ou seja, a tradução de uma sentença *input* seria feita por palavras isoladas, tendo como suporte a análise estatística. O quadro na Figura 3 apresenta as possibilidades de tradução de uma palavra do alemão para o inglês (KOEHN, 2010, p. 82):

6 How do you program a computer to “understand” a text just as a person does, and also to “create” a new text in the target language that “sounds” as if it has been written by a person?
7 “Machine translation does not have to be perfect to be useful, crummy machine translation also has its applications”.

Translation of *Haus*	Count
house	8000
building	1600
home	200
household	150
shell	50

Figura 3: Contagem de cada tradução possível da palavra alemã *Haus* para o inglês

A estatística tem papel relevante nesse contexto, na medida em que analisa a frequência de cada possibilidade de tradução, apontando a tradução mais provável. A palavra *Haus*, portanto, seria traduzida como *House,* tradução mais frequente segundo a análise apresentada no quadro acima.

Segundo Koehn (2010), empregar um modelo como esse equivaleria a abrir um dicionário bilíngue e buscar pela palavra desejada, obtendo seu sentido na língua-alvo. De fato, tal método apresenta uma análise muito simplificada e ignora o contexto no qual se encontram as palavras. Quando se trata de traduzir uma sentença inteira, o usuário teria que "montar um quebra-cabeça", juntando os significados de palavras isoladas para inferir o sentido global da sentença.

Outro conceito que merece destaque é o de alinhamento, processo responsável pelo pareamento de *input* a *output*. Num modelo de tradução baseado em palavras isoladas, o alinhamento também é feito palavra-por-palavra. Desse modo, uma sentença do alemão, como *das Hausist klein*, traduzida palavra-por-palavra, seguiria o alinhamento de cada palavra do alemão para o inglês, sendo que este processo é implícito. O esquema na Figura 4 ilustra o alinhamento realizado implicitamente ao processo de tradução de uma sentença (KOEHN, 2010, p. 84):

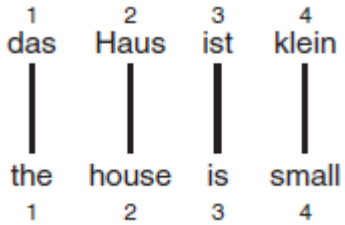


Figura 4: Alinhamento de cada palavra do alemão para o inglês

Neste esquema, todas as palavras foram alinhadas perfeitamente e a tradução de cada uma das palavras a partir de análise estatística resultou na sentença do inglês *the house is small*.

Mas o fato é que as línguas de modo geral não costumam se equiparar de tal modo, na verdade elas podem se distinguir muito em termos morfológicos, estruturais e semânticos também. Nesse sentido, o alinhamento lexical consegue recobrir alguns desses aspectos distintos entre as línguas. A mudança de ordem é um recurso necessário em casos como o mostrado na Figura 5 (KOEHN, 2010, p. 85):

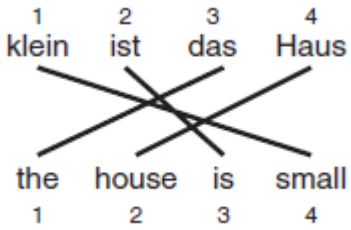


Figura 5: Reordenamento da sentença alemã para o inglês

As palavras foram reordenadas durante a tradução para garantir que a estrutura de ordem fixa do inglês seja mantida. Outro caso a se considerar refere-se à quantidade de palavras no *input* e no *output*. Quando o número de palavras difere, uma única palavra da sentença *input* pode ser alinhada a duas palavras da sentença *output*. Ou talvez seja o caso de ignorar tal palavra, quando ela não possui uma palavra equivalente na outra língua, como um verbo auxiliar, por exemplo. O alinhamento lexical pode-se tornar ainda mais complicado se traduzirmos expressões idiomáticas. A expressão do inglês *kick the bucket,* por exemplo, possui uma expressão equivalente no português, *bater as botas.* Mas não podemos afirmar que *botas* seja uma boa tradução para *bucket*, por isso, neste caso, o alinhamento lexical não faria sentido e, sim, o alinhamento sentencial.

Consideremos agora outro modelo de tradução, bastante utilizado atualmente, o qual baseia suas análises em pequenas sequências de palavras ou sintagmas que serão tomadas como unidades mínimas para a tradução. Casos, por exemplo, em que uma única palavra é traduzida por duas palavras em outra língua poderia ser um problema para o modelo anterior, mas não para este. Considere a tradução na Figura 6, do alemão para o inglês (KOEHN, 2010, p. 128):

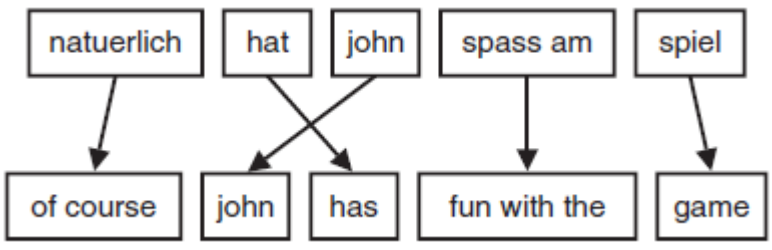


Figura 6: Tradução por máquina baseada em sequências de palavras

Primeiramente, a sentença do alemão é fragmentada em sintagmas que serão traduzidos para sintagmas correspondentes do inglês. Finalmente, tais sintagmas são reordenados. No caso da Figura 6, as seis palavras alemãs e as oito palavras inglesas foram mapeadas como cinco pares de sintagmas. Cabe notar que a divisão aqui apresentada não obedece à segmentação proposta pela teoria sintática, a qual apresentaria um sintagma nominal *fun* e um sintagma preposicional *with the game.*

Uma segunda vantagem deste modelo, segundo Kohen (2010), seria a redução da ambiguidade na tradução. No primeiro modelo, as possibilidades de tradução eram maiores, pois se desconsiderava o contexto, ampliando o número de traduções possíveis. Já no modelo baseado em sintagmas, diminui-se o espaço amostral e aumenta-se a chance de optar pela tradução mais adequada, tendo em vista que sequências de palavras juntas podem oferecer um contexto maior para a análise probabilística.

Voltemos à questão estatística, subjacente a estes modelos apresentados e que é responsável pela decodificação no processo de tradução, ou seja, pela busca da melhor tradução dentre tantas possíveis. Isso pode ser um problema, caso determinado *input* ofereça um número considerável de traduções prováveis. De acordo com Kohen (2010, p. 155), um erro de busca refere-se à falha do tradutor em tentar buscar a melhor tradução, isso pode ocorrer quando a tradução mais frequente não corresponde à melhor tradução. Em alguns casos é preciso se considerar a tradução pouco frequente. Mas, novamente, essa não é uma tarefa simples para a máquina. Enquanto somos capazes de escolher a tradução que faça mais sentido para nós, a máquina precisa fazer essa escolha numa situação como a ilustrada na Figura 7 (KOEHN, 2010, p. 159):

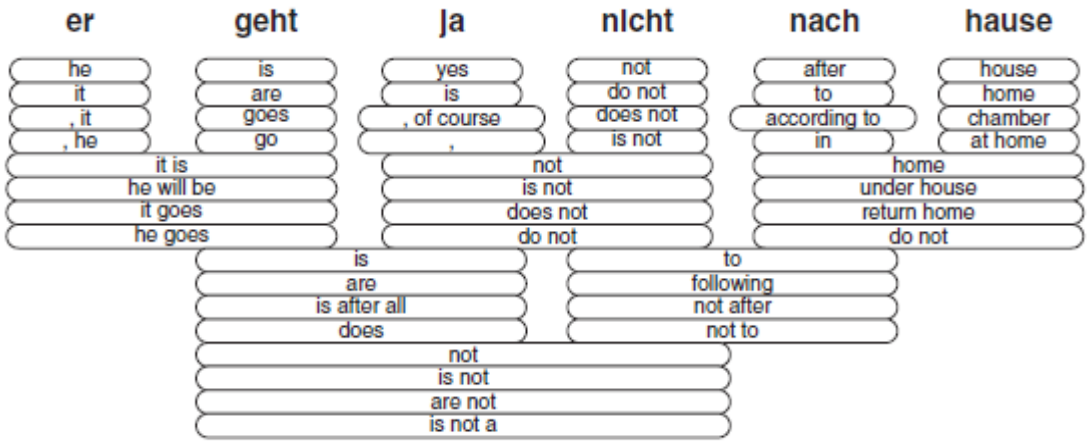


Figura 7: Opções de tradução para a sentença *input* do alemão *er geht ja nicht nach hause*

O quadro de sintagmas extraído do *corpus Europarl* apresenta todas as sequências de traduções possíveis a este *input*. A grande questão aqui seria encontrar traduções de sintagmas adequadas e colocá-las juntas numa sentença da língua-alvo. O quadro também mostra como o modelo de tradução palavra-por-palavra pode tornar ainda mais difícil este processo, como observamos na primeira tradução: *he is yes not after house,* a qual não corresponde a uma sentença do inglês e muito menos à tradução esperada *he doesn't go home.*

1.2.3.2 Tradução Híbrida

Além do modelo estatístico de tradução, há a abordagem baseada em regras linguísticas. Este modelo baseia-se no conhecimento linguístico e tem a habilidade de lidar com dependências de longa distância, concordância e ordem de constituintes, diferentemente do modelo estatístico, que não envolve nenhuma análise linguística profunda (KAMRAN, 2013, p.1).

Em relação ao modelo baseado em regras, este analisa o texto-fonte, criando uma representação simbólica, intermediária, a partir da qual o texto-alvo é gerado. De acordo Sawaf et al. (2010), tal método requer um léxico extenso com informações sintáticas e semânticas e ainda um conjunto expressivo de regras. Com todos estes dados, uma ferramenta de Tradução por Máquina Baseada em Regras consegue alcançar um nível de abstração suficiente para que um texto tenha uma tradução próxima do esperado. Entretanto, possuir uma quantidade de dados suficiente para tal abstração ainda é uma meta a ser alcançada para esta abordagem. Por isso, traduções que se pautam apenas em regras linguísticas ainda não atingem o nível de qualidade esperado.

Considerando as vantagens e as limitações de cada uma destas abordagens, nasce a motivação para se desenvolver um sistema híbrido de TM, em que se procure unir as aplicações de cada uma das abordagens e compensar suas falhas (SAWAF ET AL., 2010). Caso o tradutor se depare com uma rara combinação de palavra ou construção, esta deve ser analisada em primeira instância pelo módulo baseado em regras, visto que uma análise estatística pode não ser útil. O *Google Tanslate* é um exemplo de sistema de tradução por máquina que se utiliza deste método híbrido. O sistema de tradução com o qual opera pode ser visualizado esquematicamente na Figura 8 (SAWAF ET AL., 2010, p. 2):

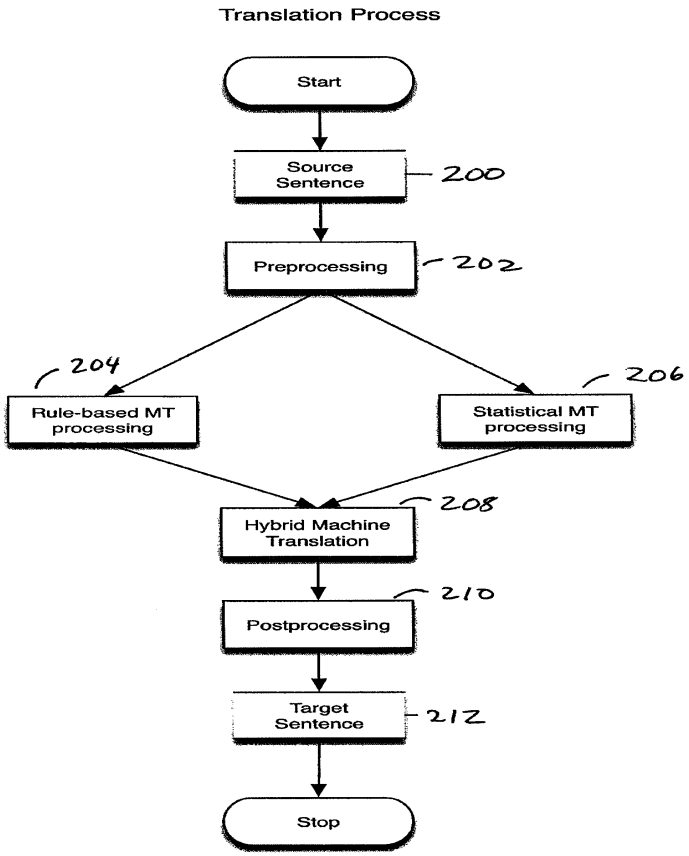


Figura 8: Processo de tradução híbrida do *Google Translate*

Neste esquema, observa-se o módulo de tradução baseado em regras, o módulo de tradução estatística e a ferramenta híbrida de tradução. Tal sistema ainda possui uma base de dados, na qual são armazenados *corpora* da língua-fonte e da língua-alvo, bem como os modelos de língua baseados em regras e os estatísticos.

No processo de tradução, o módulo baseado em regras traduz o texto-fonte através de um modelo de língua baseado em regras. Enquanto isso, o módulo estatístico traduz o texto-fonte baseado em modelos de língua estatísticos. E, combinados, tais métodos oferecem à ferramenta a possibilidade de cumprir melhora tarefa de traduzir um texto na língua-fonte para a língua-alvo (SAWAF ET AL. 2010, p.1). Regras são costumeiramente usadas, por exemplo, para a tradução de aspectos mais formulaicos dos textos, tais como datas, aberturas e fechamentos de cartas, por exemplo.

Embora este método seja mais eficiente, ainda não se tem um sistema automático de tradução de alta qualidade. O fato é que o módulo baseado em regras, por melhor que seja seu desempenho no momento, ainda não é capaz de lidar com toda a "irregularidade" das línguas naturais. Tais idiossincrasias não podem ser captadas pelo nível de abstração proposto pelo modelo. Isso porque o módulo opera com informações linguísticas básicas, com representações

muito amplas das línguas, diferentemente do que ocorre com a tradução humana, na qual o tradutor se utiliza de um repertório de informações semânticas e estruturais (os *frames* e construções) muito mais ricas do que o que se tem em tais sistemas. Em 2016, entretanto, o Google Translate passou a operar através de um novo modelo de tradução, o qual emprega a Tradução Baseada em Rede Neural, tecnologia que simula as conexões neurais humanas e permite oferecer traduções melhores que o modelo anterior. Dentre os fatores para tais melhorias, destaca-se a capacidade de o modelo considerar o contexto de uma sentença completa em vez de apenas algumas palavras.

No capítulo de análise, aponta-se para a possibilidade de utilização de um recurso computacional baseado em *frames* e construções como proposta a ser futuramente testada com vias a sanar algumas dessas deficiências.

2 SEMÂNTICA DE FRAMES E FRAMENET

2.1 O conceito de *Frame*

O estudo do significado lexical passou a ter um tratamento diferenciado com o surgimento da Semântica de *Frame*s, a partir das ideias de Fillmore e seus colaboradores (1975, 1977a, 1977b, 1982, 1985, dentre outros). As correntes semânticas anteriores, de orientação formalista, ocupavam-s e em caracterizar as condições sob as quais sentenças individuais de uma língua poderiam ser consideradas verdadeiras. Tais noções apresentadas pela "semântica das condições de verdade" eram secundárias aos propósitos de Fillmore (1985), que buscava desenvolver uma "semântica da compreensão". Esta seria baseada no conceito de *frame* e, segundo o autor, estaria interessada "*na relação entre os textos linguísticos, os contextos que os instanciam, os processos e os produtos de sua interpretação*" (FILLMORE, 1985, p.222)[8].

Segundo essa teoria semântica, há uma série de representações esquemáticas (cenários conceptuais) de nossa interação com o mundo, padrões de crenças, instituições, esquemas imagéticos, etc. que formam um pano de fundo ou *background* para a significação lexical. Tais estruturas são denominadas *frame*s e são responsáveis, portanto, por enquadrar um sentido específico de uma palavra em determinado contexto, já que uma mesma palavra pode assumir significados distintos em situações diversas. Ao assumir o *frame* como o construto mais forte e central para os estudos da significação, Fillmore amplia e reorganiza o estudo da semântica lexical, pois permite a caracterização não só de palavras, como também de sentenças e expressões de uma língua (FILLMORE, 2003a).

Cabe notar a diferença entre tal abordagem cognitivista, baseada em *frame*s, e uma abordagem tradicional da semântica lexical, representada pela Teoria dos Campos Lexicais. Esta última enfatiza as relações estruturais – sintagmáticas e paradigmáticas – que se estabelecem entre as palavras em determinado domínio semântico. O pré-requisito para a compreensão do significado de uma palavra individual seria o conhecimento do falante acerca da posição que esta ocupa no campo – na estrutura da qual participa – e dos demais competidores para esta posição (FILLMORE, 1985). Tal concepção implica que o sentido de uma palavra depende diretamente da existência de outras palavras em determinado campo semântico, as quais competem umas com as outras. O sistema de classificação de hotéis na indústria turística pode ilustrar essa competição entre os termos de um domínio bem como a necessidade de se conhecer os mesmos.

8 "(...) in the relation between linguistic texts, the contexts in which they are instanced, and the process and products of their interpretaton".

O sistema de classificação de hotéis no Brasil é relativamente simples, já que envolve categorização por estrelas, ou seja, quanto maior o número de estrelas, melhores serão as condições do hotel. Entretanto, nos Estados Unidos, a classificação pode ser um pouco confusa para um turista inexperiente. Ao buscar por um hotel da melhor qualidade, não seria estranho se este turista optasse por um hotel First Class, afinal *first* parece indicar que este estaria no topo do ranking, contudo, tal categoria de hotel seria a terceira melhor no ranking, que possui Deluxe e Luxury na primeira e segunda posição, respectivamente.

Fillmore (1985) reconhece alguma semelhança desta abordagem com a abordagem baseada em *frame*s e sugere que a noção de campo seja análoga ao conceito de *frame*, porém tais teorias adotam perspectivas diferentes ao tratar a significação lexical. Na Teoria dos Campos Lexicais, reafirma-se que o conhecimento acerca das palavras é que seja fundamental para a interpretação dos significados e não o esquema conceptual que subjaz a elas. Já para as teorias semânticas baseadas na noção de *frame*, a significação lexical só pode ser analisada a partir do background que envolve a palavra, o qual funciona como um pré-requisito conceptual para o entendimento do significado (FILLMORE & ATKINS, 1992). Nesse caso, o sentido não será depreendido pela relação que as palavras estabelecem entre si, mas pelo conhecimento prévio do falante acerca dos *frame*s que são evocados por essas palavras.

Esse pressuposto background que permite ao falante a compreensão e também a sua expressão na língua em diversas situações pode ser denominado como *frame*, esquema, cenário, *script,* narrativas culturais e memes. Embora todos estejam designando o mesmo conceito, há algumas diferenças que podem ser notadas quando estes termos são utilizados nos estudos de Inteligência Artificial, Psicologia Cognitiva e Sociologia. Por essa razão, Fillmore adota o termo *frame* que, segundo o autor, recobriria os demais, embora tenha seu foco na análise da forma linguística – palavras ou construções (FILLMORE, 2012).

Fillmore & Atkins (1992) exemplificam, a partir dos nomes referentes aos dias da semana, a diferença nos tratamentos da significação oferecidos pela Teoria dos Campos Lexicais e pela Semântica de *Frames*. O que dizer sobre a relação entre as palavras s*egunda-feira*, *terça-feira* e *quarta-feira*? Para a Teoria dos Campos Lexicais, existe uma relação de sucessão entre tais termos, de modo que configurariam um ciclo e participariam ainda da relação "parte de", tendo em vista que todos possuem relação direta com a palavra *semana,* a qual nomeia o ciclo. A descrição baseada em *frame*s, por sua vez, se interessaria em analisar como tais termos se enquadram no sistema completo de termos do calendário. Nas palavras dos autores: "*o que mantém esses termos juntos é o fato de serem motivados por, fundados em, e construídos com*

uma esquematização específica da experiência" (FILLMORE & ATKINS, 1972, p.77)[9]. No que tange aos nomes dos dias da semana, uma série de estruturas estariam sendo consideradas, como: a) o ciclo natural criado pela aparente viagem diária do Sol; b) o meio padrão de se calcular quando o ciclo de um dia termina e outro começa; c) o ciclo convencional de sete dias do calendário, com uma subconvenção especificando o membro inicial do ciclo; e d) a prática cultural de designar diferentes porções do ciclo semanal para o trabalho e para a folga. Tais estruturas calcadas na experiência são o background para outras organizações do tempo e de termos referentes ao calendário, como *semana, dia,* assim como seus derivados *dia de semana, final de semana*, além de interagir com estruturações maiores, como *mês* e *ano.*

Na próxima seção, serão apresentados alguns estudos preliminares acerca destas questões, os quais deram origem à teoria da Semântica de *Frame*s, bem como a suas aplicações tecnológicas.

2.2 Precursores da FrameNet

2.2.1 Os casos semânticos

Na década de 60, Fillmore procurava oferecer alguma contribuição à abordagem chomskiana de gramática vigente. Com a separação entre estrutura de superfície e estrutura profunda feita pela gramática gerativa, Fillmore acreditava que sua colaboração poderia se dar na descrição da estrutura profunda a partir de casos profundos, ou seja, de papéis semânticos, como Agente, Paciente, Tema, Fonte, Alvo, etc., pareados a possibilidades de manifestação sintática, como Sujeito, Objeto e Oblíquo. Surgiria então a Gramática de Casos, difundida com a publicação de *The Case for Case* (FILMORE, 1968).

Nesta teoria, os casos profundos estariam diretamente relacionados entre si. Numa sentença como "*João deu uma rosa a Maria", João* seria Agente, *rosa* seria Objeto transmitido e *Maria*, o Recipiente. Além disso, as funções gramaticais (sujeito, objeto direto, etc.) e os marcadores (preposições) seriam previstos pelas configurações dos papéis profundos; então o Agente poderia ser Sujeito, o Objeto transmitido seria o Objeto Direto e o Recipiente seria introduzido pela preposição *para*. A partir daí surge a discussão sobre Valência, a qual foi inicialmente apresentada pela análise do verbo abrir, como ilustrado em (12):

9 "What holds such words together is the fact of their being motivaed by, founded on, and constructured with a specific schematization of experience"

(12) Agente>Instrumento>Objeto

O = A porta abriu

AO = Eu abri a porta

IO = A chave abriu a porta

AIO = Eu abri a porta com a chave

Nesta representação, as funções sintáticas nucleares (Sujeito e Objeto) são aplicadas aos participantes segundo a hierarquia apresentada acima. Isso não significa, entretanto, que Agente será sempre Sujeito, como afirmam algumas gramáticas tradicionais. Como podemos perceber pelos exemplos acima, no nível sintático, a função de Sujeito pode ser atribuída a qualquer um dos participantes da cena. Estes participantes, por sua vez, assumem sempre o mesmo papel temático no *frame* evocado por *abrir: eu* será sempre Agente, *porta* será Objeto e *chave* será Instrumento.

Este foi o passo inicial que deu origem ao estudo acerca dos *frame*s. Começou-se a falar sobre participantes de uma cena, de papéis atribuídos a eles e a tratá-los como *frame*s de casos. A partir da lista de casos profundos criada, era possível definir situações específicas (o que mais tarde veio a ser chamar *frame*) e quanto maior o número de casos, maior seria a abrangência das situações descritas. A estrutura de casos de Fillmore é relevante no sentido de que pode ser intuitiva, ou seja, as propriedades lexicais por ele elencadas correspondem à forma como o falante pensa sobre suas experiências. Mais relevante ainda, como demonstra o autor em "*The case for case reopened"* (FILLMORE, 1977), é o pressuposto de que se deve fornecer uma ponte entre a descrição das situações e suas representações sintáticas subjacentes. E isso se daria a partir da atribuição de papéis sintático-semânticos a participantes particulares de uma situação representada por uma sentença. Fillmore (2012, p.18) apresenta algumas combinações de casos profundos que descrevem situações específicas em (13-15):

(13) Agente, Instrumento, Objeto: Eu consertei isso com a chave de fenda.

(14) Experienciador, Conteúdo: Eu lembrei do acidente.

(15) Estímulo, Experienciador: O barulho me assustou.

Papéis temáticos como os acima são capazes de descrever situações mais genéricas ou abstratas, mas não dizem muito a respeito de situações mais específicas. Um *frame* como o de TRANSFERÊNCIA, por exemplo, pode ser descrito a partir de elementos como Doador, Tema e Recipiente, pois o *frame* expressa uma noção genérica de transferência. Já uma cena como a de

COMÉRCIO_COMPRAR pode ser considerada uma elaboração mais específica do *frame* de TRANSFERÊNCIA, mas não pode ser descrita nos mesmos termos desta. É necessário então que se criem termos particulares que correspondam à categorização dos elementos em determinada situação específica, como COMPRADOR, VENDEDOR, MERCADORIA, etc. A semântica de casos não daria conta dessas especificidades, por causa da lista finita de casos semânticos. O falante possui em mente toda essa organização minuciosa das situações/cenas, ele não interpreta o mundo através de elementos tão gerais, como Agente, Caminho, Objeto movido, etc. E é isso que faz com que ele compreenda muito bem cada palavra em seu contexto, em seu *frame*.

Tendo isso em mente, Fillmore, em trabalhos posteriores (1982, 1985), demonstrou que a lista de casos semânticos era ineficiente para representar as propriedades lexicais e semânticas das palavras e que era preciso então colocar em evidência a cena cognitiva na qual um enunciado é produzido. Assim, desenvolve a premissa central de que os significados são relativizados às cenas (FILLMORE, 1977). Sendo assim, uma gramática de casos não precisa englobar todos os aspectos relevantes de uma situação, mas apenas uma pequena parte desta, ou seja, é preciso adotar uma perspectiva.

Seguindo essa lógica, o pesquisador conclui que "em vez de definir frames em termos de um conjunto de papéis, por que não criar os frames primeiro e então definir os papéis em termos dos frames?" (FILLMORE, 2012, p. 19). A descrição completa da cena EVENTO_COMERCIAL envolveria a identificação dos participantes: Comprador, Vendedor, Dinheiro e Bens. Contudo esta seria uma representação prototípica, tendo em vista que um enunciado envolvendo este frame adotaria uma perspectiva, diante da escolha lexical (comprar/vender/pagar/etc.). Além disso, neste mesmo evento, Comprador e Vendedor podem se comportar como Agentes, o que implica na escolha de uma perspectiva por parte do falante. Uma sentença como (16) mostra que, embora toda a cena Comercial esteja conceptualmente disponível, a escolha pelo verbo comprar coloca em evidência o Comprador e a Mercadoria. Por isso, a contraparte sintática, explorada na Gramática de Casos, mantém seu papel importante, uma vez que é esta que fornecerá pistas para a perspectiva adotada pelo frame.

(16) Meu irmão comprou um carro mais bonito.

Outra abordagem para o tratamento da perspectiva seria a noção de que alguns elementos, embora sejam ativados quando da evocação do *frame*, nem sempre são realizados linguisticamente. Tal concepção é fundamental para a anotação que se desenvolve na FrameNet

(aplicação computacional da Semântica de *Frame*s, a qual é apresentada na seção 2.3 deste capítulo), pois os dados, provenientes de uso linguístico, não constituem sentenças exemplares, na maioria das vezes. Desse modo, as sentenças podem apresentar a realização linguística de apenas alguns dos participantes de uma cena, deixando de expressar abertamente outros elementos também importantes, sem, contudo, comprometer a compreensão do enunciado.

Amparando-se na Gramática de Casos, a Semântica de *Frames* e, por consequência, a FrameNet, assumem a necessidade de relacionar informação gramatical aos papéis semânticos. Ou seja, os participantes de uma cena (Elementos de *Frame* – ou EFs) são definidos em termos de sua realização sintática e semântica. Contudo, tais EFs são representações mais específicas dos casos propostos por Fillmore em *The Case for Case*, pois são relativizados à cena da qual participam. O *frame* de COMÉRCIO, por exemplo, possui descrições específicas de seus participantes, como Comprador, Dinheiro, etc., enquanto os casos semânticos seriam generalizações das propriedades mais salientes destes participantes (Agente, Paciente, etc.).

Como será apresentado mais adiante, no capítulo 3, o estudo de Fillmore também se reflete diretamente no desenvolvimento da Gramática das Construções, uma vez que se assume que toda descrição construcional deve veicular informações de cunho formal e semântico (FRIED & OSTMANN, 2004), dada a definição de construção como pareamento forma-sentido. Além disso, a abordagem dos casos é relevante para a descrição dos elementos internos à construção. Uma evidência disso seria a representação da valência dentro dos diagramas de descrição, nos quais se encontram informações acerca da função gramatical e do papel semântico das expressões.

2.2.2 O frame de Risco

Outro estudo contribuiu consideravelmente para a teoria da Semântica de *Frames*, o estudo de Fillmore e Atkins (1992) sobre o *frame* de RISCO, em *Toward a Frame-Based Lexicon: The Semantics of Risk and its neighbors*. Esta foi a primeira tentativa de demostrar a aplicação da teoria da Semântica de *Frame*s para a prática lexicográfica, ao constituir um léxico baseado em *frames* (FILLMORE, 2003a). O estudo amparou-se ainda numa metodologia baseada em *corpus*.

A análise do *frame* de RISCO envolveu a descrição dos padrões léxico-sintáticos que envolvem o lexema *risk*, como Nome e como Verbo. Além disso, foram especificadas as categorias fundamentais para a descrição da valência de *risk*, tais como: Chance, Dano,

Objeto_valorizado, Situação_de_risco, Ato, Autor, Ganho_pretendido, Propósito, Beneficiário, Motivação.

Embora o estudo tenha demonstrado algumas análises de *risk* como Nome, a maior parte das considerações foram feitas acerca da forma verbal. Duas das principais considerações foram: i) o verbo *risk* sempre tem ou um complemento direto nominal ou gerundial, ambos tomados por objeto direto (objeto nominal e objeto gerundial); ii) há três categorias principais representadas gramaticalmente como objetos: Objeto_valorizado, Dano e Ato. A categoria Objeto_valorizado, por ser uma "coisa", é representada apenas por Objeto nominal; quanto às categorias Dano e Ato, pelo fato de a primeira ser um "evento" e a segunda, uma "ação", podem ser representadas verbalmente (na forma gerundial) ou nominalmente (nominalização da noção do verbo ou como metonímia de um evento ou ação).

Os exemplos (17-19) ilustram casos em que a categoria Objeto_valorizado é realizada e gramaticalmente representada como um típico Objeto Direto, ou seja, em sua forma nominal. Além disso, a categoria Situação_de_risco também está presente nesses exemplos, sendo expressa por um sintagma preposicional:

(17) He was being asked to risk {his good name} {on the battlefield of politics}.
(18) Others had risked {all} {in the war}.
(19) I would be foolhardy to risk {human lives} {in the initial space flights}.

Os resultados obtidos com este estudo demonstraram a possibilidade de se desenvolver um dicionário que baseie suas análises na noção de *frame* e que explore informação proveniente de *corpus* para enriquecer suas descrições. Este dicionário, que, em muitos aspectos, difere da lexicografia tradicional, é apresentado a seguir na seção dedicada ao projeto FrameNet.

2.3 O projeto FrameNet

O projeto FrameNet, desenvolvido no International Computer Science Institute (ICSI), por Fillmore e seus colaboradores, é um recurso lexicográfico que busca representar computacionalmente os *frames*, para que, através destes, sejam descritos os significados lexicais. Tal empreendimento deu origem a outros projetos, como a FrameNet Brasil, desenvolvida na Universidade Federal de Juiz de Fora, bem como outros projetos desenvolvidos para o espanhol, alemão, chinês, japonês e o sueco.

As análises realizadas na FrameNet tomam como unidade mínima a Unidade Lexical, que corresponde ao pareamento de um lexema a um *frame*, de modo que são anotadas sentenças provenientes *de corpora* que contenham tais ULs. De modo diferente do que ocorre em outros recursos lexicográficos, a FrameNet não se atém à análise de apenas uma palavra. Ao contrário, o recurso se ocupa de várias palavras, de uma só vez, que evocam um mesmo *frame*. Neste caso, o processo equivale a analisar um *frame* por vez, recobrindo todas as possíveis palavras evocadoras do mesmo. O objetivo da ferramenta, segundo Fillmore (2012), é criar uma base de dados que inclua todos os *frame*s possíveis de serem descritos, os quais são os esquemas cognitivos que subjazem à significação das palavras (FILLMORE, 2012).

Outro fator que diferencia a FrameNet da lexicografia tradicional, ainda segundo o pesquisador, é a reunião, em um mesmo conjunto, de palavras antônimas que podem instanciar, na sua localidade sintática, um mesmo Elemento de *Frame*, como é o caso das ULs *obey* e *disobey,* que, no *frame* COMPLIANCE, tomam como objeto o Elemento de *Frame* Acts. Já no caso de lexemas polissêmicos, o critério que se adota é a separação destes em *frame*s distintos, como é o caso do lexema *adhere*, que pode ser vinculado ao *frame* ATTACHMENT ou ao *frame* COMPLIANCE. Isto se deve ao fato de que os sentidos não são os mesmos e, portanto, as ULs também não são as mesmas – reiterando a concepção de que uma UL corresponde ao pareamento de uma forma (lexema) a um significado específico (*frame*), como veremos em detalhes na próxima seção.

2.3.1 Polissemia e Valência

Lidar com o fenômeno da polissemia é uma questão primária na FrameNet, isso porque, em sua metodologia, uma tarefa anterior à anotação é a escolha de um único sentido associado a determinado lexema. Uma pesquisa em *corpus* pode evidenciar que um mesmo lexema pode estar associado a sentidos diversos. As sentenças (20-22) demonstram as diferentes acepções do verbo *abrir*:

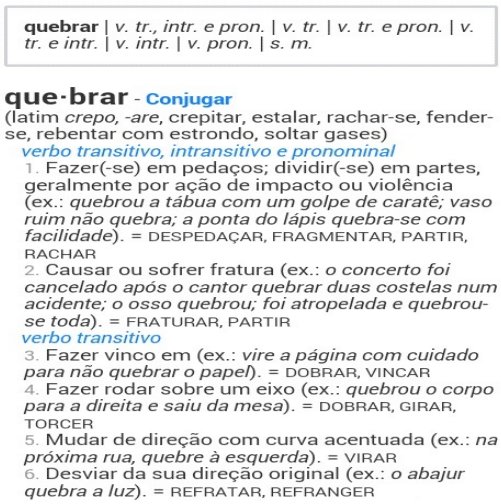
(20) Carlos abriu a janela por causa do calor.
(21) O policial abriu fogo contra os assaltantes.
(22) Ela não conseguia se abrir com a mãe.

Percebe-se em (20) que o verbo refere-se ao ato de abrir, considerado o sentido literal do lexema; em (21) e (22), no entanto, temos sentidos figurados do lexema, os quais se referem ao

ato de iniciar uma ação e ao ato de expor seus sentimentos a outra pessoa, respectivamente. Isso significa que teríamos três sentidos associados ao lexema e, portanto, três Unidades Lexicais diferentes.

De acordo com Fillmore (2008), a abordagem lexicográfica da FrameNet se opõe a abordagens monossemistas, as quais tendem a fazer generalizações acerca dos sentidos associados a um lexema. Nestas últimas, a descrição é feita palavra-por-palavra, de modo que são agrupados diversos sentidos em torno de uma única palavra. Enquanto isso, a FrameNet segue um caminho oposto, pois baseia sua descrição *frame*-a-*frame* e, por isso mesmo, não consegue recobrir, em uma só entrada, todos os sentidos de determinada palavra, porque o sentido a ser descrito será determinado pelo *frame* no qual ela se encontra. Este tipo de organização particular da FrameNet, que lança seu olhar primeiramente para o *frame*, faz com que se reúnam palavras, independentemente de suas diferenças gramaticais e distribucionais, em uma mesma lista, a lista de ULs evocadoras de um *frame*. Não se descreve uma palavra por vez, descreve-se um *frame* por vez. Ou seja, enquanto a lexicografia tradicional agrupa significados em torno de uma palavra, a FrameNet agrupa palavras evocadores de um mesmo significado de background (o *frame*).

Os sentidos relacionados ao lexema *quebrar*, por exemplo, são tratados de modo bastante distintos nestas duas abordagens. Observe, na Figura 9, a definição do verbo em um dicionário online da língua portuguesa[10]:



quebrar | *v. tr., intr. e pron.* | *v. tr.* | *v. tr. e pron.* | *v. tr. e intr.* | *v. intr.* | *v. pron.* | *s. m.*

que·brar - Conjugar
(latim *crepo, -are*, crepitar, estalar, rachar-se, fender-se, rebentar com estrondo, soltar gases)
verbo transitivo, intransitivo e pronominal
1. Fazer(-se) em pedaços; dividir(-se) em partes, geralmente por ação de impacto ou violência (ex.: *quebrou a tábua com um golpe de caratê; vaso ruim não quebra; a ponta do lápis quebra-se com facilidade*). = DESPEDAÇAR, FRAGMENTAR, PARTIR, RACHAR
2. Causar ou sofrer fratura (ex.: *o concerto foi cancelado após o cantor quebrar duas costelas num acidente; o osso quebrou; foi atropelada e quebrou-se toda*). = FRATURAR, PARTIR
verbo transitivo
3. Fazer vinco em (ex.: *vire a página com cuidado para não quebrar o papel*). = DOBRAR, VINCAR
4. Fazer rodar sobre um eixo (ex.: *quebrou o corpo para a direita e saiu da mesa*). = DOBRAR, GIRAR, TORCER
5. Mudar de direção com curva acentuada (ex.: *na próxima rua, quebre à esquerda*). = VIRAR
6. Desviar da sua direção original (ex.: *o abajur quebra a luz*). = REFRATAR, REFRANGER

Figura 9: Definição do verbo *quebrar* em um dicionário online

O dicionário apresenta uma série de possibilidades para o verbo, tanto gramaticais quanto semânticas. Já na primeira definição do verbo, a qual denota o sentido de fragmentar, temos a

10 www.priberam.pt/DLPO/Quebrar

reunião de dois eventos diferentes, uma vez que “quebrou a tábua com um golpe de caratê” envolve a ação direta de um agente, enquanto “a ponta do lápis quebra com facilidade” indica uma eventualidade não necessariamente causada por quem utiliza o lápis. Como observa Fillmore (2008), dicionários como este buscam por generalizações e, embora apresentem uma série de padrões sintáticos, apresentam como resultado um sentido básico para a palavra. Mas Fillmore entende que não só a valência pode ser múltipla como também os significados e que, nesses casos polissêmicos, é necessário que os significados sejam separados. Considere os exemplos (23-24):

(23) Eu quebrei a janela
(24) A janela quebrou

Segundo os princípios da FrameNet, em tais instâncias, *quebrar* representa duas unidades lexicais distintas, uma indicando um sentido causativo e outra, um sentido incoativo. No primeiro caso, o verbo, em sua forma transitiva, evocaria o *frame* Causar_fragmentar, no qual há necessariamente um Agente (*eu*) gerando a mudança de estado no Paciente_inteiro (*janela*). Já no segundo caso, o verbo intransitivo evocaria o *frame* Despedaçar, havendo simplesmente uma mudança de estado no Todo (*janela*), sem a presença de um agente. Por isso a FrameNet opta pela separação de tais sentidos, uma vez que eles se enquadram em *frame*s distintos. A distinção entre os sentidos de *quebrar* pode ser observada ainda pelos sinônimos a eles associados. Para o sentido causativo, temos sinônimos como *partir*, *triturar,* etc. e para o sentido incoativo, estilhaçar.

A FrameNet não só sustenta suas análises a partir de tais propriedades semânticas como também as correlaciona às propriedades sintáticas das ULs. Tais correlações são denominadas padrões de valência. A valência sintática se refere às especificações gramaticais e sintagmáticas, enquanto a valência semântica determina os Elementos de *Frame* que participam da cena evocada pela UL. Retomando o caso do verbo *quebrar,* na sentença (23), observamos a existência dos Elementos de *Frame* Agente e Paciente_inteiro, ambos expressos por um SN; gramaticalmente, *eu* possui a função de sujeito e por isso é anotado como Externo (Ext), e *a janela,* como Objeto direto (Obj). A sentença (24) apresenta apenas o EF Todo, realizado sintaticamente por um SN, com a função gramatical de Externo. Observe a associação de tais propriedades, as valências das ULs, nas colunas abaixo:

Quebrar	*EFs:*	Agente	Todo_paciente
(causativo)	TSs:	SN	SN
	FGs:	Ext	Obj

Quebrar	EFs:	Todo
(incoativo)	TSs:	SN
	GFs:	Ext

Segundo Fillmore (2008, p.129), a FrameNet oferece um tratamento mais adequado ao estudo da valência por:

> (1) ancorar-se em evidência de corpus; (2) basear a camada de valência semântica no entendimento dos *frame*s cognitivos que motivam e subjazem aos sentidos de cada unidade lexical; (3) reconhecer vários tipos de discrepância entre unidades no nível semântico/funcional e os padrões de forma sintática; (4) fornecer os meios de atribuição de interpretações parciais às valências que estão conceptualmente presentes, mas sintaticamente não expressas[11].

De fato, a ferramenta coloca em destaque muitas informações até então ignoradas por outros recursos lexicográficos. Estes últimos negligenciam muitas vezes a importância de se apresentarem devidamente as informações de cunho sintático e gramatical e, quando as apresentam, não se atentam para a ligação direta destas com a sua contraparte semântica, o que acaba gerando resultados insatisfatórios para o dicionário.

2.3.2 Relações entre frames

Da mesma forma que os esquemas cognitivos, os *frame*s estão interligados na mente do falante, numa rede cognitiva que permite a interpretação de diversas situações. A FrameNet registra a relação entre *frame*s, em uma rede que ilustra, por exemplo, o fato de um *frame* mais específico ser herdeiro de outro *frame* mais abstrato. Os *frame*s estão organizados de maneira hierárquica e se interligam a partir de relações como: Herança, Perspectiva, Sub*frame* e Uso.
No FrameGrapher, Figura 10, há a representação da cena de Causar_fragmentar e as relações que este estabelece com outros *frame*s.

11 (1) relying on corpus evidence; (2) basing the semantic layer of valency on an understanding of the cognitive frames that motivate and underline the meanings of each lexical unit; (3) recognizing various kinds of discrepancy between units on the semantic/functional level and patterns of syntactic form; and (4) providing the means of assigning partial interpretations to valents that are conceptually present, but syntactically unexpressed.

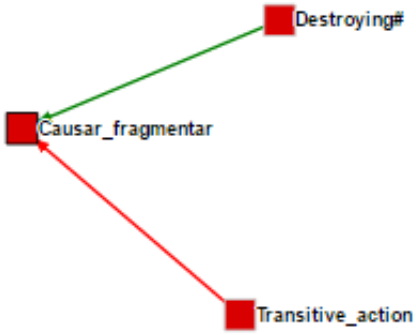


Figura 10: As relações entre Causar_fragmentar e outros *frame*s no FrameGrapher

Nessa representação, as setas indicam as relações que os *frame*s estabelecem entre si. As setas vermelhas indicam relação de Herança entre os *frame*s, de modo que podemos perceber que Causar_fragmentar herda de Transtive_action, enquanto usa Destroying. Isso significa, para o primeiro, que todos os EFs presentes no *frame*-pai (Transitive_action) devem constar no *frame*-filho (Causar_fragmentar), e que este é uma elaboração mais específica daquele, podendo apresentar mais EFs que o *frame*-pai; e, para o segundo, que Causar_fragmentar é pressuposto na ideia de Destruir, ou seja, que uma parte da cena evocada pelo *frame*-pai se refere ao *frame*-filho (RUPPENHOFER ET AL., 2010).

2.3.3 Estrutura da FrameNet e o processo de anotação

Um dos fundamentos da Semântica de *Frame*s é o de que a significação dos lexemas depende diretamente dos *frames* que evocam. Por isso, o primeiro passo para a descrição lexical constitui-se na escolha de um *frame* e na sua definição. A definição inicial do *frame* é um procedimento intuitivo, mas o linguista também baseia esta descrição em dicionários, artigos e thesauri. Retomando o *frame* Causar_fragmentar, anteriormente abordado, este pode ser definido conforme o apresentado na Figura 11.

A partir desta definição é possível analisar o sentido das palavras que evocam o *frame*. Mas a contraparte sintática também é analisada pela ferramenta, de modo que se especifica como as propriedades semânticas se realizam formalmente. Para isso, são apresentados conjuntos de anotações que exemplificam o *frame* abordado, bem como o uso de algumas das ULs. As expressões destacadas pelas cores no campo Definição da Figura 11 representam os Elementos de *Frame* (EFs), ou seja, os participantes da cena. Tais participantes são tidos como EFs centrais, pois são imprescindíveis na conceptualização do *frame* Causar_fragmentar, mas é possível haver outros participantes nesta cena que, do ponto de vista deste *frame*, são tomados como periféricos,

como: Grau, Explicação, Instrumento, Maneira, Meios, Lugar, Propósito, Superfície_resistente, Resultado, Sub-região e Tempo.

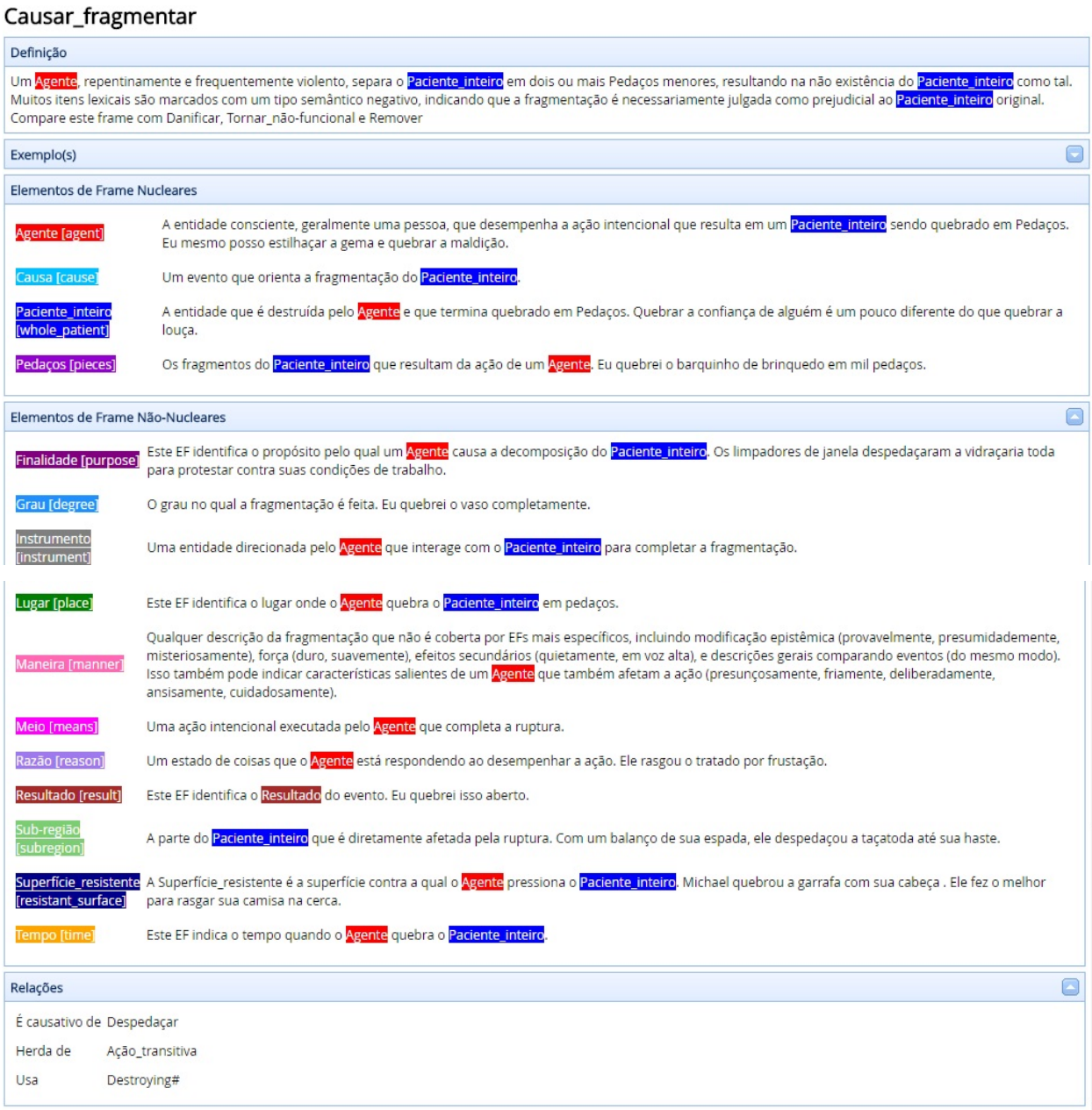


Figura 11: Descrição do *Frame*Causar_fragmentar

Para a descrição de um *frame* é necessário que se faça um levantamento, também inicialmente intuitivo, das ULs que o evocam, procedendo à anotação de sentenças que contenham os usos de tais ULs em camadas. São três as principais camadas de anotação: i) a camada Elemento de *Frame*, responsável por atribuir papéis semânticos aos constituintes da sentença; ii) a camada de Função Gramatical, correspondente à anotação de funções como Externo (que corresponde ao Sujeito), Objeto Direto, Objeto Indireto e Dependente – quando a

palavra alvo é um verbo – e Aposto e Núcleo – quando se trata de um nome ou adjetivo; e iii) a camada Tipo Sintagmático, em que é especificado o sintagma de determinado constituinte (se verbal, nominal ou preposicional). A anotação pode ainda se dar em outras camadas, quando for necessário especificar, por exemplo, casos de sentenças metafóricas, expressões idiomáticas e elementos suporte. A Figura 12 exemplifica a anotação da sentença *Eu quebrei o vidro pra poder abrir o carro e entrar.*

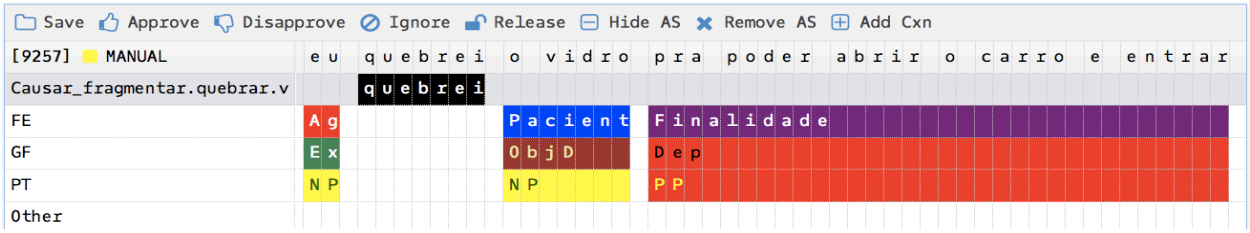


Figura 12: Anotação da UL *quebrar.v*no*Frame* Causar_fragmentar

Os tipos de anotação desenvolvidos na ferramenta são dois: a anotação lexicográfica e a anotação de texto corrido. A primeira tem por objetivo a descrição de um determinado *frame* a partir de uma palavra alvo e, nesse sentido, possui a vantagem de aprofundar a análise deste *frame*, já que são apontadas as ULs que o evocam e anotadas uma série de sentenças ilustrando seus usos. Com este método, é possível então realizar uma análise mais aprofundada do *frame* em questão. Em relação à outra proposta de anotação, o principal objetivo deve ser recobrir um grande número de *frame*s em uma mesma anotação, sem, com isso, aprofundar a análise de cada um deles. Neste último caso, a ferramenta ganha em abrangência de descrição, mas perde em profundidade de análise, já que não serão anotadas diversas sentenças ilustrando a UL evocadora do *frame*, mas apenas aquela que se encontra no texto sob análise.

O próximo capítulo aborda os pressupostos teóricos da Gramática das Construções que fundamentam as discussões acerca do padrão de quantificação indefinida.

3 GRAMÁTICA DAS CONSTRUÇÕES

As correntes construcionistas existentes compartilham entre si alguns princípios, ao mesmo tempo em que se diferenciam consideravelmente da abordagem gerativista. Uma das questões centrais de discussão é a admissão de que, do ponto de vista construcionista, não há na gramática um componente transformacional, como se assume na teoria gerativa, nem mesmo diferentes níveis de representação, com a separação entre estrutura de superfície e estrutura profunda. Haveria, na verdade, um único nível da representação, em que as sentenças são licenciadas pelas construções de uma língua. Desse modo, a semântica seria diretamente associada à forma de superfície (GOLDBERG, 2013). Não havendo um componente transformacional, a Gramática das Construções se propõe a oferecer um tratamento uniforme e integral das construções de uma língua, estejam elas na periferia ou no centro da gramática.

A postulação da existência de uma rede de construções também reúne a maioria dos gramáticos construcionistas. Aquela, também denominada Constructicon (FILLMORE, 1988), constitui o conhecimento que o falante tem sobre sua língua, um conhecimento geral que o permite utilizar e interpretar diversas sentenças, estando as construções todas interligadas via links de herança (GOLDBERG, 1995). Portanto, a gramática mental do falante, segundo essa teoria, seria formada por essa rede de construções.

Outro ponto fundamental de tal teoria é a concepção de que há um *continnuum* entre gramática e léxico, que abarca desde estruturas como palavras até as regras mais gerais da gramática. Não há, segundo essa teoria, razão para se separar tais estruturas, uma vez que estas se encontram todas interligadas numa rede de construções, a qual corresponde ao nosso conhecimento acerca da gramática da língua. Um parâmetro para a diferenciação dos tipos de construções seria a esquematicidade destas, de modo que o *continuum* vai das estruturas mais concretas (menos esquemáticas) às estruturas mais abstratas (mais esquemáticas). Outro argumento a favor deste *continuum*, por influência da Gramática Cognitiva (LAGANKER, 1987), é o fundamento simbólico das construções. Não só palavras são associadas a significados, palavras gramaticais como preposições, determinantes, dentre outras, são dotadas de significado, assim também como padrões mais gerais, como construções gramaticais. Neste caso, o valor simbólico de todos estes elementos faz com que os mesmos sejam incluídos numa gramática das construções. A tabela na Figura 13, apresentada por Goldberg (2013, p.3), resume o *continuum* discutido:

Construction	Examples
Word	*Iran, another, banana*
Word (partially filled)	*pre-N, V-ing*
Idiom (filled)	*Going great guns, give the Devil his due*
Idiom (partially filled)	*Jog <someone's> memory, <someone's> for the asking*
Idiom (minimally filled) *The* Xer *the* Yer	*The more you think about it, the less you understand*
Ditransitive construction: Subj V Obj1 Obj2 (unfilled)	*He gave her a fish taco; He baked her a muffin*
Passive: Subj aux VPpp (PPby) (unfilled)	*The armadillo was hit by a car*

Figura 13: Construções em vários níveis de complexidade e abstração

As expressões idiomáticas possuem a característica de não serem composicionais, pois o significado total da expressão não corresponde à soma das suas partes. Observa-se, pela tabela, que quanto maior o número de elementos fixos na expressão (*Going great guns*), mais esta se aproximará do léxico (das construções lexicais), pois, assim como as palavras, ela precisa ser memorizada no léxico mental do falante. Por outro lado, se a expressão admitir a introdução de palavras diversas, mais ela se aproximará das estruturas gerais e abstratas da língua, como a Construção Ditransitiva e a Construção Passiva. Nesse *continuum,* também se encontram, como apresentaremos na próxima seção, os padrões de cunhagem (cf. KAY, 2005, 2013). E, embora estes não estejam representados na tabela da Figura 13, podemos inferir que sua localização seja próxima às expressões idiomáticas minimamente preenchidas, pois, assim como estas, os padrões de cunhagem necessitam de pouco material fixo e podem ser preenchidos por uma variedade de palavras, como é o caso dos exemplos em (25-28):

(25) forte como um cavalo
(26) leve como uma pena
(27) burro como uma mula
(28) escuro como a noite

A visão clássica das construções como um pareamento forma-sentido retoma o princípio de Saussure sobre o caráter simbólico das estruturas da língua. Mas há um ponto importante nessa concepção, que divide opiniões entre alguns linguistas. Algumas abordagens construcionistas (cf. GOLDBERG 1995, 2006, 2013; CROFT 2001, 2013; BARÐDAL 2008;

MICHAELIS, 2013; BROCCIAS, 2013) assumem que todas as unidades linguísticas devam ser tomadas como signos Saussurianos, ou seja, para toda forma haverá uma contraparte semântica, de modo que não há nenhuma estrutura sintática autônoma. Por outro lado, abordagens mais heterogêneas da gramática (FILLMORE & KAY, 1993) consideram que a língua possui estruturas abstratas, princípios independentes da semântica, do mesmo modo que existem princípios semânticos independentes de uma contraparte sintática. Neste ponto de vista, ainda assim se consideram as estruturas linguísticas como signos Saussurianos, porém nem toda estrutura será um signo nesses termos (JACKENDOFF, 2013).

Tendo em vista as diversas abordagens para a Gramática das Construções, cumpre esclarecer que este trabalho se ancora, mais especificamente, na Gramática das Construções de Berkeley (BCG), por reconhecer a necessidade de assumir um viés mais formalista dentro de uma teoria semântica (Semântica de *Frames*). Esta necessidade é decorrente das análises que se desenvolvem ao longo da tese e que buscam verificar quais restrições governam a estrutura de quantificação binominal indefinida [N1 de [N2]] e que licenciam as diversas expressões de quantificação deste tipo na língua. De acordo com Fillmore (2013), a BCG assume que uma construção seja o pareamento não só de informação semântica e sintática, como também de todos os princípios que restringem e conectam tais construções. Nessa abordagem, busca-se representar os constructos e as construções através de um tratamento mais formal, e isso é possível através das matrizes de atributo e valor, a partir das quais são representadas as restrições que se aplicam sobre o licenciamento das construções.

Um conceito importante nesta corrente construcionista é o de Unificação. A BCG, assim como as demais correntes, nega a existência de um componente transformacional, de estrutura profunda e de categorias vazias, assumindo, na verdade, que existe um processo de Unificação, no qual a informação semântica é diretamente ligada à sintática. Nessa gramática, nos termos de Fillmore (2013, p.112): "*o que você vê é o que você tem*"[12]. Não havendo componente transformacional, assume-se que as construções ocorram simultaneamente, interligando-se de maneira a formar as diversas sentenças possíveis de uma língua, estando o significado diretamente ligado a tais estruturas. Em relação à matriz de atributo e valor, característica da BCG, é possível que se represente pela matriz a formação de uma construção sintática, além de evidenciar as construções menores que dela participam.

Nesse contexto, adotar um viés construcionista requer um olhar diferenciado para o tratamento das sentenças. Enquanto alguns gramáticos entendem que as palavras sejam os elementos primários que as constituem, os construcionistas, por sua vez, tomam como unidade

12 "What you see is what you get."

primária de análise a própria sentença, a qual será segmentada em unidades menores. Nesta última abordagem, a qual organiza a informação linguística através de estruturas de constituintes, as palavras não são necessariamente os elementos mais relevantes para a formação das sentenças. Embora o léxico também faça parte da gramática, é preciso que se considerem inicialmente as estruturas maiores (construções) que compõem os enunciados, as quais compreendem outras estruturas menores (também construções) e que são, então, preenchidas pelas palavras (construções lexicais). As sentenças são, portanto, um conjunto de construções as quais obedecem a uma hierarquia na estrutura de constituintes, de modo que uma ou mais construções se encaixe em outras construções. Considere o exemplo (29):

(29) O caminhão chegou.

A partir da proposta de análise da BCG, ao realizar a segmentação inicial da sentença (29) temos duas estruturas relevantes: a Construção de Sujeito-Predicado, a qual é composta por um SN (*o caminhão*) e um SV(*chegou*); e a Construção de Determinação, representada pelo SN e composta por um determinante (*o*) e um Nome (*caminhão*). A representação inicial da Construção de Determinação é feita por Fillmore & Kay através da matriz na Figura 14 (FILLMORE & KAY, 1995, p.22).

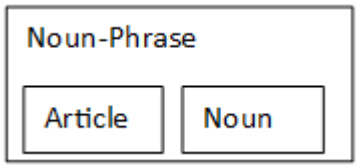


Figura 14: Construção de Determinação

Nesse esquema, temos a informação de que, para se formar um sintagma nominal, é necessário que se tenha uma palavra da classe Artigo e uma palavra da classe dos Substantivos. Quanto à representação do restante da sentença, que inclui o predicado da mesma, temos a matriz na Figura 15 (FILLMORE & KAY, 1995, p.23).

Figura 15: Construção Sujeito-Predicado

Após especificar a representação da Construção de Determinação e a Construção Sujeito-Predicado, é possível representar a estrutura que licencia o constructo (o material linguístico) *o caminhão chegou*, conforme Figura 16.

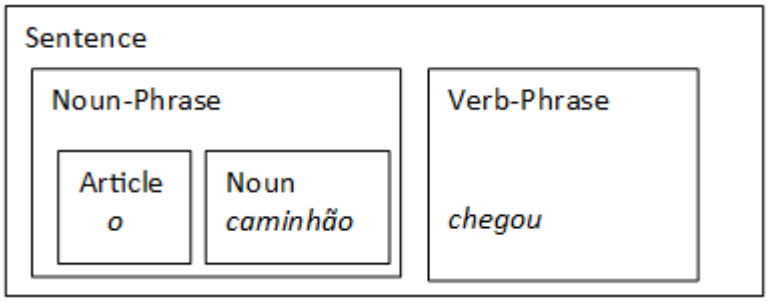


Figura 16: O constructo *o caminhão chegou*

Existem, contudo, outras especificações possíveis para a Construção de Determinação, que não foram representadas nas Figuras 14-16. A expressão de determinação (construção filha representada à esquerda), por exemplo, pode ser, além de Artigo, Demonstrativo (*este caminhão*), Possessivo (*seu caminhão*) ou Quantificador (*cada caminhão*). Isso indica que a representação desta estrutura precisa ser um pouco mais abrangente do que a apresentada na Figura 18. Existe ainda uma relação de dependência (e de compatibilidade) entre os constituintes posicionados à esquerda e à direita, assim não será todo Determinante que poderá ocorrer com qualquer Nome nessa construção. Por exemplo, não é possível dizer **estes caminhão* ou **meu caminhões*, pois há uma discrepância na informação referente ao número (singular ou plural) destes termos.

Retomando o conceito de Unificação, tão importante para esta teoria, e tendo em vista o que foi dito acima sobre a Construção de Determinação, observa-se que a informação referente ao número do constituinte à esquerda deve se unificar à informação do constituinte à direita. Isso não significa que ambos devam possuir a mesma informação singular/plural, mas, sim, que a informação de ambos não deve ser conflitante. Isto é o que observam Fillmore & Kay (1995) acerca do conceito de Unificação, ao dizerem que dois pedaços de informação devem se unificar somente se não forem contraditórios, se eles não entrarem em conflito.

As Figuras 17 e 18 apresentam a combinação de duas estruturas ou, nos termos da BCG, a Unificação entre elas:

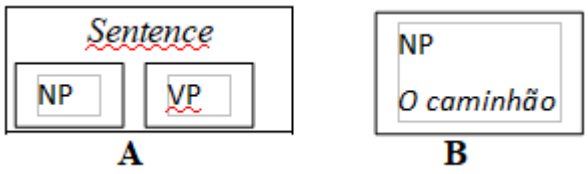


Figura 17: Representação da unificação entre construções A e B

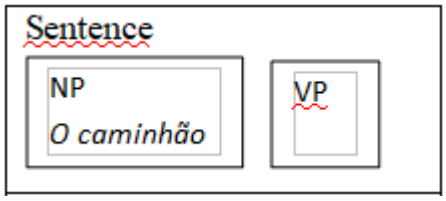


Figura 18: Resultado da unificação entre construções A e B

A unificação das estruturas A e B licencia o sintagma nominal *o caminhão* dentro da construção de Sujeito-Predicado. Do mesmo modo, embora não esteja ilustrado, um VP como *chegou,* poderia ser unificado ao VP da construção, licenciando então a sentença *o caminhão chegou.* A Unificação funciona, então, de modo a combinar diversas estruturas (ou construções), desde que sejam compatíveis, licenciando as possíveis sentenças da língua.

Outros aspectos das AVMs são revelados na Figura 19, sob o ponto de vista do constructo *o caminhão*:

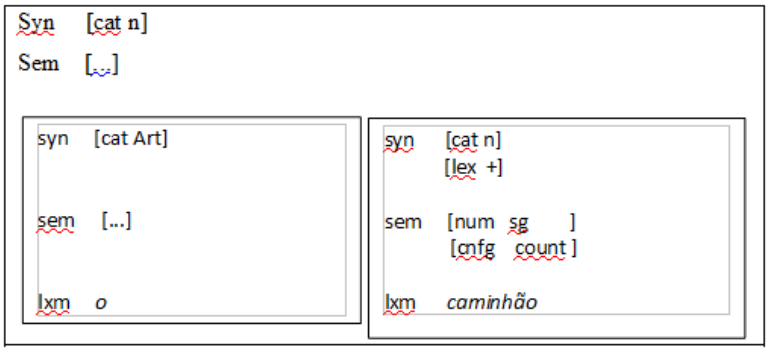


Figura 19: O constructo *o caminhão*

As atribuições de valores nas AVMs seguem a divisão entre aspectos semânticos [sem] e sintáticos [syn]. Na Figura 19, as propriedades externas do constructo são expressas pela AVM [syn [cat n], sem [...]]. Isso indica que o construto como um todo tem propriedades herdadas do núcleo nominal. Quanto às propriedades referentes aos signos filhos, temos que no signo à esquerda a AVM seria [syn [cat Art], sem [...], lxm *o*], já o signo à direita teria sua AVM descrita assim [syn [cat n], [lex +], sem [num sg], [cnfg count], lxm *caminhão*]. Assim, tem-se

que o núcleo lexical [lex +] do sintagma é um nome [cat n] singular [num sg] contável [cnfg count], o qual é antecedido por um artigo cujas propriedades sem foram subespecificadas [...] por serem as mesmas do núcleo nominal. O atributo lxm (lexema) é utilizado para indicar se o constituinte sendo descrito é uma palavra.

Na próxima seção, será apresentada a distinção entre construções e padrões de cunhagem, uma discussão relevante, que será retomada no capítulo de análise deste trabalho.

3.1 A distinção entre Construção e Padrão de Cunhagem

Ao revisar a literatura sobre a Gramática das Construções, deparamo-nos com a abordagem de Kay (2005, 2013), uma concepção estrita de gramática, na qual se deve considerar apenas a quantidade mínima de informação que o falante precisa ter para que seja capaz de entender e produzir sentenças da língua. Desse modo, assume-se que as construções de uma língua sejam estes padrões mais gerais e produtivos, os quais licenciam as mais diversas sentenças compreensíveis pelo falante. Uma expressão como *bola vermelha* configura uma estrutura de modificação do PB, altamente produtiva na língua e que permite a instanciação de inúmeras combinações do tipo. Esse é um conhecimento que faz parte da gramática, do conhecimento que o falante tem sobre a língua, que o possibilita interpretar e produzir diversas expressões com esta mesma estrutura.

Diante disso, Kay (2005, 2013) continua a discussão iniciada por Fillmore (1997) acerca de outras estruturas que, *a priori*, seriam tidas como construções, mas que, segundo os autores, seriam casos de padrões de cunhagem, os quais não se qualificariam como parte da gramática, por não serem suficientes nem necessários para interpretar qualquer conjunto de expressões da língua. Em 1997, Fillmore já apontava para a distinção entre construções e tais padrões:

> Podemos distinguir dois tipos de "criatividade" na língua. Um caso seria a habilidade dos falantes, usando recursos existentes [construções], para produzir e entender novas expressões. No outro caso, em que usamos o termo cunhagem, um falante usa padrões existentes na língua para criar novos recursos. (FILLMORE, 1997, p.2)[13]

Para ilustrar tal diferença, Kay, em *The Limits of (Construction) Gammar* (2013), coloca duas estruturas em contraste. A primeira, apresentada nos exemplos (30-32), se refere à construção do inglês *All*-cleft:

13 "We can distinguish two kinds of 'creativity' in language. In one case there is the ability of speakers, using existing resources [constructions], to produce and understand novel expressions. In the other case, the one for which we use the term coining, a speaker uses existing patterns in the language for creating new resources."

(30) All I can eat is half a pizza.
(31) All that one has to do is to start training earlier.
(32) All I want is to get it out of the flat…

A estrutura em (30-32), instanciada obrigatoriamente pelo quantificador a*ll,* indica uma leitura de baixa expectativa. Ou seja, numa implicatura escalar, em que o topo da escala seja comer uma pizza inteira, a construção *All*-cleft indica que tudo o que o falante em (30) consegue comer é metade de uma pizza. Por isso, nessa construção, *all* não funciona como quantificador universal, o que geralmente ocorre em outras configurações sintáticas. Mais importante para a discussão que se faz neste trabalho é o fato de que tal estrutura é tomada como construção, pois permite a instanciação de inúmeras proposições a partir do padrão *All*-cleft, sendo, portanto, consideravelmente produtiva na língua.

Já a segunda estrutura apresentada por Kay, não gramatical e não produtiva, refere-se a um padrão de cunhagem. As instanciações desse padrão são denominadas *collocations*, por serem combinações de palavras que se reiteram na língua, as quais podem ser observadas em (33-38):

(33) Dark as night
(34) Green as grass
(35) Free as a Bird
(36) Strong as a horse
(37) Big as a house
(38) Light as a feather

Tais exemplos são instanciações do padrão *A as NP,* que licencia muitas outras possíveis combinações ou *collocations*, indicando o sentido de "muito A". Porém, segundo o pesquisador, não é toda combinação de Adjetivo e Sintagma Nominal que resulta nesta mesma estrutura, com o mesmo sentido. Tal estrutura, portanto, não é produtiva e nem suficiente para que se interprete qualquer conjunto de expressões da língua inglesa.

De fato, cada uma destas *collocations* deve ser aprendida e memorizada. Já quando se trata de uma construção, o falante não precisa aprender e decorar todas as possíveis combinações entre as palavras licenciadas por ela. Com a Construção de Modificação (*bola vermelha)*, por exemplo, não é necessário que se aprenda cada uma das suas possíveis instâncias (*bola azul, bola roxa, sapato preto*, *camisa pequena,* etc.), é preciso apenas que se conheça o padrão de formação

destas. Em relação ao padrão de cunhagem ocorre o contrário, o falante precisa aprender todas as instâncias para que as utilize. É claro que a criatividade do falante também atua nestes padrões, mas ela é regulada pela analogia.

Considerem-se algumas expressões de quantificação indefinida do PB nos exemplos (39-48), que, do mesmo modo que as *collocations* anteriores, possuem um amplo conjunto de possibilidades:

(39) Mar de gente
(40) Oceano de calúnias
(41) Enxurrada de notícias
(42) Enchente de cartas
(43) Caminhão de dinheiro
(44) Mundo de problemas
(45) Onda de greves
(46) Navio de coisas boas
(47) Universo de notícias
(48) Tsunami de assaltos

As instanciações acima seguem o mesmo padrão de formação, a estrutura [N1 de [N2]], e várias são as possibilidades de preenchimento deste padrão por um nome do português (TAVARES, 2014). É interessante notar que muitos dos nomes que atuam como quantificadores neste padrão compartilham caraterísticas semânticas, podendo pertencer ao mesmo domínio e diferenciar-se, muitas vezes, pela leitura escalar que acionam, como em (39) e (40). Isso se justifica pelo processo de analogia subjacente à formação de tantas expressões de quantificação indefinidas no PB, ou seja, se *mar* pode funcionar como quantificador, um elemento da natureza com dimensões ainda maiores, como *oceano,* também pode. Assim também ocorre em (41) e (42), embora apenas um estudo diacrônico possa apontar qual expressão deu origem à outra, fica evidente a relação de analogia entre nomes que indicam algum fenômeno natural, funcionando como quantificadores neste padrão (*tempestade, avalanche, vendaval*, dentre outros). As expressões em (46), (47) e (48) são apontadas como possíveis candidatas a expressões de quantificação indefinida, por analogia com as expressões em (43), (44) e (45), respectivamente.

O rico campo de lexicalizações e a possibilidade de prever novas expressões para essa estrutura faz com que ela pareça um padrão muito produtivo de formação de expressões de quantificação indefinida. No entanto, não se pode considerar que haja produtividade de type em

relação a N1, uma vez que o slot não é aberto a ponto de aceitar diversos nomes do português. Existe, na verdade, uma série de restrições subjacentes à estrutura [N1 de [N2]] e a formação de novas expressões se dá por analogia com os nomes que se encaixam em tais restrições, a partir daquilo que já funciona.

Na visão estrita de gramática proposta por Kay (2013), não existe processo parcial de produtividade, sendo assim ou teríamos construções – constituintes da gramática e altamente produtivas – ou padrões de cunhagem – que estariam fora da gramática e não seriam produtivos. Já na abordagem de Goldberg (2013), elementos como *collocations* e expressões idiomáticas teriam seu lugar dentro da gramática, mas seriam tratadas como construções com generalização limitada. Segundo a autora, uma solução para estas construções de difícil generalização seria relacioná-las a construções mais gerais da língua através de herança, captando assim os aspectos regulares e irregulares de cada construção.

Quanto ao *continuum* proposto pela Gramática das Construções, os padrões de cunhagem estariam mais próximos do léxico, pois suas instâncias, as *collocations*, precisam ser reiteradas na língua para que se convencionalizem. É a partir de sua reiteração no uso que elas são aprendidas e memorizadas (assim como o léxico de uma língua) e não a partir de uma estrutura geral de formação das mesmas.

As implementações computacionais da Semântica de *Frames* e da Gramática das Construções (FrameNet e Constructicon) também assumem a continuidade entre gramática e léxico, embora, por razões metodológicas, léxico e gramática sejam tratados separadamente – informações lexicais são anotadas no recurso lexicográfico e informações gramaticais, no recurso construcional. Em relação à modelagem de construções, muitas estruturas já foram anotadas no Constructicon. O empreendimento atual da ferramenta tem sido modelar as estruturas mais gerais do PB, aquelas que, segundo Kay (2013), fazem parte do conhecimento da gramática de uma língua e que são suficientes e necessárias para interpretar e produzir as sentenças da mesma. Nesse contexto, torna-se necessário revisar algumas construções já modeladas anteriormente na ferramenta, até mesmo para que se possa verificar até que ponto a distinção entre construção gramatical e padrão de cunhagem é interessante para a modelagem computacional de construções.

A própria estrutura de quantificação indefinida, apresentada nos exemplos (39-48), foi criada no Constructicon, a partir do trabalho de Tavares (2014), mas nossa proposta é a de que esta seja reanalisada para que possamos então concluir se de fato ela é uma construção ou um padrão de cunhagem e, após isso, reavaliar seu lugar analítico. Para tratá-la como construção, é preciso que se encontrem as restrições do padrão N1 de N2 que licenciam constructos como

enxurrada de informações, *mar de gente, chuva de críticas* e descobrir se, assim como na visão de Goldberg (2013), este seria o caso de uma construção de difícil generalização. Por outro lado, se a investigação desta estrutura não oferecer pistas sobre uma possível generalização, seria viável tratá-la como um padrão de cunhagem. Neste caso, teremos ainda de propor a remodelagem do padrão e investigar quais são os requisitos para a modelagem de padrões de cunhagem, tendo em vista sua aproximação com o léxico.

3.2 O Constructicon

O projeto FrameNet, apresentado na seção 2.3 deste trabalho, evidenciou durante seu desenvolvimento a necessidade da criação de outro recurso o qual daria suporte às suas análises e que seria um produto computacional da abordagem construcionista da gramática. Este empreendimento, denominado Constructicon, daria conta de representar as estruturas da língua – as construções gramaticais[14] – com as quais a anotação lexicográfica não seria capaz de lidar. Como ponto de partida, os pesquisadores da FrameNet de Berkeley buscaram investigar como seriam anotadas as estruturas *non-core*, isto é, aquelas que não ocupam posição central na gramática da língua inglesa. Mas o objetivo deste projeto de modo geral, é ampliar-se e tornar-se, segundo Fillmore (2008), um repertório de construções gramaticais da língua.

Cabe ressaltar, primeiramente, que, embora o Constructicon seja uma aplicação computacional da Gramática das Construções, fazer um Constructicon tem um propósito diferente de se fazer uma gramática baseada em construções. Nem todo objeto linguístico, embora esteja descrito numa Gramática das Construções, será descrito num Constructicon. Isso porque, enquanto a primeira deve ser capaz de lidar com todo material linguístico – inclusive as construções lexicais –, um Constructicon descreve apenas aquelas construções não tratadas lexicograficamente pela FrameNet, para que se evite a replicação da descrição das construções e redundância na base de dados das ferramentas. De qualquer modo, a intenção é a de que cada construção seja representada de modo compatível com o desenvolvimento da gramática construcionista. Também se espera que este recurso possa ser útil para novos níveis de *parsing* e outras atividades que envolvam PLN (FILLMORE, LEE-GOLDMAN & RHOMIEUX, 2012).

A necessidade de desenvolver um Constructicon ficou evidente, de fato, quando a FrameNet buscou ampliar o escopo de análise da UL para a sentença, através da anotação de texto corrido. Originalmente, a FrameNet foi criada para dar conta da análise de diversas ULs do

14 O Constructicon ocupa-se em descrever apenas construções gramaticais, uma vez que as construções lexicais são descritas no modo lexicográfico de anotação.

inglês, de modo a oferecer descrições lexicais baseadas em seus padrões sintáticos e semânticos (valência). Mas, com o crescimento natural da ferramenta, com seu repertório de frames e ULs anotadas, seria possível iniciar um processo de anotação de texto corrido que permitisse a visualização das ULs dentro de um contexto, da interação entre elas e da estrutura maior que as une e que é responsável pela integração semântica de toda a sentença. Durante o processo, muitas sentenças foram anotadas, enriquecendo e ampliando a base de frames e ULs descritos, porém isso apenas funcionou com estruturas canônicas da língua, que não apresentam um alto grau de complexidade na relação entre os seus constituintes. Já aquelas sentenças mais complexas, frequentemente presentes nos mais diversos textos da língua, mostraram-se inviáveis à anotação de texto corrido. Este é o caso de um trecho retirado da *Economist Magazine*, de 10 de maio de 2007, apresentado por Fillmore (2008, p.5):

> For all the disappointments, posterity will look more kindly on Tony Blair than Britons do today. Few Britons, it seems, will shed a tear when Tony Blair leaves the stage on June 27th after a decade as prime minister, as he finally announced this week he would do. Opinion polls have long suggested that he is unpopular.

Há uma série de estruturas nestas sentenças que merecem um tratamento diferente do que ocorre em uma anotação lexicográfica. Algumas delas, segundo o autor, são: a) na expressão *for all disappointments, for all* é uma construção do inglês que designa a ideia de concessão (*apesar de*); b) *look kindly on* se refere a um *phrasal verb* que denota um "julgamento positivo"; c) *it seems,* por sua vez, não estabelece nenhuma relação estrutural com o resto da sentença, trata-se apenas de uma cláusula parentética modal; d) *June 27th* é uma dentre outras maneiras de representar o pareamento entre um dia e um mês em inglês. Esses são alguns dos padrões que ocorrem nesse fragmento e que merecem um tratamento diferenciado no que se refere à anotação de texto corrido. Foi a partir da identificação dessas estruturas e da necessidade de descrever propriedades tão complexas que a equipe da FN de Berkeley se propôs desenvolver um recurso computacional como o Constructicon.

Em relação à anotação feita num Constructicon, a catalogação das construções gramaticais de uma língua descreverá suas características gramaticais e sua contraparte semântica (o *frame* a que se referem), ligando cada uma destas descrições a um conjunto de sentenças anotadas para exemplificar tais características e utilizando alguns dos mesmos recursos utilizados pela FN.

Embora se fale em repertório de construções, o que se anota de fato no Constructicon são os constructos, ou seja, as expressões licenciadas pelas construções. Uma construção pode ser descrita formalmente (AVMs) ou informalmente (em prosa). Mas, novamente, a anotação é feita no nível do constructo, de modo que: "*(...)cada anotação captura as propriedades de um*

construto particular com relação a uma construção particular que o licencia"[15] (FILLMORE, LEE-GOLDMAN & RHOMIEUX, 2012, p.321).

Em relação à anotação construcional, são adotadas duas segmentações, uma dentro da outra: delimita-se a passagem que contém a construção apenas e identificam-se as entidades linguísticas dentro deste segmento, os elementos constituintes. O segmento como um todo representa o signo mãe do constructo, já os elementos constituintes são os signos filhos, ou Elementos da Construção nos termos da FrameNet (de modo paralelo aos Elementos de Frame). A identificação e rotulação dos constructos e Elementos da Construção (ECs) se dão somente em termos de descrições gramaticais parciais, uma vez que a informação semântica é relacionada aos elementos através de um link proposto – mas não implementado pela FN de Berkeley – com a base de *frames* (FrameNet). Já com relação à construção, a descrição é feita em prosa (assim como é feito com os *frames*), e a informação dada apontará para os *frames* relevantes para esta descrição. Por exemplo, na descrição da construção do inglês *Month-plus-Date* serão apontados *frames* referentes a datas, meses e anos, e ao constructo será atribuída à informação de que este designa um *day-size unit.* Este é um constructo do tipo nominal com dois ECs (mês e data), com suas restrições semânticas e sintáticas cada.

Como era de se esperar, a FN-Br também deu início ao desenvolvimento de um Constructicon para o PB, porém com algumas modificações do projeto original de Berkeley. Consideremos um constructo do português brasileiro (49), licenciado pela Construção Transitiva_Direta_Ativa, para que possamos ilustrar mais a diante o processo de anotação realizado pelo Constructicon do PB:

(49) O menino cortou o pano

Neste exemplo, todo o segmento - *o menino cortou o pano* - representa o signo mãe, enquanto os fragmentos – *o menino* – e – *cortou o pano* – são os ECs. Esta construção evoca o frame de Ação_transitiva e a anotação da sentença (49) é ilustrada na Figura 20:

15 (…) each annotation captures the properties of a particular construct with respect to a particular construction that licenses it.

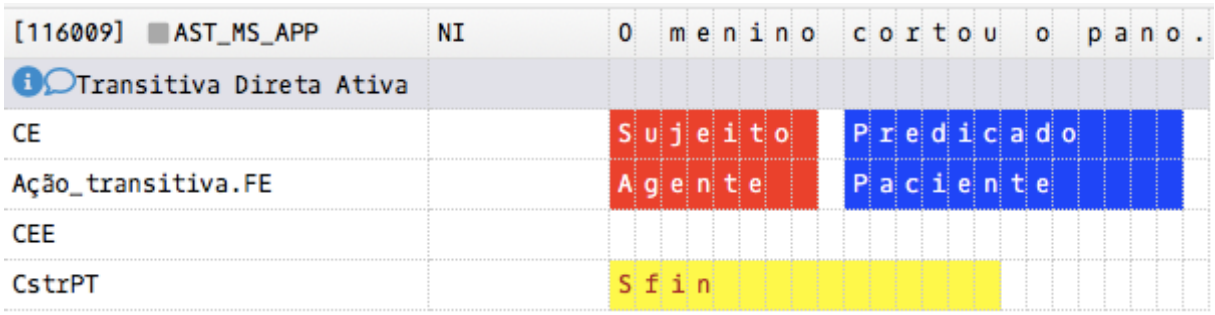


Figura 20: Anotação do constructo licenciado pela Construção Transitiva_Direta_Ativa

A anotação construcional, assim como a lexicográfica, é realizada em camadas, as quais oferecem informações acerca dos ECs, do Frame evocado pela construção e pelos seus ECs, além do Tipo Sintagmático do constructo. Como se pode observar pela figura, o primeiro EC é identificado como Sujeito, enquanto o segundo EC caracteriza-se como Predicado. Com relação ao *frame* indicado na segunda camada, temos que o signo mãe (a construção como um todo) evoca o frame de Ação_transitiva, enquanto os signos filhos são pareados aos EFs Agente e Paciente. É importante lembrar que a anotação centra-se em aspectos formais do constructo, isso porque a FrameNet Brasil promove a unificação entre os aspectos formais descritos no Constructicon aos aspectos semânticos (Frames e Elementos de Frame) presentes na base de dados FrameNet. Esta correlação entre ECs e EFs é feita através de uma relação criada para este fim, a relação de Evokes. Ao unificar os aspectos formais aos aspectos semânticos, integrando assim as duas bases de dados (FrameNet e Constructicon), estamos confirmando e modelando a continuidade entre gramática e léxico, um dos fundamentos da gramática construcionista (LAGE, 2018). Quanto à última camada, o constructo como um todo é rotulado como uma Sentença finita (Sfin).

Outra mudança significativa no Constructicon do português se deu a partir da premissa de que, assim como ocorre numa língua natural, a rede de construções modelada computacionalmente deve ser totalmente interligada por meio de relações entre as construções. No Constructicon de Berkeley, apenas relações entre *frames* e relações entre uma UL e um *frame* são modeladas, nos casos em que há uma conexão entre uma construção e um *frame* ou em que há relações de herança entre construções, tais informações são registradas na descrição textual das construções e não modeladas de fato. Percebendo a necessidade de se realizar a modelagem dessas relações, a FN-Br introduziu em sua base de dados a noção de Entidade, a qual pode ser de vários tipos – *word forms*, lexemas, lemas, ULs, Elementos de *Frame*, *frames*, Elementos da Construção, construções e relações –, de modo que o sistema permite que todas estas Entidades possam se relacionar entre si.

Segundo Laviola et al. (2017), a base de dados do Constructicon da FN-Br possui até o momento 110 construções modeladas, incluindo 13 construções da família Para Infinitivo (LAGE, 2013), 1 construção de quantificação indefinida (TAVARES, 2014), 65 construções equivalentes a construções do Constructicon de Berkeley (LAVIOLA, 2015) e 31 construções da "*core grammar*" do português brasileiro, como estruturas argumentais de Sujeito-Predicado, dentre outras (ALMEIDA, 2016).

Quanto à decisão de se anotar um objeto linguístico num Constructicon, Lage (2013, p.82) observa que nem todas as construções devem ser descritas nesse âmbito, uma vez que algumas construções podem ser tratadas via padrão de valência e, portanto, descritas num lexicon. Tendo isso em vista, a pesquisadora elaborou três critérios – abordados em mais detalhes no capítulo de análise –, os quais apontam o caminho na escolha pelo ambiente mais adequado de anotação, construcional ou lexicográfico. A investigação realizada na dissertação de mestrado apontou para a anotação construcional do padrão de quantificação, tendo em vista a adoção dos critérios então propostos por Lage, e por isso a estrutura foi modelada no Constructicon.

A investigação atual, por sua vez, coloca em questão o status da estrutura de quantificação como construção, nos termos de Kay (2005, 2013), e traz à tona a importância de se avaliar a adequação dos critérios elaborados aos padrões de cunhagem, uma vez que aqueles só recobrem a análise de construções lexicais e construções mais esquemáticas, sem, no entanto, tratar de níveis estruturais intermediários. Tal dificuldade decorre, em parte, do fato de o Constructicon, enquanto recurso, fazer um recorte sincrônico, instantâneo, da língua e propor uma abstração, um modelo, baseado nele. A dificuldade em modelar tais estruturas intermediárias em um ambiente como o Constructicon pode ser posta em perspectiva a partir da discussão de Bybee (2010) acerca do processo de *chunking* e de seu produto, os *chunks*.

Bybee (2010), em seu livro *Language, Usage and Cognition*, observa que o uso reiterado de sequências de palavras (ou *chunking*) pode dar origem a algumas construções da língua. Assim, aquilo que é frequentemente reiterado se convencionaliza, podendo dar origem a um padrão mais esquemático que, mesmo sendo um *chunk*, não é invariável. Um *chunk*, como observa a pesquisadora, tem sido evidenciado como o produto de um processo de enorme influência no sistema cognitivo, atuando na organização geral da memória, por isso tem como principal fator motivador a repetição. Estes estão presentes tanto na percepção quanto na produção e alguns efeitos interessantes são observados nos dois casos. No primeiro, observa-se a redução dos gestos articulatórios e, no segundo, a habilidade de se prever o que vem a seguir.

Quanto ao processamento das línguas, o *chunking* é responsável pela formação de sequências de palavras as quais a autora denomina como "prefabricadas" ou "*prefabs*", tais como *take a break*, *break a habit* e *pick and choose*. Segundo a autora, a língua é ao mesmo tempo estruturada e variável, ou seja, ela nos permite ser amplamente criativos e, ao mesmo tempo, a utilizamos a partir de sequências altamente frequentes que se reiteram cada vez mais e funcionam como alternativas convencionalizadas de se expressar uma ideia. É a partir dessa concepção que a pesquisadora discute a extrema relevância dos *prefabs*, tidos como qualquer expressão multi-palavra convencionalizada, os quais são armazenados e produzidos como *chunks* na língua.

Há outro ponto interessante a respeito dos *prefabs*, o de que "são convencionais no sentido de que eles são estabelecidos através de repetição no uso, mas eles não precisam ser altamente frequentes" (BYBEE, 2010, p.60). A autora justifica que, assim como aprendemos uma nova palavra com apenas algumas repetições, ou até mesmo uma única vez, no caso de falantes nativos, também é possível que possamos registrar um *prefab* após experienciá-lo poucas vezes. Essa definição deixa clara a dificuldade que tais elementos impõem aos sistemas de tradução por máquina, os quais são baseados em correspondências frequentes e estatisticamente relevantes de pares de estruturas linguísticas nas línguas-fonte e alvo.

A autora propõe ainda um *continuum* para os diferentes níveis de *chunks* presentes na memória. Segundo Bybee (2010), palavras que nunca ocorreram juntas não constituem um *chunk*, porém aquelas que tenham ocorrido ao menos uma vez podem, por sua vez, constituir um *chunk* mais fraco, uma vez que suas partes constituintes têm mais força que o todo, ocupando uma ponta do *continuum*. Do outro lado, estariam as palavras que coocorrem mais frequentemente na língua e que configurariam *chunks* mais fortes, tendo em vista que o todo teria mais força do que suas partes. Nesse sentido, os *prefabs* podem variar entre si através da força dos *chunks* que os constituem, o que é determinado pela frequência de ocorrência daqueles.

Embora variáveis, a produtividade de tais *chunks* é relativa. Segundo Bybee, a analogia seria o processo responsável por essa possibilidade de diferentes itens lexicais preencherem os *slots* de uma construção esquemática. A autora afirma que:

> Uma importante fonte de criatividade e produtividade na língua que permite a expressão de novos conceitos e a descrição de novas situações é a habilidade de expandir os *slots* esquemáticos em construções e preenchê-los com novos itens lexicais, sintagmas ou construções. (BYBEE, 2010, p. 57)[16]

16 "An important source of creativity and productivity in language that allows the expression of novel concepts and the description of novel situations is the ability to expand the schematic slots in constructions to fill them with novel lexical items, phrases or other constructions."

Tal afirmação retoma os achados cognitivistas sobre a capacidade criativa dos falantes, os quais são capazes de gerar novos usos a partir da esquematização dos *slots* de uma construção. A pesquisadora afirma que duas ou mais instanciações do *slot* na construção já autorizam a elaboração de uma estrutura mais esquemática, a qual se sobrepõe a suas instanciações. Contudo, a analogia também pode se dar no nível do constructo, das instanciações individuais, o que parece ocorrer no objeto de investigação desta tese.

A analogia é tomada como um processo oposto ao da produtividade baseada em regras, uma vez que "é fortemente baseada na similaridade entre itens existentes ao invés de regras simbólicas mais gerais" (BYBEE, 2010, p. 57). Segundo a autora, exemplificando um caso de analogia semântica, a Construção Resultativa investigada por Boas (2003), a qual apresenta o verbo *drive* junto a um adjetivo ou sintagma preposicional que expresse o sentido de loucura ou insanidade, habilitaria novas expressões que denotem sentidos próximos de *crazy* e, dificilmente, habilitaria a criação de uma nova expressão como "*It drives me happy*", devido ao seu afastamento semântico do sentido de "*It drives me crazy*".

Os produtos dos processos de analogia e *chunking*, por não serem sujeitos a regras gerais e nem necessariamente ocorrerem com muita frequência, representam um desafio aos modelos computacionais de língua, como o Constructicon, e aos sistemas de traduções estatísticos. Discuti-los e analisá-los, nesse contexto, é a proposta desta tese.

4 METODOLOGIA

4.1 Corpus

A constituição do banco de dados de exemplos para a construção se deu durante o curso de mestrado e teve como ponto de partida a introspecção dos pesquisadores envolvidos no projeto, a partir de um levantamento dos possíveis lexemas que preencheriam a posição de N1 na estrutura [N1 de [N2]], obtendo inicialmente a seguinte lista: *pilha, mar, montanha, enxurrada, porrada, avalanche, oceano, floresta, caminhão, galáxia, enchente, vendaval.* Diante da observação da profusão de tipos da construção em situações de fala espontânea em diversos meios, tal lista foi ampliada. Assim, com o intuito de oferecer uma descrição bastante abrangente do fenômeno investigado, deu-se continuidade à tarefa de elencar outros lexemas que pudessem ser instanciados pela construção, chegando ao número de 35 Nomes Quantificadores, que engloba também nomes que indicam pequenas quantidades.

Conforme descrito em Tavares (2014), o próximo passo consistiu na busca sistemática a partir de cada nome elencado no *corpus* CetenFolha, através da ferramenta de busca Sketch Engine (www.sketchengine.co.uk). O referido *corpus* contém 25 milhões de palavras em português brasileiro e tem como base o Jornal Folha de São Paulo. Já o Sketch Engine é um repositório de *corpora,* o qual pode ser acessado, dentre outras possibilidades, a partir das modalidades de busca *Word Sketch* ou *Concordance*. Na primeira modalidade, os resultados são exibidos em forma de listas de palavras com os contextos nas quais ocorrem, e, na segunda, há a apresentação de um ambiente para a instanciação do objeto linguístico. Para a investigação do nosso objeto de pesquisa, optou-se pela segunda modalidade – vide Figura 21. A informação acerca do contexto sintático no qual se observa o dado linguístico torna-se acessível devido à etiquetagem dos *corpora,* realizada pelo *parser* PALAVRAS (BICK, 2000).

A busca na ferramenta se deu a partir da estrutura **N1 de,** pois foram elencados apenas nomes que preenchessem N1, deixando o *slot* N2 aberto, tendo em vista que sua possibilidade de variação é significativamente maior que a de N1. É importante destacar que a especificação de N1 limitou a observação de outros potenciais nomes quantificadores, mas que sem tal especificação não seria possível realizar a busca. A Figura 21 ilustra a busca pela modalidade *Concordance* e a Figura 22 apresenta o resultado da mesma.

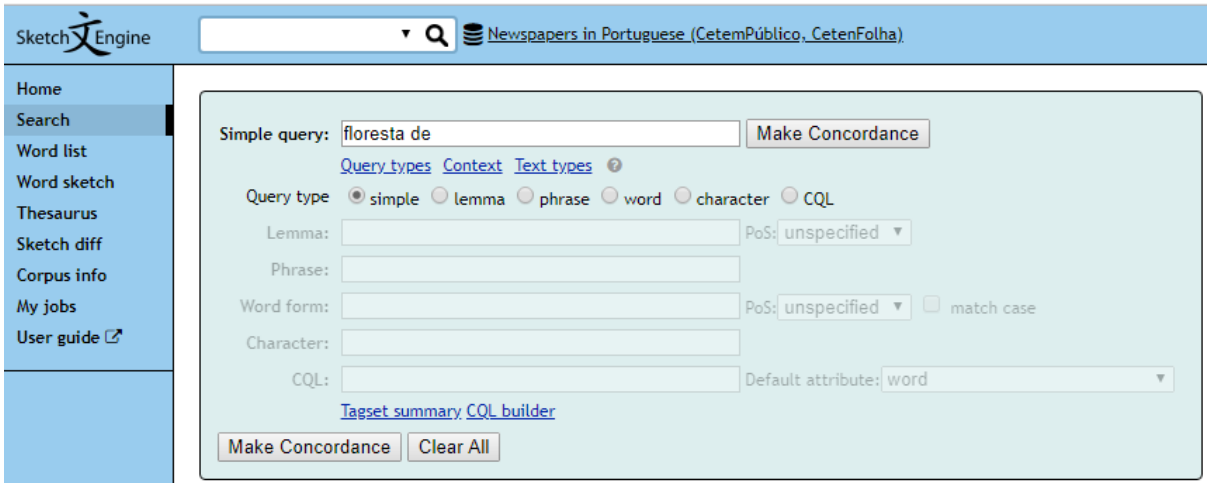


Figura 21: Busca através da modalidade Concordance

Figura 22: Resultado da busca realizada pela modalidade *Concordance*

Tal resultado não apresenta somente as instanciações da construção de quantificação indefinida, por isso passou-se à limpeza criteriosa dos dados, eliminando-se repetições e outros padrões identificados. Por exemplo, eliminaram-se os casos ambíguos, em que não se poderia determinar se o constructo equivaleria à construção de quantificação metafórica ou à literal, como a expressão *pilha de livros*[17], que, sem contexto suficiente, pode assumir ambas as interpretações.

A Tabela 1 exibe a lista de Nomes Quantificadores com os respectivos exemplos extraídos do *corpus* CetenFolha (TAVARES, 2014, p. 57):

17 A discussão acerca das construções literais e figurativas será feita no capítulo de análise.

	Types	Exemplos
01	Pilha	É uma insensatez por uma **pilha de motivos** que não cabem, todos, neste espaço.
02	Mar	Tristar já está voando desengonçado sobre um **mar de nuvens.**
03	Montanha	É a mesma contusão que fez a alemã tomar **montanha de analgésicos** no ano passado.
04	Enxurrada	A metafísica francesa das Luzes não passa de uma **enxurrada de falatório tedioso.**
05	Porrada	A juventude é uma garantia para uma **porrada de coisas.**
06	Avalanche	Fernando Henrique recebe diariamente **avalanche de adesões**, proporcional à sua posição nas pesquisas.
07	Oceano	No caso, Maradona ostentava um **oceano de razões**.
08	Floresta	Era aquele tipo de telejornal à base de «conteúdos», que ambicionava dimensionar a **floresta de notícias.**
09	Caminhão	É taxa suficiente para trazer um **caminhão de dólares** para o Brasil.
10	Galáxia	O cinema era uma **galáxia de gênios e heróis** e Veneza era um « luau » de lendas vivas.
11	Enchente	Um verdadeiro dilúvio, uma **enchente de cartas** inunda a redação
12	Vendaval	Entre a morte da mendiga e a tão aguardada pororoca fraterna... um **vendaval de amofinações** varre a vida do gêmeo bonzinho.
13	Bando	O Sport merece destaque: tem um **bando de craques**.
14	Penca	Tem **pencas de seguidores** no Rio de Janeiro, mas diz que o verdadeiro guru é o movimento.
15	Rio	O risco para os usuários é gastar **rios de dinheiro** na conta telefônica.
16	Enxame	No dia 12 de agosto um **enxame de Ovnis** invadiu uma cidade e aterrorizou seus 30 mil habitantes.
17	Inundação	O caso Simpson provocou verdadeira **inundação de debates** sobre a questão da violência doméstica nos EUA.
18	Bocado	Pegamo a CG, que era a única coisa que nóis tinha,um**bocado de removedor** e caímo na estrada.
19	Multidão	Estranho, um presidente da República pedindo apoio a uma **multidão de miseráveis** para salvá-los.
20	Pá	Sem os bordéis Freud teria uma **pá de casos** a menos e talvez tivesse se tornado arqueólogo.
21	Batalhão	... Sem falar no **batalhão defotógrafos e de cinegrafistas** que disputavam a imagem do dia.
22	Pelotão	... **pelotões de pesquisadores** pensavam que se aproximavam mais e mais do correto ao elaborarem montanhas de papéis.
23	Corja	E onde é que foi parar aquela **corja de cronistas e subcronistas de futebol** que só fazia espinafrar o homem?
24	Mundo	A Copa está chegando e... a seleção dos EUA é um **mundo de problemas**.
25	Tempestade	A esta altura, uma **tempestade de Pelés** pode ter desabado sobre a cabeça do candidato.
26	Onda	São Paulo assiste uma **onda de estréias teatrais** nesta sexta.

27	Dilúvio	Virou mania nos EUA, com direito a gibi, duas capas da “Rolling Stone”, um **dilúvio de merchandising**.
28	Punhado	... é ilusório supor que se vai, de fato, pôr ordem na casa por meio de um **punhado de operações** nas favelas.
29	Poço	“Nossa idéia é selecionar esse material e criar um grande centro de documentação”, diz, um **poço de histórias**.
30	Pingo	“Somos a melhor marca, temos os melhores produtos...” diz, sem um **pingo de modéstia**.
31	Gota	Longe dos estúdios de TV, embolsa o dinheiro do contribuinte sem liberar uma **gota desuor** por ele.
32	Ponta	Apesar de se dizer satisfeito com o empate, o técnico da Suécia não escondeu uma **ponta de decepção**.
33	Pitada	O filme é uma espécie de documentário. Só que com imagens lindas e uma **pitada de ficção**.
34	Dedo	O TSE adiou para depois do primeiro turno uma decisão a respeito, mas ninguém com um **dedo de juízo** acredita que qualquer um dos dois possa ser condenado.
35	Fiapo	Apesar da idade, não perdeu um **fiapo de sua voz aveludada**, talvez a mais aconchegante que o jazz já conheceu.

Tabela 1: Exemplos de instanciações do padrão de quantificação

Em relação ao *corpus Centenfolha*, a escolha por esta fonte se deu pelo fato de este ser um *corpus* tratado e disponível no *Sketch Engine*, contudo o mesmo apresenta apenas textos do gênero jornalístico, limitando a análise dos dados a um único ambiente discursivo. Do mesmo modo, compromete-se a análise a respeito da frequência de ocorrência dos tipos que instanciam a construção, uma vez que o gênero jornalístico favorece a ocorrência de alguns constructos em detrimento de outros. Isso, porque algumas expressões muito recorrentes no uso linguístico são pouco frequentes em nossa base de dados por serem de natureza informal (TAVARES, 2014). Tendo identificado tais limitações, durante o curso de doutorado, houve a expansão de tal banco de dados, buscando a constituição de um conjunto de dados diversificado. Constitui-se assim um *corpus* oral a partir dos dados extraídos do NURC-RJ e do Corpus do Português, e um *corpus* do gênero ficcional a partir de buscas no Corpus do Português, somando tais dados então ao *corpus* escrito/jornalístico já apresentado anteriormente.

A Tabela 2 apresenta o universo total de dados atualizado, por ordem de frequência de ocorrência dos N1s.

01	Punhado	68
02	Avalanche	59
03	Montanha	55
04	Bocado	51

05	Multidão	49
06	Ponta	46
07	Mar	43
08	Batalhão	29
09	Pitada	28
10	Bando	26
11	Enxurrada	24
12	Onda	20
13	Pilha	13
14	Mundo	12
15	Penca	11
16	Rio	10
17	Pá	9
18	Dedo	9
19	Poço	8
20	Gota	8
21	Fiapo	6
22	Caminhão	6
23	Inundação	6
24	Oceano	5
25	Tempestade	4
26	Dilúvio	4
27	Galáxia	4
28	Floresta	4
29	Porrada	4
30	Vendaval	3
31	Enxame	3
32	Pingo	3
33	Pelotão	1
34	Enchente	1
35	Corja	1

Tabela 2: Frequência de ocorrência dos tipos licenciados pela construção

Nosso conjunto de dados é composto por 623 ocorrências, sendo 100 dessas ocorrências referentes à expressão de pequena quantidade, e as demais correspondem à grande quantidade. Já o número de *types*, nomes que instanciam N1 no padrão, é de 35.

4.2 Análise Baseada em Frames

A análise da correlação entre os elementos constituintes do padrão de quantificação baseou-se em frames, tendo em vista nosso aporto teórico e também o objetivo final deste trabalho de incluir tal padrão num Constructicon. Foram identificados 8 frames evocados pelos 35 nomes que funcionam como quantificadores no padrão. A distribuição desses nomes quantificadores (N1) para cada frame identificado é representada na Tabela 3:

Agregados	corja, pelotão, batalhão, multidão, enxame, bando, penca, pilha
Medida_por_ação	bocado, punhado, pitada
Locais_naturais	mar, montanha, oceano, galáxia, rio, floresta, mundo
Clima	tempestade, dilúvio, onda, inundação, avalanche, enxurrada, vendaval, enchente
Contêiner	pá, poço
Parte_todo	dedo, fiapo, gota, pingo, ponta
Veículo	caminhão
Impacto	porrada

Tabela 3: Distribuição dos N1s entre os frames evocados

Diante da identificação de tais frames, foram investigados então os frames evocados por N2[18], a partir da análise dos contextos de uso, como evidencia a amostra na Tabela 4. O propósito desta análise foi verificar possíveis correlações entre os frames de N1 e N2, apontando preferências ou restrições no padrão de quantificação. Assim, para cada frame de N1, foram identificados os frames evocados por N2, de modo a estabelecer uma correspondência entre os frames. Por exemplo, o frame Parte_todo, evocado por *ponta, gota, fiapo, pingo* e *dedo*, apresenta N2s que evocam frames como: Expressar_publicamente, Propriedades_naturais, Propósito, Cogitação e Consciência – vide Tabela 4. Porém, como será demonstrado em detalhes no capítulo de análise, tal investigação parece oferecer resultados bastante conflitantes, de modo que se torna difícil a identificação de padrões para certos N1s e frames, o que evidencia a grande variedade e distinção entre os nomes que ocorrem no padrão.

18 Foram analisados apenas o conjunto de N2s presentes no corpus original, provenientes de dissertação de mestrado, uma vez que a expansão do banco de dados se deu após tal investigação.

	Frame_N1	Frame_N2	SENTENÇAS
2	Parte-todo	Expressar_publicamente	Apesar da idade (faz 69 anos em setembro) , não perdeu um **FIAPO DE** sua voz aveludada , talvez a mais aconchegante que o jazz j
3	Parte-todo	Propriedades_naturais	A não ser que Mazinho faça chover , diz o técnico . Olho para o céu azul , azul , sem um **FIAPO DE** nuvem , e volto para o hotel , cert
4	Parte-todo	Propósito	O técnico Palhinha classificou a derrota como humilhante. «O time não apresentou **PINGO DE** de objetividade. Desse jeito, será difí
5	Parte-todo	Cogitação	Seria didático um **PINGO DE** reflexão sobre nossa história recente .
6	Parte-todo	Consciência	No seu lugar , entrou sua antítese -- Biro , aquele lateral violento e sem **PINGO DE** imaginacão . Resultado : exatamente de seu lado
7	Medida_por_ação	Previsão	Ou seja , melhor organização de jogo , uma **PITADA DE** imprevisibilidade nos passes e nos lançamentos , mais acurada , uma visão d
8	Medida_por_ação	Sentimento	« Estou pasmo com as críticas que vários ex-ministros têm feito ao Plano FHC . Atribuo isso a uma **PITADA DE** inveja , porque nenhm
9	Medida_por_ação	Texto	O filme é uma espécie de documentário . Só que com imagens lindas e uma **PITADA DE** ficção .
10	Medida_por_ação	Foco_no_estímulo	Afinal de contas, felicidade e romance (com uma **PITADA DE** humor) são palavras conhecidas da diretora Nora Ephron .
11	Medida_por_ação	Existência	É um mundo de mutantes superpoderosos, mas com uma **PITADA DE** realismo .
12	Medida_por_ação	Propriedade_mental	Ontem o governo Itamar Franco adicionou mais uma **PITADA DE** loucura neste caldeirão de insegurança ao determinar a intervençã
13	Medida_por_ação	Foco_no_estímulo	As transformações de Carmen dão uma **PITADA DE** humor à temática espiritualista da novela que , para a atriz , é bastante oportun
14	Medida_por_ação	Fazer_barulho	Raimundos é , então , um power trio com uma **PITADA DE** triângulo em aqui , uma levada de baião ali e o hardcore em primeiro pla
15	Medida_por_ação	Criar	Falta uma dose de irreverência que é o sal do futebol . Falta uma **PITADA DE** improvisação, de fantasia, que é apanágio da cor .
16	Medida_por_ação	Sentimento	Schoti coloca uma **PITADA DE** otimismo em sua receita . Para ele , as Bolsas deverão registrar altas expressivas aos primeiros sinais
17	Medida_por_ação	Ensinar_educação	As perigosas relações professor-aluno, temperadas com **PITADA DE** educação sexual, estão dando manchetes quase diárias nos tabl
18	Medida_por_ação	Foco_no_estímulo	Contando momentos íntimos de sua vida , os livros misturam ficção e realidade , com **PITADAS DE** bom-humor .

Tabela 4: Amostra da pesquisa acerca dos frames evocados por N2

4.3 Teste de Julgamento de Aceitabilidade por Falantes Nativos

Com o intuito de analisar o julgamento de falantes nativos do português a respeito das possibilidades de combinações de N1 e N2 (variando-se apenas o N2), realizou-se um teste de julgamento de aceitabilidade de algumas expressões de quantificação indefinida. O motivo que levou ao desenvolvimento deste experimento foi a necessidade de se avaliar o potencial produtivo da estrutura de quantificação e o grau de aceitação dos novos usos, ou seja, de novas combinações para N1 e N2. O experimento é constituído por expressões frequentes no *corpus* e expressões novas criadas a partir destes usos frequentes, para que se tenha uma análise comparativa entre esses dados. Como já mencionado, a variação na estrutura se deu apenas no *slot* de N2.

Em relação aos procedimentos adotados, desenvolveu-se um questionário on-line, o qual foi acessado pelos voluntários e respondido. Ao todo, foram apresentadas a cada voluntário 30 sentenças a serem avaliadas em uma escala de 1 a 5, no que tange à possibilidade de ocorrência na língua. Como a pesquisa foi toda automonitorada via internet, não houve contato direto entre o pesquisador e os pesquisados. No entanto, elegemos para esta pesquisa uma ferramenta conhecida como segura e buscamos manter sigilo total em relação à identificação dos participantes, visto que nosso objetivo se centra apenas na informação disponibilizada a partir da análise das sentenças.

A página apresentada na Figura 23 exemplifica de que o modo o experimento se mostra ao participante da pesquisa, com as instruções de sobre como responder ao questionário e com a escala de 5 pontos, a qual permite uma avaliação subjetiva em termos da adequação ou não de determinada sentença no uso do português:

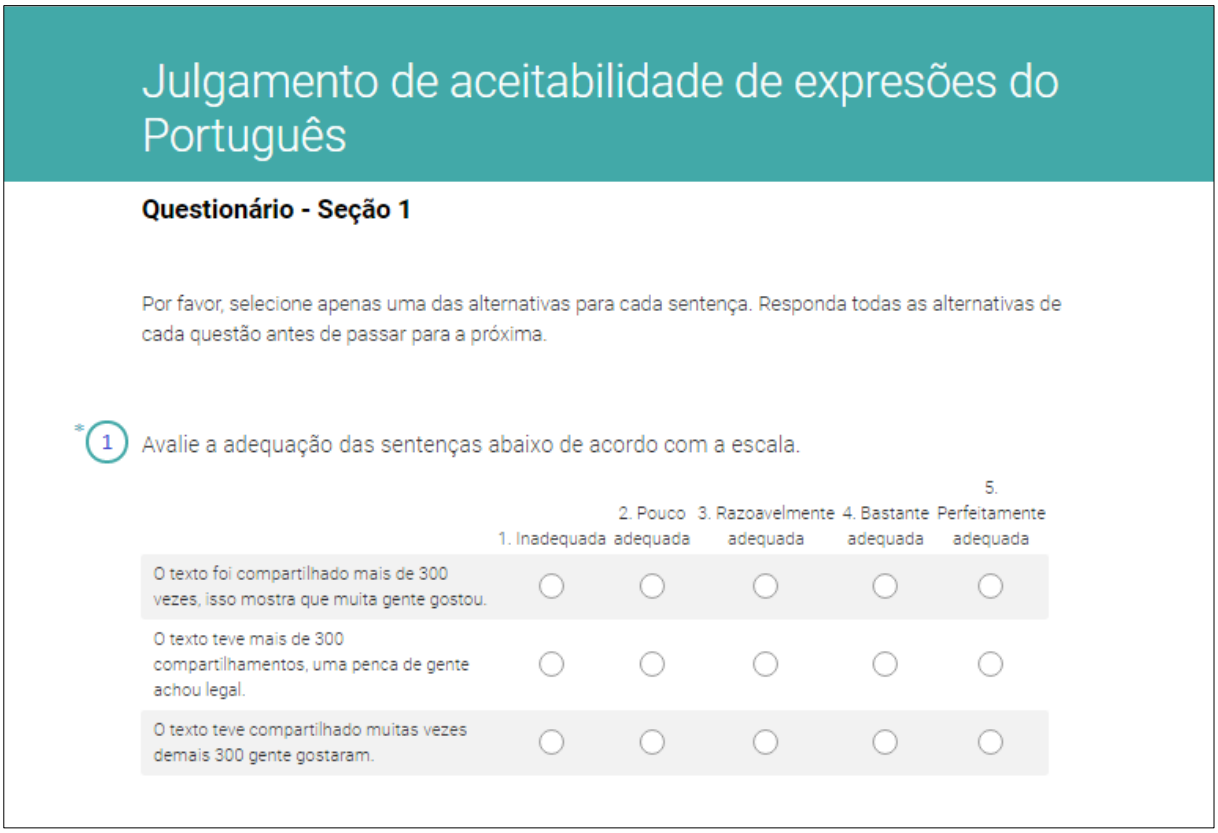


Figura 23: Página apresentada ao voluntário para julgamento das sentenças

Para cada questão foram apresentadas: 1 sentença alvo, com a expressão de quantificação indefinida licenciada pelo padrão N1 de N2 – *O texto teve mais de 300 compartilhamentos, uma* ***penca de gente*** *achou legal* –, 1 sentença com significado próximo, porém que não contenha a estrutura referida – *O texto foi compartilhado mais de 300 vezes, isso mostra que* ***muita gente*** *gostou* – e 1 sentença distratora, com desvios gramaticais, de ordem dos constituintes ou apenas incoerente – *O texto teve compartilhado* ***muitas vezes demais*** *300 gente gostaram*.

Os resultados desta pesquisa podem corroborar algumas hipóteses a respeito de restrições presentes na estrutura que governam as combinações de N1 e N2 ou podem indicar a fragilidade destas restrições e a possibilidade de expansão da estrutura no uso. Tal analise será apresentada mais adiante, na seção 5.3, do capítulo de análise.

5 REMODELAGEM DO PADRÃO DE QUANTIFICAÇÃO INDEFINIDA

Neste capítulo será analisada a estrutura binominal de quantificação indefinida à luz das correntes teóricas apresentadas nos capítulos anteriores, de modo a verificar a aplicabilidade da teoria de Kay (2013), a respeito dos padrões de cunhagem, a esta estrutura. Para tanto, inicialmente serão apresentados, em ordem cronológica, outros estudos que se debruçaram sobre o mesmo fenômeno.

5.1 Abordagens Anteriores para o padrão de quantificação N1 de N2

5.1.1 O estudo de Monte de e Chuva de

Brodbeck (2010), em sua tese de doutorado, investigou dois subtipos da construção nominal de quantificação do português, *monte de* e *chuva de,* e, dentre os principais achados desta pesquisa, destacam-se: a emergência da construção através do uso para o sistema de quantificação e o fenômeno do desencontro sintático-semântico presente na mesma.

A autora não descarta a existência de outras expressões provenientes desta mesma estrutura e afirma que a motivação da profusão de novas formas seja uma demanda discursiva. Nesse contexto, o falante pode optar por uma expressão de quantificação convencional como *monte de* ou por expressões menos convencionais, tais como *enxurrada, porrada, penca, bocado,* as quais parecem atuar na construção de maneira mais expressiva. Além disso, o falante tem a sua disposição quantificadores que estabelecem entre si uma relação de gradação: *monte, montinho* e *montão*, assim como se observa, de maneira menos explícita, em *pingo*, *chuva* e *tempestade*. Tal observação reitera a importância de se avaliar o uso de tais expressões, uma vez que a escolha dentre esses quantificadores será determinada pela necessidade comunicativa do falante.

Este estudo também revelou questões importantes sobre a evolução diacrônica das expressões investigadas. Com relação a *monte*, observa-se que os primeiros usos do nome associado a um complemento, tais como em: *monte de terra* e *monte de pedras,* surgem no século XIV, e já no século XV surgem os usos metafóricos. Segundo Brodbeck:

> o que ocorre é a reorganização semântica da expressão motivada pela metáfora (primária) MAIS É PRA CIMA, que recruta um esquema imagético piramidal, evocado pela estrutura físico-visual de monte: orientação vertical PARA CIMA a partir de uma base alargada (de empilhamento de entidades físicas). (BRODBECK, 2010, p.105)

Tendo em vista tal reorganização, o nome *monte* passa a expressar *muita quantidade.* Com a convencionalização do quantificador ao longo dos séculos*,* aumenta-se sua produtividade de type em relação a N2, assim são diversas as possibilidades de combinação do nome com entidades, sejam elas massivas ou contáveis, na forma plural ou singular.

No que se refere a *chuva*, o nome emerge como expressão quantificadora a partir do século XVI, tornando-se de fato frequente no século XIX (BRODBECK, 2010, p. 110). A pesquisadora identifica a presença da metáfora Movimento Massivo de Fluidos é Quantidade neste quantificador, a qual também integraria outras expressões de quantificação, como *enxurrada, onda, avalanche,* dentre outras. Além da metáfora, também é relevante que se considere o perfil sensório-motor da expressão, que envolve movimento massivo vertical, assim, o perfil sensório-motor de *chuva de pedras* motivará o surgimento de chuva de perguntas, por exemplo.

A pesquisadora também identifica a coerção dos elementos constituintes da construção, tendo em vista a presença do fenômeno do **desencontro/*mismatch*** linguístico. Tal fenômeno é entendido como um conflito entre forma e significado, por isso é traduzido como desencontro sintático-semântico, de modo que as expressões nas quais se observam este conflito são "incongruentes em relação aos padrões mais gerais de correspondência na língua" (FRANCIS & MICHAELIS, 2003, p. 2).

Como exemplifica Brodbeck (2010), nas expressões de quantificação, ocorre a harmonização do Atributo de Quantificação (definido ou indefinido) ao Atributo de Perfilamento do Nome (contável ou massivo), de modo que se tem expressões como: *Ele tomou* ***muito café*** (Incontável Indefinida) e *Comprei* ***dois livros*** (Contável Definida). Mas também é possível que tal harmonização não ocorra como se espera, diante de expressões como: *Me vê* ***dois cafés*** *e* ***três leites****, Compra* ***duas águas*** *pra gente,* o que sugere a coerção do Determinante sobe o Nome. Assim, *café,* antes massivo, passa a nome contável por coerção do Determinante que possui dimensões enumerativas bem definidas. Em relação às expressões *monte de* e *chuva de* tal processo ocorre de modo distinto. Observem-se as expressões em (50-51):

(50) Chuva de granizo

(51) Chuva de protestos

Enquanto em chuva de granizo há uma combinação harmoniosa entre o núcleo *chuva* e seu elemento constituinte *de granizo;* em *chuva de protestos* o mesmo não ocorre, pois não é possível se conceber uma chuva (literal) de protestos. Aqui atuaria o processo da coerção,

através da reanálise morfossintática de *chuva*, que passará de Núcleo a Sintagma Modificador, isto é, o nome mudará seu status de entidade modificada para nome quantificador. A coerção teria sido realizada pelo complemento (de protestos) sobre o núcleo (chuva).

Assim como propõe Bybee (2010), Brodbeck sugere que uso frequente da expressão *monte de* como indicador de quantificação indefinida tenha se convencionalizado na língua, dando origem à esquematização [N1 de [N2]]. A partir do estabelecimento de uma estrutura de quantificação no PB, inicia-se o processo criativo da linguagem, sendo possível admitir novos nomes atuando como quantificadores neste padrão. Daí a admissão também de *chuva* como um quantificador. Tendo a analogia como base do processo de expansão da construção, surgem novas expressões como: *montanha de, montão de, enxurrada de, tempestade de,* dentre tantos outros os quais são objeto de análise da presente tese.

5.1.2 Construções Binominais Quantitativas e os processos de extração de porção e multiplexização

Alonso (2010) analisa, por uma perspectiva baseada no uso, quatro construções binominais quantitativas que se formam pela estrutura *um N1 de N2*, evidenciando que tais construções se diferenciam não só em termos formais e discursivos, mas também nos processos cognitivos mais complexos inerentes à interpretação da quantificação. Tais processos são observados nas construções exemplificadas em (52-55):

(52) Num N1 de Nsing2 (um quilo de farinha)
(53) Num N1 de Npl2 (um grupo de pessoas)
(54) Art Indef N1 de Nsing2 (um pouco de manteiga)
(55) Art Indef N1 de Npl2 (uma cacetada de crianças)

Observa-se que as construções compartilham uma sintaxe básica, a estrutura **um N1 de N2**, e, segundo a autora, estas construções estariam inter-relacionadas numa rede de construções através laços de herança. Já a análise de cada estrutura se deu a partir dos seguintes tópicos: i) a categoria gramatical de *um,* como Numeral ou Artigo Indefinido; ii) as propriedades semânticas de N1, envolvendo sua referência mais ou menos determinada; iii) o grau de gramaticalização das estruturas; iv) as propriedades semânticas de N2, principalmente em relação a sua configuração contínua/incontável ou discreta/contável e, consequentemente, ao seu número, singular ou plural.

Outro ponto central para sua investigação foi a identificação dos processos cognitivos de extração de porção ou unidade e de multiplexização, relacionados, respectivamente, às estruturas em (52-53) e (54-55). Observa-se que a pista sintática para tais processos se encontra no número do N2, uma vez que as estruturas relacionadas ao processo de extração de porção ou unidade apresentam N2 singular, enquanto as estruturas relacionadas ao processo de multiplexização apresentam N2 plural. Conforme observa Alonso (2010), isso se explica pelo fato de que o primeiro processo possui a função de delimitar um referente tomado a priori como não-delimitável, ou seja, de quantificar elementos incontáveis, como se observa pela expressão *quilo de farinha;* desse modo, é compreensível que N2 (*leite*) apresente-se no singular, tendo em vista sua configuração (ao menos a princípio) incontável. Ao passo que o segundo processo consiste na multiplicação do referente, que será, necessariamente, uma entidade contável; como ocorre em *grupo de pessoas*, no qual o N2, que poderia se apresentar de forma uniplexa (pessoa), foi multiplicado pelo modificador *grupo* e, portanto, apresenta-se na forma plural. Assim, a representação da estrutura de quantificação com N2 singular determina que tal estrutura seja mais restrita ao uso de N2 singular, embora, no que concerne à estrutura com N2 plural, não há restrição quanto ao número de N2, que, por ser discreto, pode se apresentar em ambas as formas (mesmo que a forma plural seja a preferida).

De acordo com a pesquisadora, os processos cognitivos mencionados estariam mais explícitos nas construções com Numeral, as quais envolveriam noções mais objetivas (ou até mesmo mais precisas de quantificação). As construções com Artigo Indefinido, por sua vez, por corresponderem a interpretações mais subjetivas da quantidade, redimensionariam tais processos para o âmbito das avaliações subjetivas.

A construção com Numeral e N2 singular – Num N1 de N2sing –, corresponderia a uma estrutura clássica já bastante descrita na literatura, a classe dos partitivos (*um litro de leite, uma xícara de café, uma colher de sopa*). Alonso (2010), no entanto, oferece uma perspectiva construcionista à estrutura de quantificação e evidencia a relação desta a outras menos centrais, como as construções com Artigo Indefinido. Tais construções também correspondem ao objeto de investigação do presente trabalho e, como observa a pesquisadora, elas indicam "uma dimensão avaliativa mais subjetiva, em que o falante julga determinada quantidade dentro de uma escala quantitativa" (ALONSO, 2010, p. 86). Dentre os subtipos instanciados pelo padrão com N2 singular estão os nomes: *um pouco, um pouquinho* e *um bocadinho*; para o padrão com N2 plural a pesquisadora investigou os subtipos: *um bocado, um monte, uma porção, um montão, um bando, um mundo*. O primeiro padrão estaria atrelado ao processo de extração de porção ou unidade e teria a função de avaliar a quantidade de N2 em um nível baixo na escala da

quantidade (*João comeu um pouco de queijo)*. O segundo padrão, associado ao processo de multiplexização, estaria envolvido na avaliação de N2 em um nível alto na escala (*Eu sei um bocado de coisas sobre você*).

Como afirma Alonso (2010), as construções com *um pouco de N2* e *um pouquinho de N2* seriam mais cristalizadas na língua, e talvez este seja um dos pontos centrais que diferenciam a análise da autora da presente tese, uma vez que os subtipos aqui investigados fariam parte de um grupo de nomes relativamente frequentes no uso, porém não amplamente convencionalizados, como *dedo de N2*, *fiapo de N2, ponta de N2*, etc. Outro ponto divergente é a não formalização do Artigo Indefinido como parte integrante do padrão, isto porque em nossa base de dados observaram-se instâncias da construção com outros Determinantes ou até mesmo sem qualquer Determinante acompanhando as expressões de quantificação.

Por fim, a autora estabelece uma relação entre tais padrões de quantificação e construções de modificação de grau, do tipo *um N ADJ* (*um pouco triste*, *um bocado cansado*), de modo que estas seriam licenciadas por aquelas. De modo paralelo, como será evidenciado na seção 5.1.3 deste capítulo, Tavares (2014) demonstrou que o polo discursivo da construção de quantificação, ao menos no que se refere aos subtipos investigados, já envolve a noção de grau/intensidade, demonstrando a correspondência entre as propostas.

5.1.3 A Construção de Quantificação Binominal no Espanhol

O trabalho de Verveckken (2012) também se mostrou relevante para esta tese, pois trouxe reflexões acerca da Construção Binominal no espanhol, a partir de uma abordagem centrada no uso, a qual buscou explicar a mudança linguística e a frequência de ocorrência de algumas expressões de quantificação.

A autora identificou que a quantificação no padrão [N1 de [N2]] pode ser feita por nomes das seguintes categorias: os contêineres (barril, tonel), as configurações de massa (monte, pilha), os coletivos (bando, multidão), os fenômenos naturais (enchente, oceano, etc.) e as nominalizações denotando agrupamento ou extensão (amontoamento). Tais nomes, segundo a autora, possuem implicaturas escalares em adição ao seu significado lexical. Assim, apenas nomes que expressem quantidades expressivas preencheriam o *slot* N1, como: montón (montão), hatajo (bando), pila (pilha), racimo (cacho), tropel (multidão), letanía (ladainha), alud (avalanche), aluvión (enchente) e barbaridad (barbaridade). E, de modo semelhante ao que propomos neste trabalho, nomes abstratos que já instanciem a noção de quantidade (quilo, quantidade, número, parte, maioria) não são recobertos pelo estudo de Verveckken (2012).

A análise diacrônica da pesquisadora evidencia questões semelhantes às descobertas por Brodbeck, a respeito de *monte,* que no estágio de pré-gramaticalização teria a função de núcleo do sintagma, enquanto o sintagma preposicional seria o modificador, como se observa em (56-57):

(56) Se sentaron a reposar sobre um **montón de piedras**
(Se sentaram a repousar sobre um montão de pedras).

(57) Silvia y François se habían conocido em um café de Paris hacía um **montón de ãnos**.
(Silvia e François haviam se conhecido em um café de Paris fazia um montão de anos).

Observa-se que no exemplo (56) o N2 (*pedras*) atua como modificador do núcleo N1 (*montão*), indicando as entidades constituintes de N1. Em (58), *anos* não são interpretados como as entidades constituintes de montão, mas considera-se que *montão* passe a modificar *anos*, ou seja, N1 deixa de ser núcleo e passa a modificador (quantificador) de N2. Assim como se observou no trabalho de Brodbeck (2010) para a construção do PB, esta mudança corresponde aos usos gramaticalizados da construção de quantificação. Em uma expressão como *Chuva de verão*, por exemplo, *chuva* é o núcleo sintático da expressão, o qual é modificado por *de verão;* já em *chuva de críticas*, *chuva* é que se torna o modificador de *críticas* (TAVARES, 2014). Assim, em relação às propriedades que envolvem a gramaticalização da construção, Verveckken afirma que:

> A mudança de status de núcleo de N1 para quantificador (premodificador) de N2 envolve uma reanálise morfossintática em termos da estrutura interna do sintagma binominal: de [Det + NQnúcleo + [de + N2] para [[Det + NQ + de] +N2núcleo] (VERVECKKEN, 2012, p. 436, tradução nossa).[19]

Um ponto fundamental deste trabalho é a constatação de que as propriedades dos nomes quantificadores em seu sentido literal são mantidas na construção, adicionando algo além da noção de quantidade, como se observa pela sentença (58):

(58) Su discurso desató un **alud de críticas**.
(Seu discurso desatou uma avalanche de críticas)

19 "The shift from head-status of N1 to quantifying (/premodifying) N2 involves a morphosyntactic reanalysis in terms of the internal structure of the binominal syntagm: from [Det. + QNhead + [de + N2]] into [[Det. + QN + de] + N2head]"

Neste exemplo, pode-se argumentar a respeito da ativação de algumas propriedades de uma avalanche literal, uma vez que se interpreta *críticas* como um fenômeno dinâmico, surpreendente (VERVECKKEN, 2012, p. 433). A descoberta da permanência de tais propriedades, ou como denomina a autora, Permanência da Imagem Conceptual (PIC), é extremamente importante para o presente trabalho, como identificou Tavares (2014), uma vez que a PIC governa as restrições de preenchimento dos *slots* da construção e a combinação entre N1 e N2. Ou seja, segundo essa abordagem, N2 precisa se ajustar à imagem ativada por N1, que já apresenta configuração própria. O exemplo em (62) ilustra de que maneira *alud* (avalanche) sugere uma propriedade típica de uma avalanche enquanto fenômeno natural, de modo que o N2 é quantificado e também conceptualizado desta maneira.

(59) Me siento sepultado y cautivo en un **alud de hipocresías, estratégias, tácticas** (...)
(Me sinto sepultado e cativo em uma avalanche de hipocrisias, estratégias, táticas...)

Conforme observa Verveckken (2012), a PIC é importante porque oferece uma conceptualização particular para a entidade quantificada. Além disso, segundo a autora:

> O fato de que a função comunicativa ou pragmática dos quantificadores binominais depende da PIC, conduz à hipótese de que os NQs provavelmente não se tornarão quantificadores completamente transparentes como montón de (VERVECKKEN, 2012, p. 455).[20]

Verveckken admite que a construção investigada é bastante produtiva, assumindo que qualquer nome que tenha inferências escalares possa preencher a posição de N1. No entanto, esta questão pode ser mais complexa, pelo fato de não se saber ao certo o que se entende por inferências escalares. No português, um nome como *caminhão* funciona frequentemente na estrutura de quantificação, e é possível observar que tal entidade apresenta uma noção escalar pelo fato de ser utilizado como unidade de medida (um caminhão de areia numa obra por exemplo). Mas outros veículos como avião, navio, caminhonete, dentre outros, também poderiam ser entendidos nesses termos, porém não são nomes quantificadores frequentes no português. Posto isso, é preciso delimitar os critérios de preenchimento do padrão de quantificação e rever o pressuposto de que o padrão de quantificação seria produtivo.

20 "The CIP is the key factor in the productivity of the binominal quantifier construction: byyielding an individual conceptualization of the group of entities expressed byN2, the binominal quantifier construction provides the Spanish native speakerwith a useful tool for expressive quantification. The fact that the pragmatic or communicative function of binominal quantifiers hinges on the CIP, leads to the hypothesis that QNs are not likely to become completely bleached quantifiers like montón de."

Retomando o estudo de Brodbeck, observa-se que este possui alguns pontos em comum com o de Verveckken. No primeiro, verificou-se que *monte* é um quantificador muito frequente na língua e em um estágio avançado de gramaticalização. Concomitantemente, o segundo demonstrou que *montón* foi o primeiro a funcionar como quantificador e a ter um crescimento radical na frequência de ocorrência, facilitando a associação sistemática da construção binominal com a expressão de quantidade. Ou seja, o uso frequente de *montón de* como quantificador teria sido o responsável pelo surgimento do padrão "N1 de N2" de quantificação. A parir daí, outros NQs menos frequentes são autorizados, via analogia, a participar da construção. Além disso, a falta de especificidade semântica de *montón de* é também um forte argumento para o papel prototípico que este assume, pois com um conteúdo semântico tão baixo ele pode substituir todos os NQs.

Muitas dessas questões levantadas por Verveckken, em relação aos dados do espanhol, se mostram também relevantes para a descrição e análise da construção de quantificação do português e serão devidamente retomadas no próximo capítulo.

5.1.4 O estudo das expressões de quantificação indefinida do Português

Motivada pela pesquisa de Brodbeck (2010) acerca dos quantificadores *monte* e *chuva*, Tavares (2014) dá continuidade ao estudo de outras expressões de quantificação indefinida que se formam a partir da estrutura [N1 [de N2]]. Tratam-se de dois subtipos, a saber, N1 de N2 singular – que se refere aos quantificadores de pequena quantidade – e N1 de N2 plural – apresentando nomes que indicam grande quantidade –, os quais licenciam expressões metafóricas que veiculam a noção de quantidade bastante expressiva – seja para mais ou para menos. De modo diferente do que postula Alonso (2010), aqui o padrão não é representado necessariamente com o Artigo Indefinido, pois sendo o mesmo opcional (embora bastante frequente), entende-se que o padrão sintático da construção seria melhor descrito pelo esquema N1 de N2. As sentenças (60-63) ilustram a variedade de determinantes que ocorrem com a construção:

(60) No Brasil, **<u>qualquer</u> punhado de enquetes feitas na TV**, com resposta instantânea via telefone, virou TV interativa. (CetenFolha)

(61) Ele afirmou que vai mandar milícias para « derrotar **<u>este</u> bando de foras-da-lei**. (CetenFolha)

(62) Entre ele e seu destino, curvas fatais vêm precedidas **pelo mar de buracos** e pela falta de sinalização. (CetenFolha)

(63) Falta-lhe o ar, interrompendo **a enxurrada de palavras**. (CetenFolha)

Tavares (2014) observa que a instanciação de N1 no singular é a mais comum, mas que a possibilidade de haver ocorrências na forma plural impossibilitaria a inclusão do artigo indefinido na expressão, como evidenciam as sentenças (64-67):

(64) Mas não é a diversidade do conceito que faz do Brasil um país de heróis? É, creio, porque temos mesmo **multidões de heróis**. (CetenFolha)

(65) Tem **pencas de** seguidores no Rio de Janeiro, mas diz que o verdadeiro guru é o movimento. (CetenFolha)

(66) O BC está realizando leilões de **rios de compra** no mercado de dólar comercial ...(Cetenfolha)

(67) Reabriu ontem com todo o entusiasmo, na certeza de que uma África do Sul reconhecida como democrática atrairá **pilhas de investimentos estrangeiros. (CetenFolha)**

O estudo recobriu um grupo de 35 nomes comuns do português que preenchem a posição de N1 no padrão de quantificação e passam a funcionar como Nomes Quantificadores. Como mencionado anteriormente, tratam-se de dois subtipos, divididos em termos da expressão de pequena ou grande quantidade. A Tabela 5 apresenta todos os nomes analisados por Tavares (2014, p.70):

Grande quantidade	Pequena quantidade
enxurrada, avalanche, enchente, tempestade, dilúvio, vendaval, inundação, onda, mar, montanha, oceano, rio, floresta, mundo, galáxia, caminhão, poço, pá, bando, batalhão, penca, multidão, pelotão, corja, enxame, montão, pilha, porrada, (bocado e punhado).	pingo, gota, ponta, dedo, pitada, fiapo, (bocado e punhado).

Tabela 5: Nomes Quantificadores de grande e pequena quantidade

Como se pode observar, os nomes *bocado* e *punhado* encontram-se entre parênteses e isso se justifica pelo fato de que ambos flutuam entre as interpretações de quantificadores de pequena e de grande quantidade, como evidenciou a análise do *corpus*. Considerando-se uma escala quantitativa, a pesquisadora observa que tais NQs distanciam-se consideravelmente nesta escala e a interpretação de uma quantidade muito grande ou muito pequena parte da avaliação subjetiva do falante do que seria uma quantidade considerada normal ou razoável. Assim, seu estudo considera relevante para esta construção a noção de grau intensivo, postulada por Silva (2008) de modo que:

> a quantidade considerada "normal" se situaria numa posição mediana da escala de intensificação, enquanto a quantidade representada pelos subtipos da CBQI seria posicionada em níveis inferiores ou superiores da escala. (TAVARES, 2014, p. 71)

Segundo Silva (2008), há uma relação muito próxima e cognitivamente motivada entre intensidade e quantidade, como pode-se perceber pela utilização dos mesmos termos para se expressar ambos os conceitos – muito, bastante, mais ou demais –, por isso a teoria de que exista um tipo de intensificação do grau quantitativo. Tavares (2014) assume, portanto, a noção intensificadora na construção de quantificação indefinida sobretudo quando comparada ao quantificador *muito.*

(68) Ela trouxe **muitos** livros para a escola

(69) Ela trouxe uma **porrada/penca/pilha** de livros para a escola.

Observa-se que o uso dos quantificadores em (69) intensifica a noção de grande quantidade de *livros*, comparando-se esta sentença a (68). Assim, numa escala de intensidade, considera-se que *muito* esteja numa direção ascendente, porém muito inferior a *porrada, penca* e *pilha.* E, assim como identificou Brodbeck (2010), Tavares (2014) argumenta a favor da distinção no grau de intensificação entre os próprios subtipos da construção. A distinção é expressa pelos pares: chuva / tempestade, mar / oceano, mundo / galáxia, monte / montão. A Figura 24 apresenta a escala na qual alguns nomes ocupariam posições extremas e outros, posições intermediárias (TAVARES, 2014, p. 73).

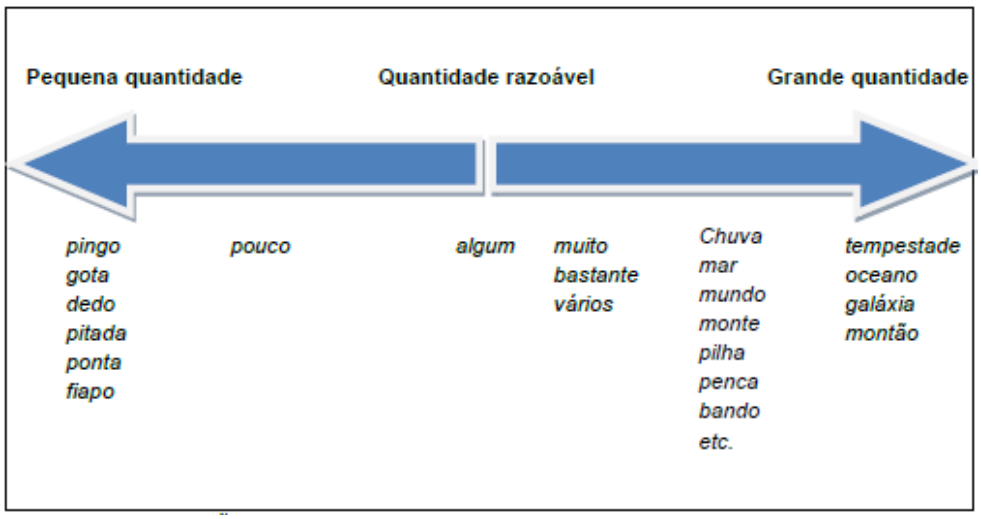


Figura 24: Distribuição dos Nomes Quantificadores na escala quantitativa

A postulação desta escala reitera o aspecto discursivo da construção, uma vez que ela que funciona como uma alternativa bastante expressiva de quantificação indefinida na língua.

No que se refere às restrições do padrão, a pesquisa identificou que os NQs que denotam grande quantidade privilegiam contextos nos quais N2 seja uma entidade Contável (78% dos dados) – vide sentenças (70-72) –, enquanto NQs que indicam pequena quantidade selecionam, na maioria das vezes, N2 Massivo (92%), conforme pode ser observado em (73-75).

(70) O Sport merece destaque: tem um **bando de craques**. (CetenFolha)

(71) Os responsáveis pelo programa vasculharam uma **multidão de arquivos de filmes**. (CetenFolha)

(72) Sabia que o choque de interesses entre os aliados eliminaria uma **penca de nomes**. (CetenFolha)

(73) Seria didático um **pingo de reflexão** sobre nossa história recente. (CetenFolha)

(74) ...Ou seja, melhor organização de jogo, uma **pitada de imprevisibilidade** nos passes e nos lançamentos… (CetenFolha)

(75) Não é com uma **ponta de dor** no coração que o poder econômico do Estado vai abandoná-lo à própria sorte. (CetenFolha)

Tais indícios apontam para as mesmas descobertas de Alonso (2010) acerca dos quatro padrões de quantificação indefinida estudados, de modo que os padrões com N2 singular estariam relacionados ao processo de extração de porção e, por isso, selecionariam entidades Massivas e os padrões com N2 plural, por sua vez, selecionariam entidades Contáveis. Há, no entanto, casos em que eventualmente o falante deseja expressar grande quantidade de entidades

Massivas ou pequena quantidade de entidades Contáveis, configurando casos de *mismatch*/desencontro linguístico, como se observa em (76-77).

(76) Tem um **pingo de aluno** na sala de aula.

(77) Mas há uma originalidade naquele **rio de belezas** da exposição.

Os desencontros semânticos ocorrem pelo fato de em (76) *pingo*, um partitivo, estar quantificando uma entidade multiplexa, isto é, Contável; em (77), *rio*, que expressa grande quantidade e que geralmente seleciona entidades individuadas, está quantificando uma entidade Massiva ou Contínua. A coerção então parece ser fundamental neste processo de harmonização dos atributos, de modo que *aluno* passa a admitir o traço [+ contínuo] e a ser interpretado como uma entidade Massiva, ajustando-se ao sentido da construção. O N2 *belezas,* por sua vez, admite o traço [+ discreto], o que justifica a forma plural, e adequa-se à combinação com o quantificador *rio* (TAVARES, 2014, p.76).

A análise também se baseou em domínios cognitivos elementares da experiência, nos esquemas imagéticos, que foram identificados como responsáveis pela conceptualização da quantificação indefinida. Observou-se que, assim como as propriedades de N1 provenientes de seu frame básico, os esquemas imagéticos seriam preservados na construção, impondo algum tipo específico de conceptualização sobre o N2. Assim, o estudo dos Nomes Quantificadores foi realizado de modo a agrupá-los nos seguintes esquemas: Verticalidade, Movimento, Contêiner, Extensão, Impacto, Coleção, Parte-Todo. O esquema de Movimento, por exemplo, que apresenta o maior número de NQs, assim também como possui os dados mais frequentes no *corpus* (cerca de 40%), inclui nomes como: *avalanche, enchente, onda, dilúvio, enxurrada, inundação, tempestade, vendaval* e *rio*. Tavares (2014, p. 81) observa que estes NQs compartilham o esquema de MOVIMENTO em seu frame básico, embora cada NQ tenha uma orientação espacial própria. Os nomes *tempestade, avalanche* e *dilúvio* possuem direção vertical DE CIMA PARA BAIXO – (78-80); *enchente* também possui direção vertical, porém com sentido DE BAIXO PARA CIMA – (81); *vendaval, enxurrada* e *rio* se movem na direção horizontal PARA FRENTE – (82-84); enquanto *onda* se move na horizontal com o sentido PARA FRENTE e PARA TRÁS – (85).

(78) A esta altura, uma **tempestade de Pelés** pode ter desabado sobre a cabeça do candidato. (CetenFolha)

(79) Essa verdadeira **avalanche de capitais** em direção aos países em desenvolvimento tem sido explicada pelos seguintes fatores conjunturais. (CetenFolha)

(80) Enquanto Goraz de era enterrada por um **dilúvio de fogo** nós éramos informados de que um acordo de cessar-fogo havia sido assinado. (CetenFolha)

(81) Um verdadeiro dilúvio, uma **enchente de cartas** inunda a redação. (CetenFolha)

(82) Entre a morte da mendiga e a tão aguardada pororoca fraterna, um **vendaval de amofinações** varre a vida do gêmeo bonzinho (...). (CetenFolha)

(83) Houve uma **enxurrada de gols**: Hungria 8 x 3 Alemanha Ocidental, Inglaterra 4 x 4 Bélgica, Áustria 7 x 5 Suíça. (CetenFolha)

(84) O risco para os usuários é gastar **rios de dinheiro** na conta telefônica. (CetenFolha)

(85) **Onda de assaltos** a carro-forte chega a PE. (CetenFolha)

Quanto às restrições do padrão, a pesquisa revelou que alguns aspectos relacionados aos esquemas imagéticos podem oferecer algumas pistas relevantes. Observem-se as sentenças (86-91):

(86) Empresas que não dão conta da **enxurrada de currículos** que recebem estão preferindo examiná-los por computador. (CetenFolha)

(87) Ele imaginava que um **dilúvio de bombas** V-1 sobre Londres levaria os aliados a pedirem a paz. (CetenFolha)

(88) Eles se sentem valorizados e recebemos **avalanche de ideias** que estão aumentando nossa produtividade. (CetenFolha)

(89) Paralisação das obras públicas federais, desemprego de mais de 600 mil trabalhadores e uma **enxurrada de ações** na Justiça. (CetenFolha)

(90) Disputa por liderança de mercado entre Antártica e Brahma provoca **onda de descontos** em pleno verão. (CetenFolha)

(91) Não podemos nos esquecer que o grande escrete de 1970 despediu-se do Brasil derrotado pelo Atlético Mineiro e sob um **vendaval de vaias**. (CetenFolha)

Em (86-87), é perceptível a combinação harmônica entre os NQs atrelados ao esquema de Movimento e as entidades *currículos* e *bombas,* por estar implícito nestas entidades a ideia de movimento. Em (88-91), no entanto, observa-se a combinação dos NQs com entidades abstratas

ligadas à informação, entidades jurídicas, econômicas e aos eventos, e, embora o N2 *ideias* possa ser metaforicamente compreendido como uma entidade projétil, o mesmo não é tão óbvio para os demais N2. Isso demonstra que o padrão possui preferências combinatórias, mas a análise dos dados parece contradizer qualquer possibilidade de regularização. Este é um desafio abordado nesta tese, uma vez que se pretende descrever tal estrutura de quantificação adequadamente e modelá-la computacionalmente, e isso significa não só modelar seus aspectos formais e semântico-funcionais, mas também suas restrições, como preconiza a Gramática das Construções de Berkeley.

Finalmente, como mencionado na seção 3.2, quanto ao modelo computacional Constructicon, durante a pesquisa de mestrado optou-se pela anotação da estrutura nesta ferramenta computacional, tendo em vista que tal estrutura foi descrita como Construção da língua e que por isso não seria tratável de modo lexicográfico. Ainda, os critérios definidos por Lage (2013) foram suficientes para a escolha do Constructicon como melhor ambiente de anotação do padrão de quantificação. Contudo, as análises desta tese apontam para o fato de que o padrão investigado não se comporta como as estruturas mais gerais da língua, pois não permite a criação de diversas expressões de modo produtivo. Na verdade, ele possibilita a geração de novas expressões, porém estas são limitadas ao processo de analogia, o que aponta para a configuração dos padrões de cunhagem. Tendo isso em vista, foi preciso realizar um novo estudo acerca dos critérios de anotação e posteriormente a remodelagem da estrutura de quantificação.

5.2 Análise das Correlações entre Frames de N1 e N2

O primeiro passo no novo estudo acerca da modelagem do padrão de quantificação indefinida buscou investigar a correlação entre os frames evocados por N1 e N2 nos dados extraídos de *corpora*. Uma vez que as análises anteriores apresentadas são unânimes em apontar para algum papel da semântica residual de N1 nas restrições selecionais aplicáveis a N2, buscamos verificar tal fato com base em dados.

As análises utilizaram os frames constantes na FrameNet Brasil como referência. Foram analisadas 549[21] sentenças com ocorrências da estrutura de quantificação e propostos dois grupos de análise: o primeiro relativo aos frames evocados por N1 e o segundo referente aos frames de N2. Para os 35 N1s foram identificados 8 frames evocados.

21 A análise das correlações entre frames foi realizada antes da expansão do corpus, que ao todo compreende 623 ocorrências da construção.

O Gráfico 1 apresenta a distribuição de todas as 549 ocorrências entre os 8 frames evocados:

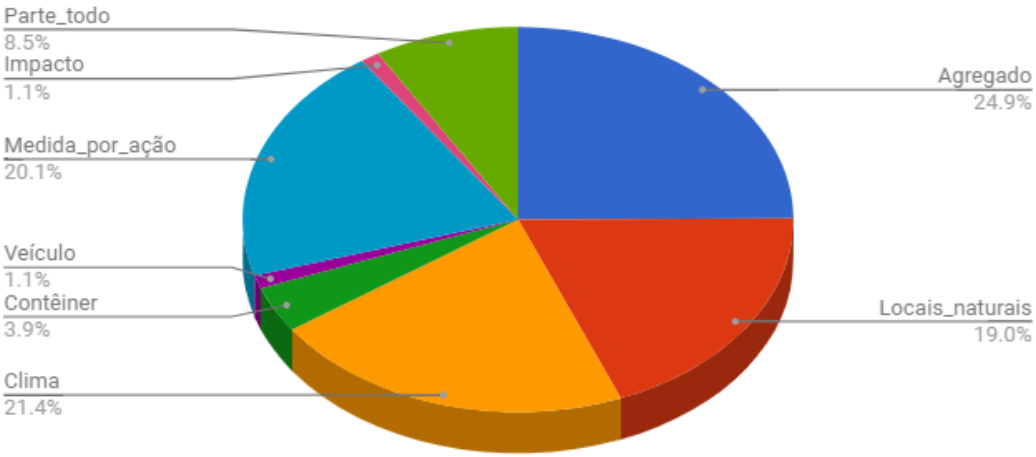


Gráfico 1: Frames elencados para N1

O frame Agregados reúne quase 25% do total de ocorrências, sendo evocado pelos Nomes Quantificadores: *corja, pelotão, batalhão, multidão, enxame, bando, penca* e *pilha*. Como o próprio nome sugere, este frame refere-se ao agrupamento de indivíduos ou coisas, o que pode justificar a proeminência deste frame nos dados, tendo em vista que a noção de grupo é uma categoria clássica de quantificação. O frame Medida_por_ação também possui grande quantidade de dados (20%) e envolve a noção de delimitação (quantificação) de uma entidade a princípio não delimitável, daí ser evocado por nomes como *bocado, pitada* e *punhado*. Por outro lado, os frames Contêiner – evocado por *poço* e *pá* -, Impacto – evocado por *porrada*- e Veículo – evocado por *caminhão* –, representam as menores porcentagens do gráfico, juntas somam 6% dos dados.

Em relação ao número de N1s para cada frame, observa-se (retomando aqui a Tabela 3) que os frames Locais_naturais, Agregados e Clima reúnem o maior número de quantificadores:

Locais_naturais	mar, montanha, oceano, galáxia, rio, floresta, mundo
Medida_por_ação	bocado, punhado, pitada
Agregados	corja, pelotão, batalhão, multidão, enxame, bando, penca, pilha

Clima	tempestade, dilúvio, onda, inundação, avalanche, enxurrada, vendaval, enchente
Contêiner	pá, poço
Parte_todo	dedo, fiapo, gota, pingo, ponta
Impacto	porrada
Veículo	caminhão

Tabela 3: Distribuição dos N1s entre os frames evocados

O principal dado para esta análise, entretanto, é a correlação entre os frames de N1 e N2. Tal informação é relevante no sentido de que permite a investigação da estrutura de quantificação como um padrão regular da língua para a expressão de quantificação (construção), desde que seja possível identificar regularidades nas combinações entre os frames de N1 e N2; ou como uma estrutura imprevisível e difícil regularização (padrão de cunhagem), caso não seja possível encontrar preferências de combinação entre os frames e nenhum padrão possível.

Tratando, portanto, da correlação entre os frames de N1 e N2, identificamos duas situações distintas e, para os propósitos desta análise, conflitantes. Considerem-se inicialmente os Gráficos 2 e 3.

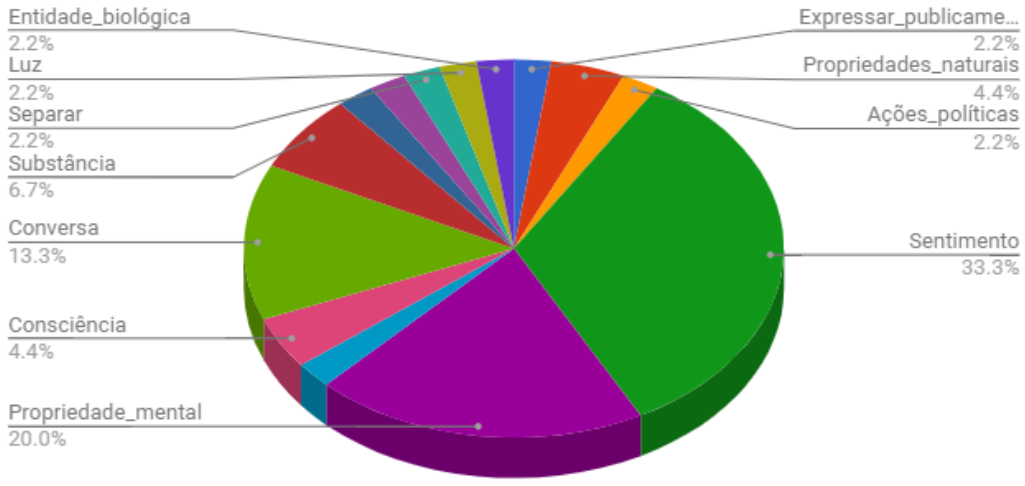


Gráfico 2: Correlação entre o frame de N1 Parte_todo e frames de N2

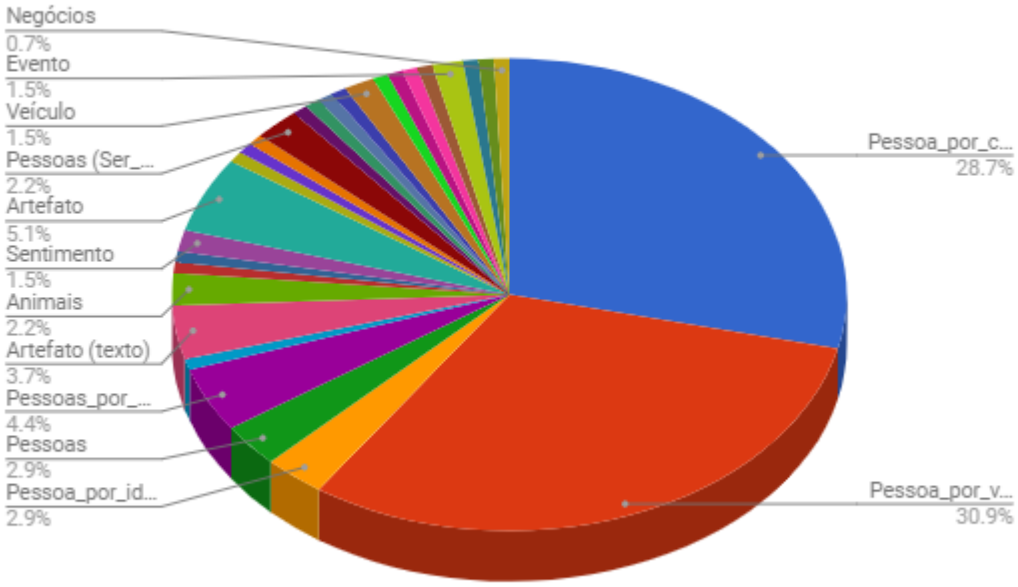


Gráfico 3: Correlação entre o frame de N1 (Agregados) e frames de N2

Os frames Parte_todo e Agregados possuem uma divisão desigual na distribuição das ocorrências com relação ao frame evocado por N2, com predominância de um ou dois frames sobre os demais. Em relação ao frame Parte_todo, no Gráfico 2, o frame Sentimento é evocado por 33% dos N2, com uma porcentagem relevante também para o frame Propriedade_mental (20%). Veja os exemplos (92-93), referentes ao frame mencionado:

(92) Apesar de se dizer satisfeito com o empate, o técnico da Suécia não escondeu uma **ponta de decepção**. (CetenFolha)

(93) Sobrou até uma **ponta de ironia** para o projeto da MTV com a gravadora Warner: “Eles estão sempre reinventando a gente”. (CetenFolha)

O Gráfico 3, referente ao frame Agregados, do mesmo modo, apresenta boa parte de seus N2s distribuídos em torno de dois frames, Pessoas_por_vocação e Pessoas_por_característica – como exemplificado, respectivamente, em (94-95), com o restante sendo distribuído entre diversos frames.

(94) Envolvendo **multidões de intérpretes**, os espetáculos de rua do Bead and Puppets nasceram há 32 anos, com o propósito de fazer arte ao alcance de todos. (CetenFolha)

(95) São **multidões de vítimas**, com quantidade pavorosa de casos fatais, sobretudo de pacientes que não quiseram ou não tiveram como recorrer a hospital público. (CetenFolha)

Os dados apresentados nos Gráficos 2 e 3, embora demonstrem alguma diversidade, possuem certas propriedades que poderíamos apontar como possíveis padrões. O fato de os N1s que evocam o frame Parte_todo se combinarem em 50% dos casos com N2 pertencentes aos frames Sentimento e Propriedade_mental já indica certa preferência da estrutura de quantificação. Do mesmo modo, N1s que evocam o frame Agregados possuem preferência por N2s que se relacionam aos frames de Pessoas, sendo Pessoas_por_vocação e Pessoas_por_característica os mais frequentes.

Embora seja possível identificar alguma regularidade nos frames anteriores, os Gráficos 4 e 5 apresentam casos em que não há predominância de determinado frame sobre os demais.

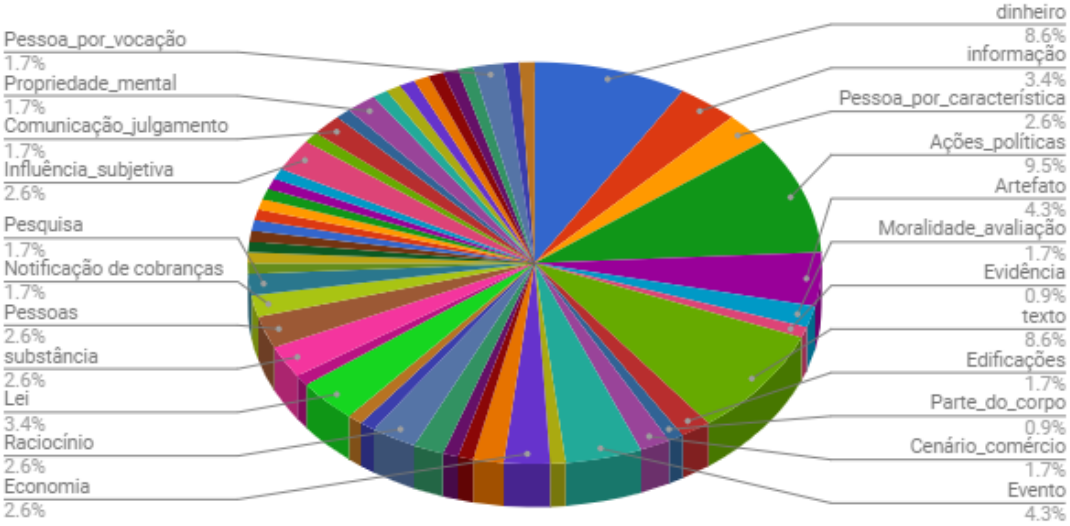


Gráfico 4: Correlação entre o frame de N1 (Clima) e frames de N2

Para o frame Clima existem cerca de 50 frames a partir dos quais se distribuem as ocorrências de N2, ou seja, a distribuição é bastante proporcional entre os frames de N2. Os frames Ações_políticas, Dinheiro e Texto – mais proeminentes nesse grupo –, são exemplificados com as sentenças (96-98).

(96) A ideia foi descartada porque o governo teme provocar uma **avalanche de reivindicações** semelhantes de outras estatais. (CetenFolha)

(97) Ele acredita até que o governo precisará realizar controles para segurar essa **enxurrada de dinheiro externo** a partir do próximo ano. (CetenFolha)

(98) Se tudo correr como prevê a Riofilme, distribuidora subordinada à prefeitura carioca, no segundo semestre deste ano uma **avalanche de filmes brasileiros** chegará aos cinemas. (CetenFolha)

Quanto ao frame Locais_naturais, há também uma distribuição bastante equilibrada, como se observa no Gráfico 5.

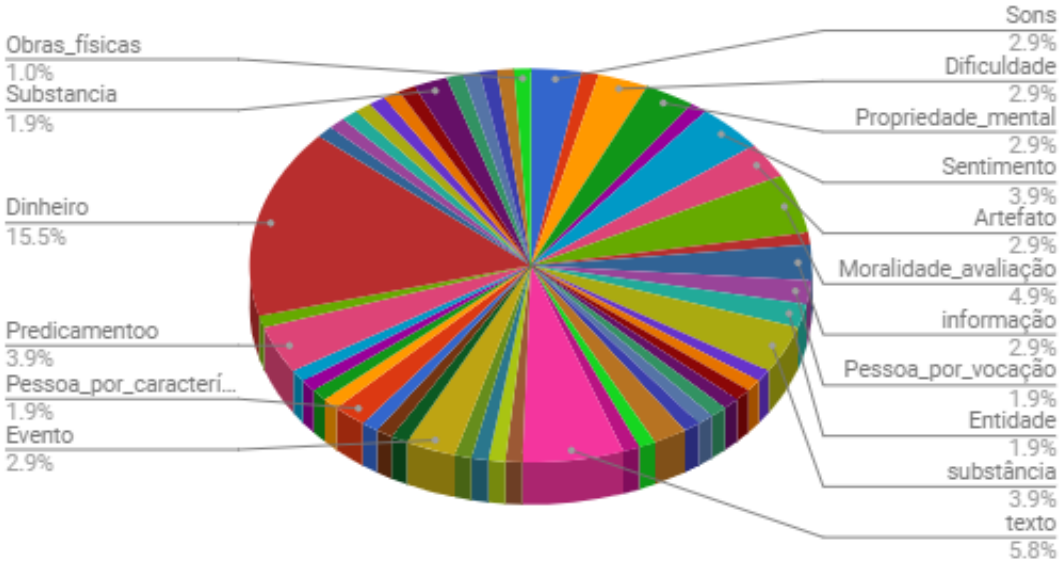


Gráfico 5: Correlação entre o frame de N1 (Locais_naturais) e frames de N2

Há pouco mais de 40 frames identificados para N2, de modo que alguns frames possuem uma frequência relativamente maior, como é o caso de Moralidade_avaliação e Sentimento, exemplificados com as sentenças (99-100).

(99) Mais difícil tão difícil que nem se menciona mais é a idéia de que foi derrubado pelas ondas do **mar de corrupção**, que se demonstrou inexistir. (CetenFolha)

(100) Foram **rios de indignação** contra o preconceito implícito na analogia. (CetenFolha)

Como pôde ser visto nos gráficos anteriores, é grande a diversidade tipológica de frames do N2 e a distribuição desses nos dados analisados é bastante equivalente. Nestes casos específicos, observa-se o comportamento mais produtivo da construção e algum afastamento dos padrões de cunhagem.

Entretanto, o estudo de Alonso (2010), apesentado na seção 5.1.2 deste capítulo, traz à tona uma possibilidade de regularização dos padrões de quantificação através dos processos cognitivos de extração de unidade ou porção e de multiplexização. Retomando brevemente este estudo, a autora evidencia que os padrões estariam vinculados a tais processos, os quais seriam responsáveis por determinar as restrições de combinação entre os elementos constituintes da construção. Assim, um padrão de quantificação relacionado ao processo de multiplexização selecionaria prototipicamente N2 massivo, uma vez que a função deste processo é a de quantificar entidades que a princípio não seriam quantificáveis; como consequência, tem-se o N2 prototipicamente singular. Por outro lado, quando se trata de multiplexização, o processo consiste em multiplicar entidades concebidas como discretas, portanto, contáveis, o que determina que N2 seja preferencialmente plural (nomes contáveis podem sofrer ou não a flexão de número, porém para o sentido multiplexo da construção o número plural seria preferido).

Tais observações também podem ser evidenciadas na análise de nosso conjunto de dados, uma vez que o padrão de pequena quantidade indefinida seleciona, em sua grande maioria – em 73 de 77 ocorrências –, N2 massivo/singular. De modo oposto, o padrão de grande quantidade indefinida seleciona dentre as 545 ocorrências, 453 N2 plural/contável.

De fato, os nomes quantificados no padrão de pequena quantidade não seriam *a priori* quantificáveis, por exprimirem noções abstratas, dentre eles figuram: *inteligência, humor, ironia, ódio, modéstia, respeito*, etc. A interpretação imediata deste padrão nos remete a uma parte bastante pequena do todo (*pitada de inteligência*), às vezes nos remete à parte mínima (*ponta de ironia)*, de qualquer modo o padrão envolveria o processo cognitivo de extração de porção, mesmo que de modo não tão explícito, como no caso de *um quilo de farinha* e *uma xícara de açúcar.* Os únicos casos em que o padrão de pequena quantidade selecionou N2 plural foram aqueles em que o nome sofreu flexão de número pelo fato de ser muito utilizado no plural, são eles: *ciúmes* e *saudades*. Ademais, houve casos de seleção de alguns N2 contáveis, embora estes tenham se apresentado no singular, como: *teoria, dúvida, crítica, testemunho, conversa* e *palestra*.

Em relação ao padrão de grande quantidade, dentre os casos de seleção de N2 singular, encontram-se nomes como: *gente*, que não vai para o plural por ser uma entidade multiplexa por natureza; *tempo* e *dinheiro*, categorias essencialmente concebidas como abstratas e massivas e que não sofreriam flexão de número possivelmente devido à existência de suas contrapartes concretas e contáveis, como *horas*, *minutos, segundos* e *notas/cédulas* e *moedas*, respectivamente; há ainda uma quantidade insignificante de nomes que podem ser flexionados, porém mantiveram-se sem flexão. Sendo assim, como propõe Alonso (2010), esta seria uma

estrutura que seleciona preferencialmente nomes no plural, justificando o processo de multiplexização subjacente. É importante destacar que 46% das ocorrências que apresentam N2 singular pertencem à construção com *bocado,* que, como evidenciou Tavares (2014), trata-se de um quantificador ambíguo, que ora atua como quantificador de grande quantidade ora como quantificador de pequena quantidade. Embora tenha sido considerado o contexto no qual a expressão é instanciada e a partir deste contexto tenha se interpretado *bocado de* como grande quantidade, percebe-se em alguns casos que a interpretação de pequena quantidade também seria possível.

Já pela abordagem bottom-up, a análise das restrições que governam a construção pode levar a diversos padrões, estes, por sua vez, podem estar relacionados às restrições de cada microconstrução. Assim, seria preciso considerar que *uma pilha de,* embora funcione como um quantificador e se combine a nomes como *pessoas, emoção* e *motivos*, ainda seleciona uma grande quantidade de entidades passíveis de serem empilhadas. Isso se justificaria pela Persistência da Imagem Conceptual do item lexical sobre a construção, citada no início deste capítulo, que, segundo Verveckken (2012), consiste na permanência de propriedades conceptuais do nome, as quais são responsáveis pelas restrições que governam as combinações entre N1 e N2. Nesse sentido, a entidade designada por N2 tem de se ajustar à imagem ativada por N1, o qual já apresenta uma configuração própria (TAVARES, 2014). A análise de *corpus* demonstra que *pilha* quantifica, preferencialmente, entidades que possam ser empilhadas (nos casos de N2 concreto). Entre estas entidades estão nomes como: *livros, dinheiro, lixo,* etc., representando 68% dos dados. De modo semelhante a *pilha*, *montanha* privilegia a quantificação de entidades que, se acumuladas umas sobre as outras, podem assumir a configuração de uma pequena montanha: *documentos, dinheiro, brinquedos, livros, cartas*, etc. Contudo, apesar de tais propriedades restritivas, o N1 admite em alguma medida a expansão da sua categoria e passa a quantificar também outros elementos não esperados (pessoas, entidades abstratas).

Esta expansão torna desafiante a tarefa de identificação de quais elementos seriam possíveis candidatos a ocorrerem no *slot* N2, e a dificuldade é ainda maior por tratar-se de uma construção com uma interpretação tão subjetiva. Além disso, embora não se acredite que, em *uma pilha de N*, N possa ser qualquer nome, existem situações em que o contexto é capaz de habilitar a leitura de um nome muito pouco provável de ocorrer na construção. Observe o exemplo (101) retirado do Corpus do Português.

(101) Oie gente ^^ (...) perdão por demorar sempre pra postar, mas eu to **um poço de nervos** gente '-- (sim, eu sei que a expressão é "**pilha de nervos**", mas eu gosto de falar poço u.ú) (Corpus do Português – Web/Dialetos)

Quando se trata de quantificação, *nervos* combina-se exclusivamente com *pilha*, tratando-se de uma expressão cristalizada do português, e como veremos mais adiante, tal expressão faz parte de um pequeno conjunto de expressões que possuem produtividade alta de token, mas não de type, uma vez que nem o *slot* de N1 e nem o de N2 costumam sofrer variações. No entanto, no exemplo em (101), houve uma variação, a qual foi explicada apenas em termos de preferência do falante pela expressão com *poço*. Por ser uma expressão bastante convencionalizada, a aceitabilidade de *poço de nervos* seria bem reduzida. Contudo, este exemplo demonstra que com o contexto necessário (neste caso a explicação entre parenteses) até expressões mais convencionalizadas permitem ser modificadas de modo criativo pelo falante. A Figura 25, extraída do Corpus do Português, demonstra a alta frequência de *pilha de nervos* em relação aos demais padrões:

		CONTEXT	FREQ
1	☐	ATAQUE DE NERVOS	170
2	☐	PILHA DE NERVOS	43
3	☐	CRISE DE NERVOS	28
4	☐	PARES DE NERVOS	24
5	☐	GUERRA DE NERVOS	19
6	☐	GÁS DE NERVOS	19
7	☐	COMPRESSÃO DE NERVOS	18
8	☐	CONJUNTO DE NERVOS	15
9	☐	ATAQUES DE NERVOS	10
10	☐	REDE DE NERVOS	9

Figura 25: Frequência da expressão *pilha de nervos* no Corpus do Português

Por isso, reconhece-se que uma análise minuciosa das restrições de cada subtipo da construção de quantificação indefinida seja necessária para o propósito de modelagem computacional do padrão de quantificação e sua aplicação num modelo de tradução, mesmo que alguns casos específicos, amparados pelo contexto, venham contradizer algumas das análises. Como será explicitado no próximo capítulo, a modelagem da construção no Constructicon possibilitará também a modelagem de suas restrições, as quais constituem evidências adicionais ou restrições *soft* acerca do padrão.

5.3 Resultados do Teste de Julgamento de Aceitabilidade

Retomando brevemente os procedimentos e objetivo deste experimento, os voluntários responderam a um questionário contendo ao todo 30 questões. Para cada questão havia 3 sentenças: 1 sentença com o objeto de estudo (SC), 1 sentença com sentido equivalente (SE) – sem a presença da construção de quantificação investigada – e 1 sentença distratora (SD). Tais sentenças deveriam ser classificadas conforme a escala: 1. inadequada, 2. pouco adequada, 3. razoavelmente adequada, 4. bastante adequada e 5. perfeitamente adequada.

O principal objetivo deste teste foi avaliar a possibilidade de variação do *slot* de N2 para determinadas combinações com N1, a partir do julgamento de aceitabilidade proposta aos participantes, tendo em vista que o padrão de quantificação parece se expandir na língua. Para isso foram apresentadas aos voluntários expressões de quantificação provenientes do *corpus* e expressões novas, formando-se então três grupos: Grupo 1 – sentenças contendo expressões de quantificação encontradas nos corpora (N2 bastante provável); Grupo 2 – expressões de quantificação novas, criadas por analogia com aquelas que ocorrem no *corpus* (N2 razoavelmente provável); e Grupo 3 – expressões de quantificação novas, criadas sem levar em conta a analogia ou a preservação domínio semântico de N2 (N2 pouco provável).

A Figura 26 apresenta o resultado da classificação das três sentenças contidas na questão 01 feita por 26 voluntários.

	1. INADEQUADA	2. POUCO ADEQUADA	3. RAZOAVELMENTE ADEQUADA	4. BASTANTE ADEQUADA	5. PERFEITAMENTE ADEQUADA	TOTAL	WEIGHTED AVERAGE
Moro no bairro há vinte anos e nunca vi uma onda de assaltos tão assustadora, diz Citelli.	0.00% 0	0.00% 0	7.69% 2	0.00% 0	92.31% 24	26	4.88
Moro no bairro há vinte anos e nunca tinha visto tantos assaltos assustadores, diz Citelli.	0.00% 0	3.85% 1	7.69% 2	15.38% 4	73.08% 19	26	4.58
Moro no bairro há vinte anos e nunca vi assaltos tanto apenas assustador, diz Citelli.	84.62% 22	11.54% 3	0.00% 0	3.85% 1	0.00% 0	26	1.23

Figura 26: Resultado de uma das questões do teste de julgamento de aceitabilidade

Evidencia-se a alta taxa de aceitação da expressão contendo a construção (92%), embora a sentença equivalente também tenha sido classificada de modo satisfatório (73%). E, como era de se esperar, a sentença distratora foi classificada como “inadequada” em 84% das vezes, uma vez que continha desvios gramaticais.

A Tabela 6 compara de modo mais explícito os resultados obtidos com sentenças contendo uma expressão de quantificação (SC) e sentenças com sentido equivalente (SE). Os números correspondem às porcentagens de classificação das sentenças como “perfeitamente adequadas”, tendo em vista as possibilidades de classificação citadas anteriormente.

Grupo 1: N2 bastante provável	SC	SE	Grupo 2: N2 razoavelmente provável	SC	SE	Grupo 3: N2 pouco provável	SC	SE
caminhão de dinheiro	65	88	caminhão de memes	53	84	caminhão de vergonha	42	88
onda de assaltos	92	73	onda de sorrisos	61	80	ondas de dinheiro	42	80
dedo de prosa	76	73	dedo de reggae	65	61	dedo de gente	53	73
punhado de gente	84	76	punhado de tempo	57	73	punhado de deboche	65	92
mar de corrupção	61	69	mar de lixo	61	65	mar de filhos	57	88
pitadas de bom humor	84	76	pitada de medo	30	40	pitada de maquiagem	69	84
porrada de coisas	38	84	porrada de descrédito	38	88	porrada de infelicidade	34	80
penca de gente	61	80	penca de lançamentos	34	80	penca de impunidade	46	84
rios de dinheiro	76	73	rio de situações	50	88	rio de turistas	53	84
pilha de material de propaganda	88	80	pilha de problemas	53	88	pilha de tempo	34	88

Tabela 6: Porcentagens relativas à aceitabilidade de sentenças contendo a construção (SC) e sentenças equivalentes (SE)

No que se refere ao grupo 1, caso em que N2 é bastante provável de ocorrer no padrão, tem-se que os participantes demonstraram melhor aceitação das expressões de quantificação, ou seja, tais expressões apresentaram boas porcentagens para a classificação "perfeitamente adequadas". Tal resultado já era esperado, uma vez que o falante pode ter tido contato anteriormente (textos orais ou escritos) com as expressões apresentadas – algumas destas expressões inclusive possuem frequência de ocorrência relativamente alta na língua[22]. Comparando-se as sentenças em que ocorre a construção (SC) e aquelas equivalentes (SE), em 6 dos 10 casos, as sentenças SC foram mais bem ranqueadas que as sentenças SE, o que está

22 De acordo com análise dos dados extraídos do Corpus CetenFolha.

destacado em cinza na Tabela 6. Nos Grupos 2 e 3, entretanto, as expressões de quantificação não tiveram o mesmo desempenho. Nestes casos, prevaleceu a sentença equivalente como a mais bem aceita pelos voluntários. O Gráfico 6 resume tais observações a partir da análise dos resultados de aceitação de *onda de N* e de suas sentenças semanticamente equivalentes:

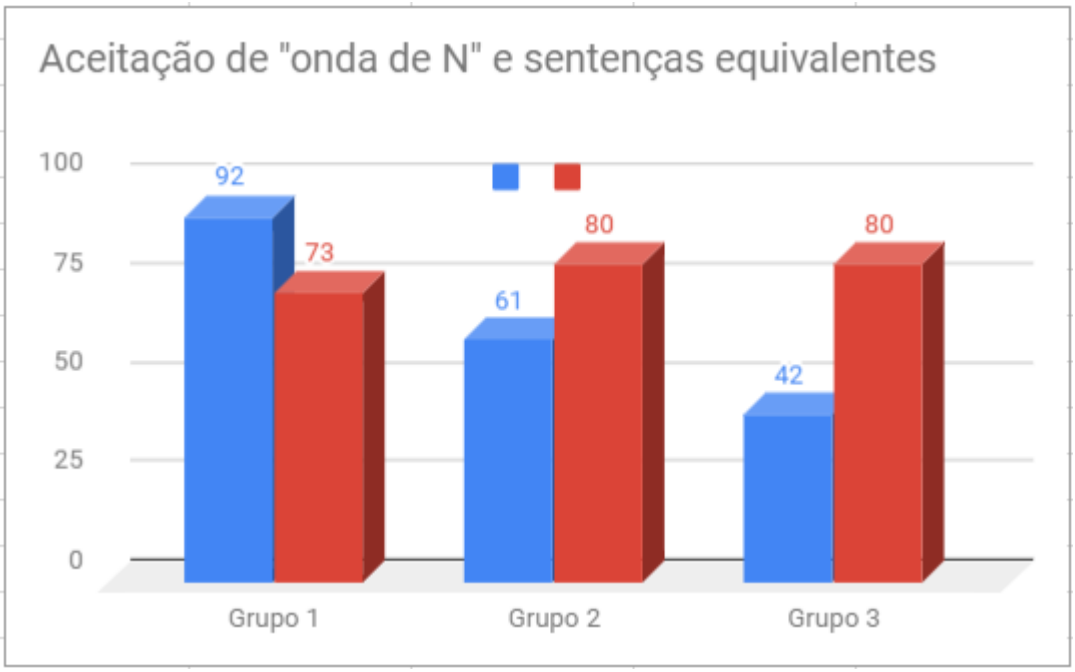


Gráfico 6: Comparação entre a aceitação de *onda de N* e de sentenças equivalentes

O Grupo 1, o qual contém uma expressão convencional do português, é classificada como perfeitamente adequada em 92% das vezes, enquanto sua sentença equivalente recebe a mesma classificação 73% das vezes. Já nos Grupos 2 e 3, a aceitação das expressões de quantificação indefinida com *onda de N* recebem a classificação de perfeitamente adequadas em 61% e 42% das vezes, respectivamente, de modo que as sentenças equivalentes são mais bem-aceitas, com uma taxa de 80% de aceitação.

É possível pensar em algumas hipóteses para este resultado. As expressões dos grupos 2 e 3 foram criadas e não atestadas em *corpus,* o que sugere que o falante pode nunca ter tido contato com esta nova forma, gerando um estranhamento e a consequente recusa da expressão. Por serem metafóricas, as expressões investigadas podem ser tomadas como recursos estilísticos por quem as lê (em alguns casos a linguagem figurada fica bastante evidente, basta comparar *pilha de material de propaganda* e *pilha de problemas*), tal observação pode levar o participante a interpretar a sentença equivalente – mais próxima do sentido literal – como a opção mais adequada. Também é possível que o participante procure fazer uma análise das sentenças e ao se deparar com uma sentença contendo a expressão de quantificação, que

funciona como um recurso discursivo enfático, e uma sentença equivalente, porém sem a presença de qualquer recurso linguístico que lhe chame a atenção, é provável que ele dê uma classificação melhor para a segunda, considerando-a mais adequada.

É perceptível a limitação da estrutura de expandir seu uso para a combinação com N2 de domínios diferentes, como *caminhão,* um quantificador de difícil regularização – pela quantidade insuficiente de dados encontrados não foram identificadas as preferências de combinação deste nome com N2 –, porém sabe-se que o quantificador parece se combinar adequadamente com muitos nomes concretos e abstratos; nomes que evocam o frame de Dinheiro e Pessoas são frequentes entre os dados, porém nomes que evocam Sentimento parecem ser menos prováveis de ocorrerem nessa combinação, como *caminhão de vergonha* (42%)*,* menos aceito que *caminhão de memes* (53%) e *caminhão de dinheiro* (65%).

Os resultados obtidos demonstram que as variações do *slot* N2 não são tão bem-aceitas, especialmente quando o N2 pertence a um domínio semântico distinto daqueles atestados em *corpus*, o que contribui para o tratamento de tais expressões como padrões de cunhagem.

5.4 Extensão metafórica entre as expressões de quantificação indefinida

A quantidade indefinida expressa pelos nomes investigados neste trabalho é metafórica, isto é, o nome quantificador (N1) é interpretado de modo figurado como uma unidade de medida. Nomes como *gota, pingo, pitada* já funcionam *a priori* como unidades de medida e ocorrem geralmente na posição de Modificador num padrão N1 de N2 (*gota* d'água, *pingo de chuva, pitada de sal)*. Porém, há nomes como *caminhão, pilha* e *bando*, que embora também possam comumente funcionar como Modificadores neste padrão (*caminhão de leite, pilha de livros, bando de prisioneiros*), não são diretamente interpretados como unidades de medida.

Não se pode negar, entretanto, que, para além da interpretação destas expressões como Modificação, há também a noção de Medida ou Quantidade. A leitura dessas expressões como unidades de medida pode não ser tão clara devido ao fato de que a medida expressa por elas, como por exemplo, em *pilha de livros,* envolve uma interpretação subjetiva. Isto é, uma pilha geralmente compreende a sobreposição de objetos, porém não há uma determinação exata de quantos objetos seriam necessários para se constituir uma pilha. Bastariam dois livros sobrepostos para que tenhamos uma pilha? A subjetividade desta expressão também é decorrente de aspectos culturais e de experiências individuais, assim uma determinada quantidade de livros empilhados sobre a mesa pode ser pequena para um leitor frequente, mas enorme para quem não tem o hábito da leitura.

Nesse contexto, por mais variável que possa se apresentar a configuração real de uma pilha, o que está em jogo nesta investigação é sua imagem prototípica, ou seja, sua concepção como uma sobreposição considerável de objetos, como mostra a Figura 27, um dos resultados mais frequentes com a busca pela expressão “pilha de livros” no Google Imagens.

Figura 27: Resultado da busca no Google Imagens por “pilha de livros”

A leitura de uma pilha como uma grande quantidade de objetos empilhados também se justifica a partir da Teoria da Metáfora Conceptual (LAKOFF & JOHNSON, 1980), mais especificamente, pela metáfora primária MAIS É PARA CIMA, devido ao seu esquema imagético vertical. E é a partir dessa ancoragem metafórica que pilha estende seu uso como um quantificador da língua, admitindo a quantificação de seres animados, eventos e entidades abstratas, como se pode observar nas sentenças (102-104).

(102) Para os milanistas, havia um meio mais romântico de fazer a viagem: de barco, até a ilha grega de Corfu, um desses paraísos que atraem **pilhas de turistas** agora que o sol saiu na primavera europeia. (CetenFolha)

(103) É claro, no meio do filme haverá congestionamentos e semáforos em número suficiente para garantir **pilhas de trombadas**, carros destruídos e emoção. (CetenFolha)

(104) É uma insensatez por **uma pilha de motivos** que não cabem, todos, neste espaço. (CetenFolha)

Também é possível visualizar essa questão através do nome *caminhão.* No contexto de uma obra civil, é comum que se façam encomendas de um ou mais caminhões de areia, e o que se está comprando de fato não é o caminhão e sim uma grande quantidade de areia – quantidade bastante variável a depender do tamanho do caminhão. Fica clara neste caso a perspectiva sobre a quantidade de areia transportada, ficando em segundo plano o tipo de caminhão (de areia).

A hipótese, portanto, é a de que a interpretação literal de *pilha de livros* como um indicador de grande quantidade possa ter dado origem a expressões metafóricas como *pilha de desempregados*, assim também como haveria a extensão metafórica entre *caminhão de areia* e *caminhão de filhos*.

Como dito anteriormente, a escolha de nomes como *pilha* e *caminhão*, por exemplo, para a construção aqui investigada justifica-se pela experiência do falante com o meio em que vive e sua interação com os esquemas imagéticos. Conforme postulado por Lakoff & Johnson (1999, p. 49), as metáforas primárias "*emparelham a experiência e o julgamento subjetivos com a experiência sensório-motora da vida cotidiana*". Daí a relação entre pilha e a concepção de que Verticalidade é Quantidade. Além disso, um indivíduo vê muito mais frequentemente caminhões no dia a dia do que outros veículos de carga. Um foguete espacial dificilmente seria recrutado para esta construção pelo fato de não fazer parte do nosso cotidiano e também por não ser um veículo de carga prototípico. Assim, também seria pouco provável que a construção permitisse nomes como *carro, ônibus, moto*, pois, embora esses sejam inúmeros no nosso cotidiano, também não são veículos de carga prototípicos.

5.4.1 A continuidade entre as construções quantitativas literal e figurada

Grady (2007) discute algumas questões que envolvem a diferenciação entre linguagem literal e figurada e que são relevantes para esta análise. Segundo o autor, há na língua casos em que não se pode estabelecer com clareza uma linha divisória entre o que é literal e o que é figurado, e o que é comumente concebido como metáfora pode ser na verdade um caso em que a categoria é "esticada" para acomodar novos itens. Um conceito intricado como o da palavra inglesa *question* em *Life is a question*, pode ser entendido: i) de maneira figurada, ou seja, seu sentido literal deu lugar a uma interpretação metafórica, na qual um elemento concreto, uma frase interrogativa, passa a se referir a um conceito mais abstrato e complexo (de que a vida é incerta); ii) ou de maneira literal, assumindo-se que a palavra *question* tenha expandido seu conceito para abarcar também a concepção abstrata de incerteza, sendo portanto associada, pelo menos, a dois sentidos diferentes.

A hipótese sobre o estiramento da categoria é interessante neste trabalho pois suporta as análises acerca da expansão de possibilidades de combinações entre N1 e N2, ou seja, retomando o quantificador *pilha*, este amplia seu escopo de combinações possíveis, por isso tem sua categoria esticada para acomodar outras entidades. Isso não exclui, entretanto, sua interpretação metafórica. Na verdade, é a leitura metafórica que permite sua abertura para novos elementos.

Outra razão pela qual a abordagem figurada é preferível à literal é a de que caso a categoria de pilha tivesse sido apenas ampliada para abarcar também coisas não empilháveis, sendo então interpretado literalmente como "pilha" e como "grande quantidade", o mesmo deveria ser válido para os demais 35 nomes investigados nesta tese. Ocorre que, de fato, os demais nomes que participam do padrão de quantificação também veiculam a noção de grande quantidade mesmo em seu frame básico, ou seja, *avalanche, enxurrada, enchente, tempestade*, por exemplo, sempre envolvem uma quantidade considerável de água. Contudo, sabe-se que tal análise perderia em generalidade, sendo preferível assumir uma rede conceptual metaforicamente motivada, na qual todos os nomes quantificadores estariam interligados uns aos outros.

Discussão semelhante é feita por Lakoff (1992, p. 81), em relação ao termo *cold,* em *John is cold,* propondo três possibilidades de interpretação para o mesmo: i) a leitura literal, em que *cold* seria uma palavra com dois possíveis sentidos (frio e insensível), que tem como obstáculo o fato de *warm* e *cool* serem outros sentidos aprendidos separadamente; ii) a leitura metafórica, que parece ser a mais adequada, tendo em vista a possibilidade de relacionar os termos *cold, warm* e *cool* num mesmo sistema conceptual que habilita, via metáfora, a interpretação de tais termos como propriedades psicológicas; e iii) por último, o autor sugere a possibilidade de se considerar inicialmente *cold* como expressão metafórica que ao longo do tempo se convencionalizou e tornou-se uma expressão idiomática.

No que se refere às expressões de quantificação, tomando agora para exemplificação o domínio dos fenômenos da natureza, é possível atribuir a *chuva de reclamações* uma interpretação literal? Acredita-se que não, pelo mesmo motivo de *cold,* uma vez que os demais fenômenos da natureza que funcionam como quantificadores (*tempestade, enxurrada, enchente, avalanche*, etc.) deveriam ser aprendidos separadamente. A interpretação metafórica, portanto, parece a mais satisfatória. Quanto à última proposta de Lakoff, não é possível afirmar que tais palavras tenham se convencionalizado a ponto de se tornarem expressões idiomáticas e se distanciarem da interpretação metafórica como ocorre com *cold,* tendo em vista que este se refere a uma experiência humana muito mais elementar.

As autoras Brodbeck (2010) e Verveckken (2012) evidenciaram o papel pioneiro de *monte* no surgimento do padrão de quantificação [N1 de [N2]]. Segundo Verveckken, *monte* teria se tornado um quantificador completamente transparente diante do seu grau de convencionalização. Parece razoável assumir uma distinção de grau na leitura figurada de *monte* em relação aos demais quantificadores, ou seja, a expressão *enxurrada de gente* seria claramente uma metáfora, enquanto *monte de gente* teria uma interpretação mais sutil. Também é plausível mencionar tal diferença entre as expressões *os preços subiram* e *os preços estão salgados,* pois a

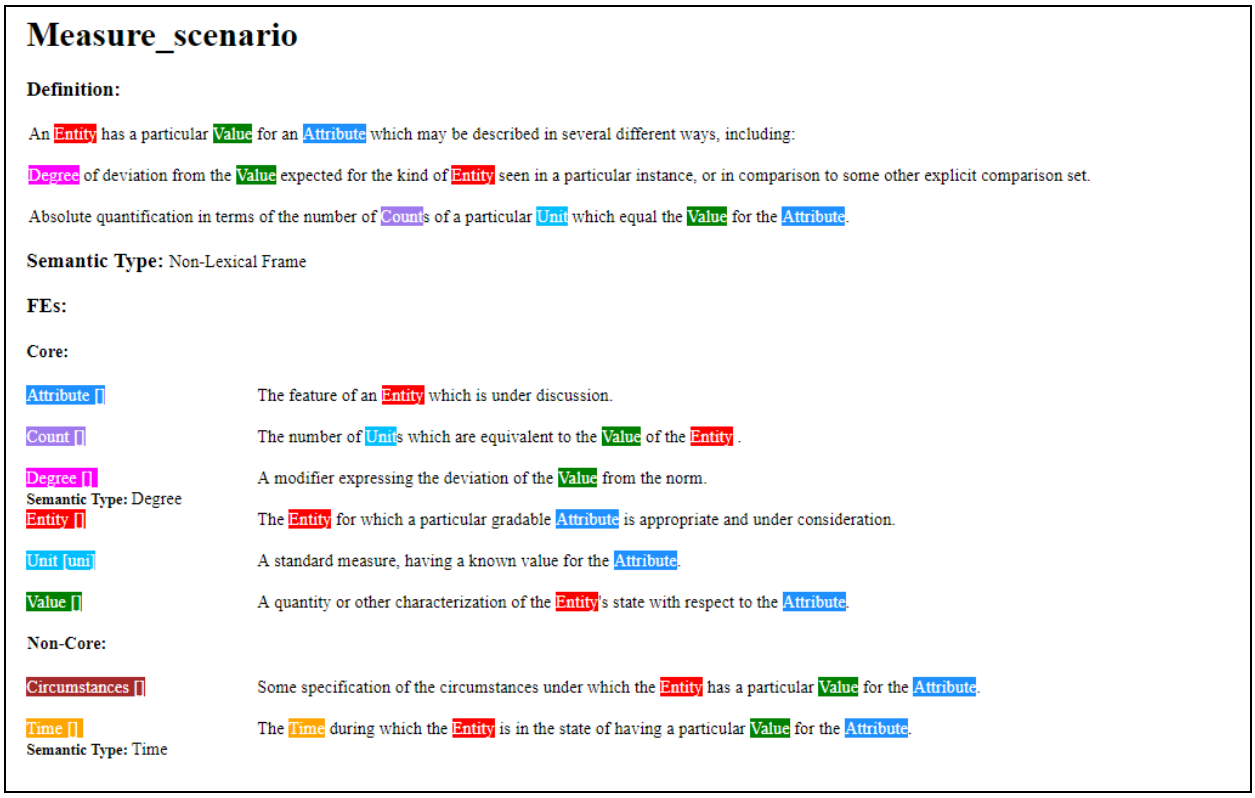
primeira faz parte de um sistema de conceitos envolvendo uma metáfora primária com origem na experiência imagética (verticalidade é quantidade), o que a torna muito mais comum no uso cotidiano e sua interpretação metafórica mais sutil em relação à segunda, que possui uma referência metafórica mais explícita.

Por estarmos falando em distinção de grau, é bastante pertinente a esta análise a proposição de uma escala, que promova o *continuum* entre linguagem literal e figurada, o qual contribuirá para as observações subsequentes acerca da construção de quantificação indefinida. Nesse sentido, admite-se que, embora o escopo deste estudo seja a construção de quantificação indefinida metafórica, é relevante relacioná-la à construção literal. A Tabela 7 apresenta algumas das relações responsáveis pela extensão metafórica entre as expressões.

pilha de livros	pilha de gente
punhado de moedas	punhado de pessoas
penca de bananas	penca de gente
caminhão de leite	caminhão de filhos
bando de pássaros	bando de sentimentos
multidão de pessoas	multidão de telespectadores
gota de água	gota de humor
pitada de sal	pitada de loucura
pingo de suor	pingo de medo
fiapo de cabelo	fiapo de voz
bocado de bolo	bocado de tempo

Tabela 7: Relação entre as expressões de quantificação literal e abstrata

Assim, assume-se a existência de duas construções abstratas de quantidade indefinida, uma literal (*pilha de livros, caminhão de areia*, etc.) e outra metafórica (*pilha de desempregados, caminhão de filhos,* etc.), ambas relacionadas por uma construção também abstrata de Medida Subjetiva. Esta, por sua vez, relaciona-se ao frame não lexical Cenário_de_medida, descrito pela FrameNet de Berkeley da forma apresentada na Figura 28.



Measure_scenario

Definition:

An Entity has a particular Value for an Attribute which may be described in several different ways, including:

Degree of deviation from the Value expected for the kind of Entity seen in a particular instance, or in comparison to some other explicit comparison set.

Absolute quantification in terms of the number of Counts of a particular Unit which equal the Value for the Attribute.

Semantic Type: Non-Lexical Frame

FEs:

Core:

Attribute []	The feature of an Entity which is under discussion.
Count []	The number of Units which are equivalent to the Value of the Entity.
Degree [] **Semantic Type:** Degree	A modifier expressing the deviation of the Value from the norm.
Entity []	The Entity for which a particular gradable Attribute is appropriate and under consideration.
Unit [uni]	A standard measure, having a known value for the Attribute.
Value []	A quantity or other characterization of the Entity's state with respect to the Attribute.

Non-Core:

Circumstances []	Some specification of the circumstances under which the Entity has a particular Value for the Attribute.
Time [] **Semantic Type:** Time	The Time during which the Entity is in the state of having a particular Value for the Attribute.

Figura 28: Descrição do frame Cenário_de_medida pela FrameNet de Berkeley

Tendo como background o frame de Cenário_de_medida, as construções abstratas compartilham o padrão N1 de N2 e diferem-se na interpretação literal ou figurada da quantidade que expressam. Assim, no nível da macroconstrução temos a construção abstrata de Medida Subjetiva e a nível das mesoconstruções, as construções abstratas de Quantidade Indefinida Literal e Metafórica. A Figura 29 ilustra a relação entre tais níveis, bem como a extensão metafórica que ocorre no nível das mesoconstruções.

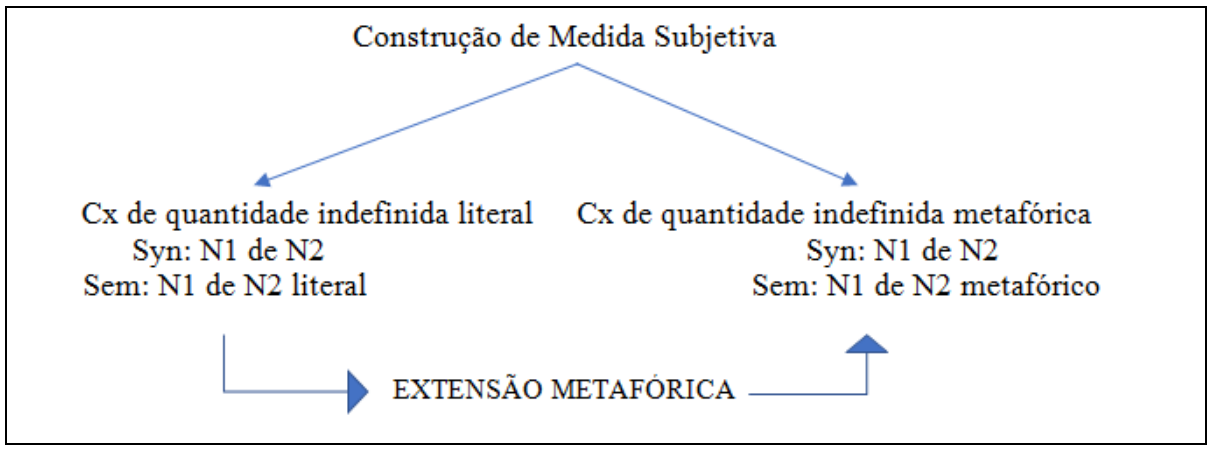


Figura 29: Extensão metafórica da construção abstrata de quantidade literal

Segundo essa análise, teríamos, portanto, dois possíveis usos de *pilha* ou dois constructos: um como unidade de medida subjetiva literal (Construção de Quantificação

Subjetiva Literal) e outro como unidade de medida subjetiva metafórica (Construção de Quantificação Subjetiva Metafórica), representados na Figura 30.

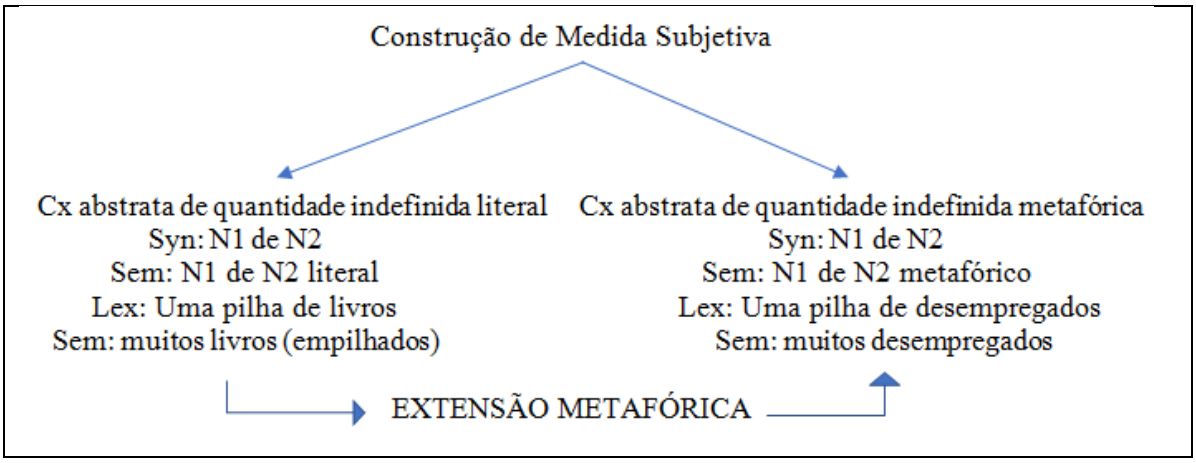


Figura 30: Extensão metafórica do constructo *pilha de livros*

Considerem-se agora as sentenças (105-106), as quais apresentam tais constructos.

(105) (...) A ele, disse, apontando para Starzl, que se afastava em direção **à pilha de livros** que esperavam o seu autógrafo. (CetenFolha)

(106) Consequência inevitável: **uma pilha de desempregados**, que, só nos 24 países mais ricos do mundo, afeta 35 milhões de pessoas. (CetenFolha)

Em (105) percebe-se que a expressão destacada refere-se a uma pilha literal de livros à espera dos autógrafos pelo autor[23], enquanto que em (106) o nome *pilha* refere-se somente à grande quantidade de desempregados, não havendo nenhuma restrição a respeito da disposição física da entidade quantificada.

No que se refere ao nome quantificado (N2), este pode apresentar propriedades semânticas bem diferentes nas construções literal e metafórica. Para a primeira, é necessário que o nome quantificado atenda aos critérios: i) ser Contável – com algumas exceções (lixo, madeira, detritos); ii) ser Concreto; iii) e ser um objeto do tipo *empilhável*. Alguns nomes que atendem a estes critérios seriam: livros, roupas, ossos, cartas, caixas, travesseiros, papéis, discos, etc. A construção metafórica, por sua vez, rompe as barreiras de tais restrições: i) embora nela também haja predominância de nomes Contáveis, há um aumento significativo de nomes Massivos ocorrendo na construção; ii) não há restrições em relação ao aspecto Concreto ou Abstrato; iii)

23 Cabe a ressalva de que neste caso há sobreposição entre linguagem literal e figurada, devido à personificação dos livros.

também não há a exigência de que os nomes concretos sejam passíveis de serem empilhados. Exemplos de nomes que ocorrem nesta construção são: turistas (concreto/contável), dinheiro (concreto/massivo), motivos (abstrato/contável), humor (abstrato/massivo).

Em relação às expressões literais com *caminhão de,* estas possuem menos restrições. A entidade quantificada pode ser Contável (melancias, bois, eletrodomésticos) ou Massiva (areia, leite, cimento, refrigerante) e, como uma expressão literal, requer entidades Concretas. As expressões metafóricas, por sua vez, diferenciam-se das literais apenas por admitir também entidades Abstratas. Exemplos dessa construção apresentam entidades: contáveis-concretas (dólares, filhos, jogadores), contáveis-abstratas (votos, problemas, informações), massivas-concretas (gente), massivas-abstratas (injustiça, paciência, experiência).

Outros nomes também foram investigados quanto a suas restrições. A expressão literal *punhado de,* por exemplo, admite entidades Concretas, Contáveis e Massivas; com a restrição de que as entidades Contáveis sejam pequenas o suficiente para caber na mão em grande quantidade. Ex.: *moedas, pedras, anéis / farinha, terra, água, etc*. Já a expressão metafórica tem extrema preferência por entidades Contáveis (99% dos dados, embora seja possível prever combinações com nomes Massivos). As entidades podem ser Concretas, sem restrição de tamanho – na verdade precisam ser maiores para que seja possível diferenciar esta da Construção literal (estradas) – e Abstratas. Exemplos dessa seleção são: *pessoas, canções, notícias, votos, papéis, ideias, livros, etc*. Enquanto isso, *penca de* literal combina-se com nomes de frutas e flores, embora banana seja a única combinação encontrada. Admite, portanto, nomes Contáveis e Concretos. E *penca de* metafórico abre-se para entidades abstratas. Nos dados não há nomes Massivos, embora seja possível pensar em alguns. Há preferência (85% das ocorrências) por quantificar pessoas, como: *autores, garotos, irmãos, nomes, oradores, gigantes, craques, agregados, parlamentares, sorrisos, etc*.

Retomando o pressuposto acerca da continuidade entre linguagem literal e figurada, observa-se também casos em que uma mesma expressão pode ser interpretada de forma literal ou metafórica, como é o caso da sentença (107).

(107) Este excerto de 16 segundos diz mais sobre a política dos anos 60 do que **uma pilha de livros**. (CetenFolha)

Neste caso não há evidências tão claras a respeito de a estrutura referir-se a uma pilha literal. Esta pilha de livros pode ser hipotética e o falante pode estar somente comparando a relevância do excerto de 16 segundos a uma grande quantidade de livros. Isso pode tornar a

interpretação da estrutura ambígua em relação ao tipo de construção a que pertence (literal ou metafórica), no entanto isso pouco interfere na função comunicativa da mesma, uma vez que em ambos os casos ela compreende a noção de grande quantidade[24]. Ainda assim, casos limítrofes como esse são importantes pois reiteram o princípio de continuidade entre linguagem literal e figurada. Haveria, portanto, um *continuum* de interpretações, através de uma categoria difusa que englobaria os casos ambíguos, sendo necessário para tanto postular uma construção intermediária e delegar ao contexto o papel desambiguador.

5.5 A continuidade entre microconstruções e a construção de quantificação indefinida

Segundo a análise das microconstruções proposta por Boas (2010), seria possível estabelecer um *continuum* entre léxico e sintaxe na medida em que se considerem os padrões de valência de itens lexicais, relacionando-os às construções esquemáticas nas quais ocorrem. Além disso, é possível estabelecer relações entre ULs que, mesmo evocando frames distintos, possuam uma mesma configuração sintática (o que indica algo sobre a construção esquemática que instanciam).

Cada padrão de valência de uma UL associado a um frame representa uma perspectiva bastante particular da cena evocada. A análise das microconstruções permite relacionar ULs que evocam frames distintos, mas que se relacionam em diferentes níveis de abstração. Nesta abordagem não há a ideia de fusão entre item lexical e construção abstrata, como proposto por Goldberg. Isso é relevante em termos teóricos, pois não há separação entre léxico e gramática, uma vez que não se trata separadamente as entradas lexicais e as construções de estrutura argumental.

A abordagem de Goldberg para a Construção Ditransitiva seria inadequada, segundo Boas, por não captar as diferenças semânticas e sintáticas inerentes aos verbos que possuem sentidos relacionados (próximos), como evidenciam os exemplos (108-109).

(108) She told Jo a fairy tell

(109) *She assured Jo her love

Segundo a análise de Goldberg (1995), verbos de comunicação como *tell, wire, quote,* e *give* acionam, dentro da Construção Ditransitiva, o sentido de transferência, a partir da metáfora

24 Ainda é possível perceber a existência de uma comparação metonímica da mídia pelo conteúdo em relação à excerto,

do conduto. Boas pontua, entretanto, que verbos como *assure, advise*, *inform* e *notify,* embora semanticamente próximos aos verbos mencionados acima*,* não permitem que a metáfora licencie a Construção Ditransitiva, como se observa em (109). Para lidar com a diferença sutil entre estes verbos seria necessário considerar entradas lexicais mais refinadas e relacionadas à estrutura argumental.

Ainda de acordo com o pesquisador, cada padrão de valência individual associado a uma UL seria tomado como a parte formal de uma microconstrução: pareamento forma-sentido/função que se refere ao evento descrito pelo frame semântico tomando uma perspectiva específica.

No que se refere à estrutura de quantificação, é preciso que se comparem os seguintes pares:

i) caminhão de mudança, chuva de verão
ii) caminhão de areia, chuva de granizo
iii) caminhão de gente, chuva de reclamações

O primeiro par de expressões refere-se a uma estrutura qualitativa, ou seja, trata-se do tipo de caminhão e de chuva. O segundo par, por sua vez, flutua entre a interpretação qualitativa e a quantitativa (ainda que de maneira subjetiva e mais sutil), pois, como mencionado no início deste capítulo, *caminhão de areia* envolve uma grande quantidade de areia, do mesmo que *chuva de granizo*, uma grande quantidade de granizo; porém também se pode inferir que se trata do tipo de caminhão e do tipo de chuva, logo isso é uma questão de perspectiva. O terceiro par de expressões possui natureza apenas quantitativa, ou seja, trata-se de uma grande quantidade de gente e de reclamações.

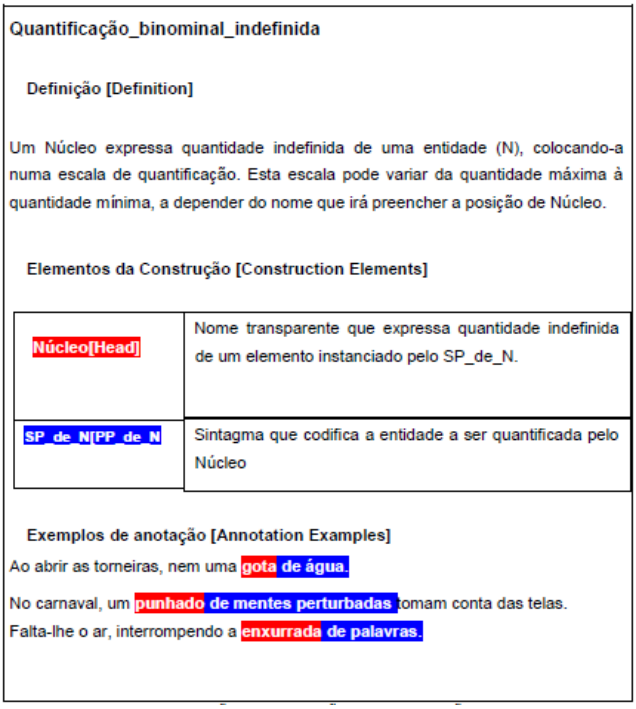
Tendo isso em vista, cabem algumas reflexões sobre o padrão de valência destas expressões. Em (i), *caminhão* e *chuva* evocam o frame de Precipitação e de Veículo e são Núcleos de uma estrutura de Modificação [N1 [de N2]], e o nome que ocupa o sintagma preposicional [de N2] funciona como Modificador. Em (ii), com a leitura qualitativa, o padrão semântico e sintático dos nomes será o mesmo de (i), porém será diferente caso a leitura seja quantitativa. Neste caso, *caminhão* (de areia) e *chuva* (de granizo) evocariam o frame de Quantidade Indefinida e possuiriam o mesmo padrão sintagmático que as expressões metafóricas em (iii), qual seja de Modificadores do padrão [N1 de [N2]]. Neste caso, é possível verificar a continuidade entre unidades lexicais, entendidas como microconstruções, e a construção de quantificação indefinida, uma vez que N1 aponta para a noção de quantidade, ainda que em uma configuração sintática diferente da construção.

Tais considerações são retomadas na próxima seção, quando tratamos da modelagem da construção, bem como dos frames evocados pelas ULs que dela participam como Nomes Quantificadores.

5.6 Remodelagem do Padrão de Quantificação Indefinida do português

Como observa Lage (2018), é extremamente relevante modelar as restrições que se aplicam às construções e tais restrições podem envolver desde a especificação do material lexical que deve preencher a construção até os aspectos semânticos que operam sobre ela. Tal modelagem tem a função de habilitar um *parser* a tomar como constructos apenas aqueles de fato licenciáveis pela construção. O Carma (MATOS ET AL., 2017) – analisador sintático-semântico alimentado pela base de dados da FrameNet Brasil – é capaz de "ler" as construções modeladas no Constructicon e, com isso, aprimorar as tarefas em PLN, no caso desta pesquisa em especial, as tarefas de Tradução por Máquina. Nesta seção analisaremos quais exigências são requeridas pelo padrão de quantificação para sua remodelagem no Constructicon.

Fala-se em remodelagem, pois a construção foi anteriormente incluída no Constructicon, como demonstra a Figura 31, durante pesquisa de mestrado (TAVARES, 2014), porém com os avanços da pesquisa sobre o padrão bem como do desenvolvimento da própria ferramenta e do *parser* construcional, verificou-se que o modelo proposto não era adequado para recobrir todas as peculiaridades da construção. Em 2014, o padrão [N1 [de N2]] foi modelado como uma Construção Binominal de Quantificação Indefinida produtiva ao nível dos *types*, uma vez que não se incluiu qualquer restrição para o preenchimento dos *slots.*



Quantificação_binominal_indefinida

Definição [Definition]

Um Núcleo expressa quantidade indefinida de uma entidade (N), colocando-a numa escala de quantificação. Esta escala pode variar da quantidade máxima à quantidade mínima, a depender do nome que irá preencher a posição de Núcleo.

Elementos da Construção [Construction Elements]

Núcleo[Head]	Nome transparente que expressa quantidade indefinida de um elemento instanciado pelo SP_de_N.
SP_de_N[PP_de_N	Sintagma que codifica a entidade a ser quantificada pelo Núcleo

Exemplos de anotação [Annotation Examples]

Ao abrir as torneiras, nem uma gota de água.

No carnaval, um punhado de mentes perturbadas tomam conta das telas.

Falta-lhe o ar, interrompendo a enxurrada de palavras.

Figura 31: Proposta de definição da Construção de Quantificação Binominal Indefinida no Constructicon[25]

Além disso, a decisão sobre a anotação da construção foi guiada pelos critérios de anotação definidos por Lage (2013). São eles: *(1) Sendo X um material lexicalmente especificado, existe X na construção em potencial?* De acordo com este primeiro critério, se o dado analisado não apresentar material lexicalmente especificado, o mesmo deverá ser tratado no Constructicon. No caso oposto, o analista deve partir para o segundo critério: *(2) Sendo F um frame e X um material lexicalmente especificado, X evoca F?* Sendo negativa a resposta, ou seja, caso este material lexicalmente especificado não evoque o *frame* esperado, a construção deverá ser anotada no Constructicon. Isto se justifica pelo fato de que X pode estar evocando um *frame* distinto daquele evocado pela construção e, por isso, a construção não poderia ser anotada no modo lexicográfico. Caso a resposta seja positiva, deve-se partir para o terceiro critério: *(3) Sendo F um frame e X um material lexicalmente especificado, X evoca F em outro padrão de valência?* Se mesmo com a mudança do padrão de valência X evocar F então a construção pode ser anotada em uma ferramenta lexicográfica; caso contrário, o tratamento deverá ser Construcional.

25 Apesar de reconhecermos, na análise que antecedeu a modelagem, que a estrutura de quantificação pressupõe uma reanálise do N1 como um quantificador N1 de X, quando da modelagem, foi mantida a separação N – SP, uma vez que ela encontrava respaldo nas demais construções modeladas no Constructicon. Retomaremos esta discussão mais adiante.

Já na aplicação do primeiro critério à construção, verificamos que a resposta era negativa. Isso devido à presença de algum grau de esquematicidade do padrão sintático em questão [N1 de [N2]], pois o mesmo não apresenta qualquer item lexical específico que possa ser considerado responsável pela evocação do frame de quantificação indefinida, que é, na verdade, evocado pela construção como um todo (TAVARES, 2014). Mas sabe-se agora que existem restrições para os subtipos da construção e que tais restrições variam conforme o tipo de N1 e o frame evocado pelo mesmo. Sabe-se também que não se trata de um padrão homogêneo, uma vez que os subtipos se distribuem num *continuum* entre padrões produtivos, semi-produtivos e idiomatizados. Assim, o problema não está em escolher o Constructicon como modelo de anotação, mas em não especificar todas as restrições e peculiaridades que acompanham a construção.

A FrameNet americana possui o frame Massa_quantificada (Quantified_mass), o qual apresenta diversos nomes comuns e adjetivos que quantificam massa e também entidades discretas. Dentre estes nomes, há diversos Nomes Quantificadores que aqui investigamos, são eles: *avalanche, enchente, enxurrada, punhado, monte, montanha, pilha, pitada, rio* e *inundação.* Esse foi o modo como a ferramenta desenvolvida em Berkeley resolveu a questão dos inúmeros nomes (bem distintos entre si semanticamente) que funcionam como quantificadores no inglês. Para nossa análise, no entanto, tal solução não é válida por dois motivos. Primeiramente, o foco desta tese recai justamente sobre o lugar da modelagem de expressões que se distribuem no *continuum* de padrões de cunhagem e construções (semi-)produtivas, logo, não seria aceitável maquiar tal análise, apenas colocando como evocadoras de um frame de quantificação unidades lexicais que só o fazem em um padrão muito específico. Em segundo lugar, procura-se investigar o frame básico destes NQs e observar os vestígios destes frames na construção de quantificação, uma vez que se assume a hipótese da Persistência da Imagem Conceptual (VERVECKKEN, 2012). Assim, cabe propor uma reanálise dos frames evocados por estes nomes. Além disso, a análise a partir do frame original de N1 não exclui o fato de que este também se relacione em alguma medida com o frame de Quantidade, pois, como visto anteriormente, existe uma continuidade entre a construção lexical e a construção de quantificação. Assim, nomes como *avalanche, enxurrada, tempestade* evocam o frame de Clima e todos envolvem um movimento massivo (ou seja, uma grande quantidade) de neve, água ou vento.

Por isso, a modelagem da construção de quantificação envolverá a modelagem dos frames básicos das ULs que dela participam. Propõe-se a modelagem dos 8 frames evocados por todos os 35 NQs, bem como das restrições acerca das ULs destes frames que de fato participam

da construção, uma vez que a modelagem a partir do frame pode levar a uma generalização sobre o tipo de N1 que não corresponde às possibilidades reais de preenchimento da construção.

Ilustrando a questão do parágrafo anterior, seguem algumas considerações sobre a modelagem dos frames de N1 tendo em vista a descrição destes pela FrameNet de Berkeley:

Agregados: há um grande número de ULs, alguns deles correspondem aos NQs aqui investigados (*bando, multidão, enxame, batalhão, penca*), porém é preciso que se realize a modelagem a nível da UL, pois há nomes como *tripulação, sexteto, círculo, escola, colheita*, dentre muitas outras que não funcionam na construção.

Locais_naturais: o frame não é suficiente para a delimitação do N1, pois possui ULs como: *caverna, praia, depressão*, dentre muitos outros, que não funcionam como quantificadores no padrão investigado. Assim, também é preciso que se modele as ULs *mar, montanha, oceano, galáxia, rio, floresta* e *mundo,* as quais, de fato, ocorrem na estrutura [N1 de [N2]].

Clima: um frame bastante genérico, pois é evocado por nomes como tempestade (de neve/de chuva/de raios), clima, chuva (de granizo), (brilho do) sol e adjetivos como claro, chuvoso. Possui Precipitação como Subframe (no qual incluiríamos *tempestade, enxurrada, enchente, dilúvio, inundação*), mas não é relacionado aos demais fenômenos da natureza (*onda, avalanche, vendaval*), assim propõe-se a modelam-se das ULs.

Contêiner: outro frame bastante genérico, o qual necessita ser modelado com as ULs *pá* e *poço* que participam da construção.

Veículo: dentre os NQs analisados apenas *caminhão* evoca este frame, porém é possível prever outros nomes evocadores deste frame (avião, navio, submarino) instanciando o padrão de quantificação. Como será explicitado ao final deste capítulo, a modelagem a nível da UL é capaz de ativar outros nomes que compartilhem das mesmas propriedades da UL de origem, criando a possibilidade de nomes próximos preencherem o padrão.

Medida_por_ação: comparado a Contêiner, este frame é mais específico e menor é a quantidade de ULs evocadoras do mesmo (*bite, pinch* e *splash*). Para duas destas ULs há equivalentes em português (*bocado* e *pitada*) e caberia adicionar a UL *punhado*. Neste caso específico, a modelagem pode ser feita a nível do Frame.

Impacto: o frame reúne em sua maioria ULs verbais: bater, quebrar, colidir, esmagar; e alguns poucos nomes: colisão, batida e impacto. Porrada, no entanto, foi o único mome encontrado no *corpus,* o qual também será modelado a nível da UL.

Parte todo: este é um frame não lexical, porém representa em um nível mais genérico as ULs participantes da construção de quantificação. *Pingo, gota* e *fiapo* não estão na base, enquanto *dedo* está anotado no frame Partes_do_corpo e *ponta* em Parte_orientacional, mas ambos evidenciam a parte (bem pequena) do todo e são herdeiros de Parte_todo. Por essa razão, a modelagem é feita a partir do frame genérico Parte_todo, para que se englobe todos os subframes.

No que se refere à modelagem da combinação entre NI e N2, ou seja, dos possíveis constructos do padrão de quantificação, depara-se com uma distribuição desigual no *continuum* entre padrões mais e menos esquemáticos. Ou seja, há NQs mais seletivos em sua combinação com N2, sendo tal restrição derivada da Persistência da Imagem Conceptual, o que aponta para a análise destas expressões como padrões de cunhagem. Por outro lado, há evidências em direção a uma análise mais esquemática, quando temos casos de NQs que ocorrem com N2 de diversos frames, como observado com os NQs que evocam Locais_naturais (*montanha, mar, mundo*, etc.), contribuindo para uma análise mais direcionada para a ponta construcional do *continuum*, nos termos da distinção proposta por Kay (2005; 2013). Em relação ao modelo computacional, é mais fácil modelar um padrão produtivo, pois as regras funcionarão para um grande número de Construções.

Consideremos novamente a análise por frames, porém observando agora a combinação entre o frame de N1 e os frames de N2. Dentre os possíveis padrões construcionais, foram identificados subtipos produtivos[26], semi-produtivos e padrões bastante restritivos, além de combinações de N1 e N2 já cristalizadas na língua.

Agregados (subtipos semi-produtivos): O grupo de NQs que evocam este frame é bastante coeso em relação ao tipo de N2 que seleciona: com exceção de *pilha*, os demais N1s combinam-se majoritariamente com os frames de Pessoas, Pessoa_por_característica, Pessoa_por_vocação, Pessoa_por_idade, Pessoa_por_origem, Pessoa_por_religião. Neste caso, a orientação do padrão

26 Cabe ressaltar que se trata de produtividade de type em relação a N2. A variação de nomes que ocupam N1 se dá por analogia, por isso não se considera aqui a noção de produtividade para N1.

é construcional, com a restrição de que o frame evocado por N2, num nível mais genérico, seja o de Pessoas (e seus herdeiros). Assim, identificam-se subtipos semi-produtivos de quantificação. Diante da pouca quantidade de frames de N2 neste grupo de Agregados, foi possível ilustrar tais restrições de N1 a partir da Tabela 8:

Frame de N2	N1							
	Batalhão	Bando	Multidão	Penca	Enxame	Corja	Pelotão	Pilha
Pessoas				2				
Pessoa_por_característica	3	14	17	3	2			2
Pessoa_por_vocação	22	7	9	2		3	1	2
Pessoa_por_idade		2	1					
Pessoa_por_origem		1	1					
Pessoa_por_religião			1					
Artefato								7
Outros			7	1	1			3

Tabela 8: Frames evocados por N2 em relação a N1 (Agregados)

Os frames Pessoa_por_característica e Pessoa_por_vocação ocorrem em 86 dos dados dentre as 113 ocorrências deste grupo, representando 76% dos frames evocados por N2. Nota-se que Multidão apresenta maior número de ocorrências de N2 evocando outros frames (Animais, Estradas, Texto, Edificações, Lei, Informação, Sentimento) e Pilha, que não tem preferência por N2 Pessoas, seleciona, na maioria das vezes, N2 relacionado a Artefato.

Ainda com relação ao NQ *pilha*, os N2s por ele selecionados – *papéis, livros, pessoas (desempregados, jornalistas), carros,* dentre outros – são também selecionados por outros NQs, porém o mesmo não ocorre com *nervos*, que forma uma expressão cristalizada com pilha: *pilha de nervos,* sendo quantificado apenas por este N1. Não se trata de uma entidade (nervos) que prefere ser quantificada por *pilha* e que eventualmente se combina com outros NQs, como ocorre com *dinheiro*, que é nitidamente frequente na combinação com *rio*, mas que também é

combinado com, pelo menos, outros seis N1s. *Nervos* forma com *pilha* uma collocation que, embora possa ser modificada por um processo criativo, natural e inevitável da língua (*poço de nervos*), faz com que qualquer modificação gere uma quebra de expectativa. Comparando tal análise àquela feita por Kay (2005), em relação ao padrão [Adj como SN] - que no PB geraria expressões como *verde como a grama*, *leve como uma pena*, *escuro como a noite*-, observa-se que, em ambas as análises, têm-se expressões idiomatizadas, com clara preferência combinatória e que, portanto, possuem um grau de produtividade *type* muito baixo.

A existência de um padrão de quantificação que engloba construções (semi)-produtivas e collocations com possibilidade de combinação muito restrita, como *pilha de nervos*, reitera a hipótese de que o padrão de quantificação se distribui no *continuum* entre construções mais esquemáticas e expressões idiomatizadas. Além disso, corrobora-se a hipótese a respeito do padrão geral de quantificação indefinida ser um padrão de cunhagem, por uma perspectiva genérica do padrão, ou seja, pela sua configuração esquemática como N1 de N2.

Locais_naturais (subtipos produtivos): Os N1s que evocam Locais_naturais combinam-se, numa análise baseada no grupo de N1s, com mais de 50 frames evocados por N2. Tais frames são distribuídos de modo proporcional entre os N2, havendo apenas alguns frames com percentagens mais significativas: o frame Texto representa 6% dos N2, porém está presente apenas na construção com *montanha*; Dinheiro representa 16% dos dados e 11% pertencem a *rio*; os 5% referentes ao frame Moralidade_avaliação são relativos à expressão *mar de corrupção,* que se repete algumas vezes no *corpus*.

A recorrência de algumas combinações (*montanha de papéis, rio de dinheiro, mar de corrupção*) sugere que alguns constructos possuem um grau de convencionalização maior em relação aos demais, e isso pode ser explicado, dentre outros fatores, pela relação conceptual que algumas destas expressões estabelecem entre si – a concepção de dinheiro como uma entidade fluida parece se combinar perfeitamente à fluidez de um rio. Ainda assim, *dinheiro* e *papeis* são frequentemente quantificados por outros N1s. Por outro lado, *mar de corrupção* é muito mais convencional do que qualquer outra combinação de N1 com *corrupção* e, de fato, não há ocorrência de outras combinações no *corpus*. Por analogia, seria possível se deparar com *oceano de corrupção* ou *mundo de corrupção,* embora tais expressões não sejam formas usuais de se quantificar e conceptualizar este N2.

A despeito da combinação preferencial entre *mar* e *corrupção*, os N1s pertencentes a este frame também apontam para um tratamento construcional do padrão N1 de N2, diante das diversas possibilidades de preenchimento de N2.

Clima (subtipos produtivos): Neste grupo também há pouco mais de 50 frames evocados por N2, com uma distribuição bastante equivalente dos frames, a não ser Ações_políticas que representa 9% dos dados, Dinheiro e Texto representam 8% cada. A saliência de tais frames pode se justificar pelo fato de que o *corpus* no qual tais N1s ocorrem com maior frequência seja um *corpus* jornalístico, os demais *corpora* (oral e ficção) não apresentam números tão significativos dos N1s que evocam o frame Clima. Este é um dos grupos de N1 com maior variedade tipológica de frames evocados por N2 e que não apresenta restrições aparentes para o preenchimento deste *slot,* sendo por isso um subtipo produtivo do padrão de quantificação.

Contêiner (subtipos semi-produtivos): neste frame as ULs *poço* e *pá* não selecionam seus N2s de modo tão coeso como os frames anteriores. *Poço* demonstra clara preferência por N2 que evoque o frame Propriedade_mental (60% dos dados); *pá,* por sua vez, combina-se mais frequentemente com Pessoas (75% dos casos). Isso aponta para dois padrões distintos, ambos evocam Contêiner, mas *poço* coloca a perspectiva sobre a profundidade do contêiner e *pá* é um tipo específico de contêiner, no caso um utensílio. No modelo que apresentado mais adiante, os padrões com *poço* e *pá* são subtipos do padrão genérico Quantificação_indefinida_Contêiner.

Veículo: *Caminhão,* muito pouco frequente no *corpus* (apenas 6 ocorrências), seleciona N2s de 5 frames diferentes, e como não foram identificadas restrições pela análise dos dados toma-se o padrão como produtivo.

Medida_por_ação (subtipos semi-produtivos): neste frame cada N1 seleciona seu N2 de modo particular. *Punhado* combina-se com Pessoas em 55% das vezes, assim uma das evidências adicionais na modelagem deste N1 é a de que Pessoas é um frame bastante provável para N2, mas também é possível notar uma diversidade tipológica de N2 neste padrão. *Bocado* possui uma característica distinta, o *corpus* demonstra a seleção dos nomes *coisas* (11 vezes), *gente* (8 vezes), *dinheiro* (8 vezes) e *tempo* (6 vezes), ou seja, de categorias superordenadas, num montante de 50 ocorrências. Há, no entanto, outros frames para N2 os quais não se repetem com a mesma frequência. Essa é uma característica que aponta para uma função discursiva específica, que é a de se combinar com nomes genéricos, enquanto que *punhado* alia-se a nomes mais específicos. A Tabela 9 ilustra tal distinção a partir do frame Pessoas, evocado por ambos os N1s.

N1 Punhado	Homens (3), gente (3), jornalistas (2), craques (1), assassinos (1), padres (1), empresários (1), telespectadores (1), escritores (1), sonhadores (1), candidatos (1), senadores (1), governadores (1), marginais (1).
N1 Bocado	Gente (8), prefeito (1).

Tabela 9: Quadro comparativo entre as ULs selecionadas por *punhado* e *bocado*

Pitada, por outro lado, seleciona N2 relacionado a Propriedade_mental (35%) e a Instrumentos e Estilos_musicais (24%), além disso, este NQ possui a função de expressar pequena quantidade. É perceptível que existam papéis específicos para cada N1, embora estes evoquem o mesmo frame. Por esse motivo, foram propostas duas construções, uma de pequena e outra de grade quantidade, ambas herdeiras da estrutura genérica de Quantidade_indefinida_Medida_por_ação.

Impacto: há apenas 6 ocorrências com o N1 *porrada*, o único encontrado no *corpus* para este frame, por isso limita-se a observação da restrição deste subtipo, que dentre as poucas ocorrências seleciona N2 de frames como: Pessoas, Entidade, Dinheiro e o frame Questionar.

Parte_todo (subtipos semi-produtivos): os N1s pertencentes a este frame são extremamente coerentes na seleção por N2 que, na grande maioria das vezes, evoca Sentimento (25%) e Propriedade_mental (21%). A porcentagem destes frames só não é maior porque alguns N1s também demonstram preferências específicas, ou seja, há boa porcentagem do frame Conversa, que é selecionado pela construção com o N1 *dedo* – vide novamente Gráfico 2. É notável, ainda, as particularidades de alguns N1s, como *ponta*, por exemplo, que tem uma seleção expressiva de N2s pertencentes aos frames Sentimento e Propriedade_mental (85% dos casos) e que se combina em 76% dos casos com N2 negativo.

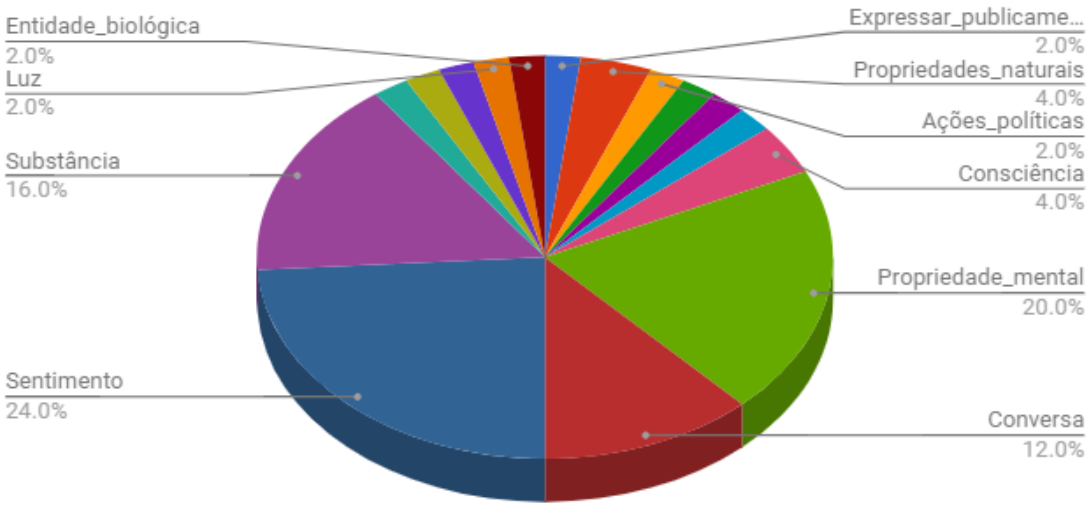


Gráfico 2: Correlação entre o frame de N1 Parte_todo e frames de N2

Dedo configura um caso excepcional dentre os N1s, uma vez que tem extrema preferência pela UL *prosa*. No sentido inverso, *prosa* também só parece se combinar com *dedo*, ao menos no universo de dados considerado para análise, e isto aponta para outra expressão cristalizada. Observem-se os exemplos (110-111).

(110) Passava um, parava para um **dedo de prosa** e aproveitava para tomar um refresco, que nunca faltava. (Corpus do Português)

(111) Dá-lhe crochê, tricô, **dedos de prosa**, conflitos familiares, amores de meia-idade, dificuldades financeiras. A empatia é imediata. (Corpus do Português)

Foi demonstrado na análise do frame Medida_por_ação que *bocado* também apontava para uma preferência por nomes específicos (coisa, gente, dinheiro e tempo), porém, diferente do que ocorre com *dedo*, tais nomes também são quantificados por outros N1s, cabendo, portanto, a distinção entre esses dois casos. Fica evidente no Gráfico 6 a preferência de *dedo* pelo frame de Conversa, o qual engloba, além de *prosa, as* ULs *palestra* e *conversa.* Diante disso, *dedo* é modelado separadamente dos demais N1s, por selecionar um tipo específico de frame para seu N2. Apesar desta predileção, *dedo* também seleciona N2 relacionado ao frame Propriedade_mental, o que é evidenciado no *corpus.*

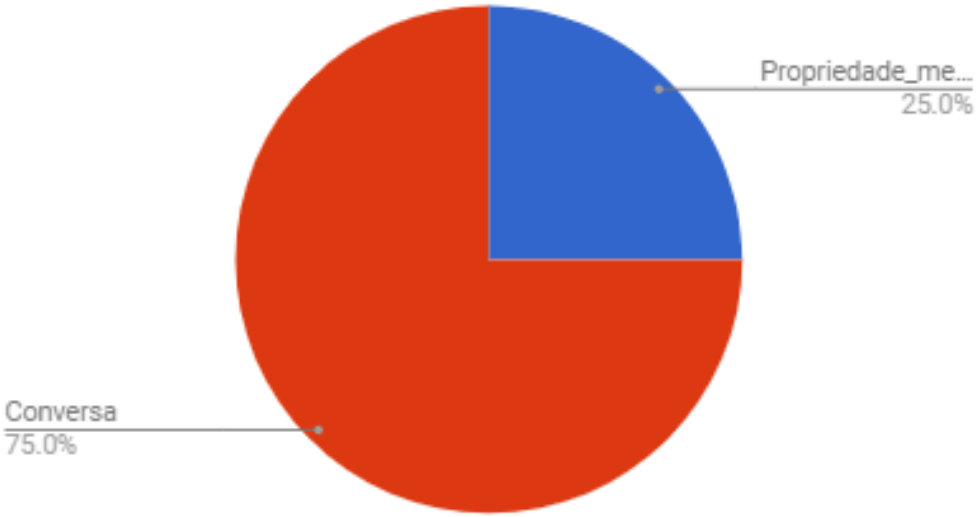


Gráfico 6: Correlação entre Dedo e frames selecionados por N2

Diante de tais particularidades de *dedo* em relação aos outros NQs, foram propostos dois padrões de quantificação indefinida: Quantificação_indefinida_dedo e Quantificação_indefinida_outros, ambos herdeiros de um padrão genérico de Quantificação_indefinida_Parte_todo.

Tendo em vista todas as considerações feitas em relação aos tipos de padrões encontrados nesta investigação, cabe discutir novamente o status da estrutura de quantificação como uma construção ou como um padrão de cunhagem. Sabe-se que a principal questão em torno do padrão de cunhagem é a não produtividade do mesmo. Assim, vimos que existem subtipos com uma produtividade alta de type, em relação a N2, e outros que são bem mais restritos na seleção do nome que preenche tal *slot,* sendo parcialmente produtivos. O que não está claro, até o momento, é a produtividade de type em relação a N1. Intuitivamente, sabe-se que o falante não pode simplesmente preencher o padrão N1 de N2 com qualquer nome do português, o que, segundo a perspectiva de Kay, já contribui para o tratamento daquele como padrão de cunhagem. É evidente que N2 é muito mais aberto nesse aspecto, afinal é possível, a priori, que qualquer entidade seja quantificada (o padrão aqui investigado apresenta exemplos interessantes disso: *rio de belezas, fiapo de voz, bocado de fisionomia, multidão de misericórdias*, etc.), mas, em relação a N1, há dois fatores a se considerar: i) não são todos os nomes do português capazes de ocupar o *slot* e ii) a combinação entre N1 e N2 também não é livre.

Antes que se trate tais expressões de quantificação como instanciações de um padrão de cunhagem, cumpre fazer algumas suposições acerca da possibilidade de se identificarem as restrições que governam o preenchimento do *slot* N1, para que se possa refutar a hipótese de que se trata de um padrão produtivo (ou uma construção, no sentido restrito de Kay). Caso fosse

possível delimitar as restrições aplicáveis à estrutura de quantificação com o objetivo de tratá-lo como um padrão produtivo, diríamos inicialmente que o *slot* N1 precisa ser preenchido por nomes que evoquem os frames anteriormente listados. Considere-se, por exemplo, o frame Veículo, o qual é evocado pelo N1 caminhão. Neste caso, observa-se que a medida expressa pelo contêiner que faz parte do veículo é focalizada, então se poderia pensar em outros veículos com medidas semelhantes ou maiores:

(112) Caminhão de dinheiro
(112a) ?Avião de dinheiro
(112b) ?Navio de dinheiro
(112c) ?Carreta de dinheiro
(112d) *Ônibus de dinheiro
(112e) *Metrô de dinheiro
(112f) *Bicicleta de dinheiro

Em (112) observa-se que não são todos os veículos habilitados a expressar grande quantidade nesta construção: *carreta, navio* e *avião* são prováveis de ocorrerem no padrão – de fato, há duas instâncias de *avião* e *navio* encontradas a partir de buscas feito no Google – e *ônibus* e *metrô* não são apropriados para preencher o *slot* N1, uma vez que não são veículos de carga. Por fim, *bicicleta* nem é um veículo de carga e nem possui um contêiner com dimensões escalares suficientes para torná-lo um NQ de grande quantidade indefinida.

Analisando agora os N1s *pá* e *poço*, estes evocam o frame Contêineres e figuram como quantificadores da CBQI, porém não são todos os contêineres que podem fazer o mesmo – alguns dos muitos exemplos que não funcionam são: *cesta, colher, xícara, caixa, garrafa, tigela, bolsa, tanque*. Em (113-114) observam-se os casos que funcionam no padrão juntamente a expressões criadas a partir de outras palavras (contêineres) próximas semanticamente.

(113) Pá de problemas
(113a) *Espátula de problemas
(113b) *Colher de problemas

(114) Poço de humildade
(114a) *Reservatório de humildade
(114b) *Cisterna de humildade

Em (113) evidencia-se que não são todos os utensílios que podem ser utilizados como NQ na construção de quantificação indefinida. As expressões de (114a) e (114b) também são rejeitadas pelos corpora utilizados como fonte de busca.

Com estas evidências, é possível perceber que a restrição ao nível do frame não é capaz de explicar o preenchimento de N1, então é preciso reformular a regra que governa tal preenchimento. Uma questão relevante observada a respeito dos NQs é que tais nomes já compreendem, de modo implícito, a noção de quantidade ou, nos termos de Verveckken (2012), implicaturas escalaras, basta observar os NQs que evocam o frame Clima: *tempestade, vendaval, enxurrada, enchente, dilúvio, onda, avalanche* e *inundação*.

Assim, parece que a restrição funcionaria em termos gerais, delimitando o preenchimento de N1 por nomes que além de evocarem o frame Clima (ou os demais frames levantados), tenham como background o frame de Grande_quantidade. Já em relação ao frame Parte_todo, que inclui NQs de pequena quantidade, o frame que estaria em segundo plano seria evidentemente o de Pequena_quantidade, pois tais nomes evidenciam de fato uma parte muito pequena de algo maior (*fiapo, ponta, pingo,* etc).

Tendo em vista essas restrições, e reconsiderando os N1s anteriormente analisados, nota-se que *caminhão* denota, em segundo plano, a noção de grande quantidade, considerando a capacidade de seu contêiner. Do mesmo modo, *poço* remete à noção de quantidade, quando a inversão da esquematização MAIS É PRA CIMA, que dá origem à MAIS É PRA BAIXO, relaciona profundidade à quantidade[27]. Em contrapartida, o NQ *pá* não aciona uma dimensão escalar que o coloque num nível alto na escala quantitativa, ou seja, não há neste caso a incorporação da ideia de grande quantidade[28]. Kay (2013) observa que é esperado que as construções gramaticais possuam restrições para sua aplicabilidade, mas que a generalização deve ser suficiente para que não se tenha que listar os casos que funcionam. Também, de acordo com Fillmore (1997): *"um processo gramatical ou padrão ou regra (ou construção') pode ser tido como produtivo se as condições de sua aplicabilidade não requerem uma lista de exceções"*. E o que se evidenciou aqui, com o exemplo de *pá*, é que existem exceções às restrições postuladas. Quanto às implicações para o modelo, tais considerações sugerem que a modelagem computacional do padrão deve especificar o que funciona e o que não funciona, de modo que se perde em generalidade (característica das construções gramaticais). Se anotássemos

27 O conceito de profundidade estaria inicialmente vinculado à intensidade e, como discute Silva (2010), também é intrínseca e cognitivamente motivada a relação entre intensidade e quantidade.

28 É evidente que comparado a colher, concha espátula, pá pode ser considerado um utensílio com proporções maiores, mas este critério não está sendo adotado aqui, uma vez que o NQ precisa possuir como plano de fundo a noção de grande quantidade.

apenas a restrição acerca do frame e da noção de Quantidade em segundo plano – o que já é um conceito bastante subjetivo para se tratar computacionalmente –, o modelo não seria capaz de reconhecer *pá (de gente),* por este utensílio não remeter à grande quantidade de algo, bem como *porrada (de gente)*, outro NQ que dificilmente seria compreendido em termos de Quantidade.

Se as expressões não são formadas livremente, então, nos termos de Kay (2005, 2013), não se trata de construção gramatical. Nesse ponto, cabe a ressalva de que o conceito de produtividade é inexistente dentro da teoria dos padrões de cunhagem – segundo Kay, numa construção gramatical não se deve considerar nem mesmo o processo parcial de produtividade – porém, como exposto neste capítulo, assume-se que alguns subtipos sejam mais produtivos/ semi-produtivos. Isso foi necessário diante da observação da heterogeneidade dentro do padrão, que possui subtipos consideravelmente abertos (produtivos) em relação ao *slot* N2 – *caminhão de N2*, subtipos que impõem algumas restrições (produtividade parcial) à seleção de N2 – *pá de N2 (pessoas)* – e casos em que os subtipos possuem formas cristalizadas – *pilha de nervos, dedo de prosa*. Então, quando se fala em produtividade, estamos nos referindo à produtividade de *type* de N2 em relação a N1, ou seja, avalia-se o tipo de nome com o qual cada N1 se combina e quão aberto é o *slot* N2. Assim, ao mesmo tempo em que tratamos a estrutura de quantificação indefinida como um padrão de cunhagem, admitimos a existência de subtipos com graus distintos de produtividade em relação a N2.

Assumindo-se que seja padrão de cunhagem, é preciso analisar o processo por trás da profusão de expressões de quantificação desse tipo, o que o autor identificou como neologismo analógico. A cunhagem de novas expressões pode ser explicada então pelo fenômeno da analogia. Os exemplos em (113-114) demonstraram que os sinônimos ou palavras semanticamente relacionadas a *pá* e *poço* foram rejeitadas pelo padrão, mas em relação a *caminhão* surgiram algumas possibilidades, como exibem os resultados da busca no Google com os nomes *navio, avião*, em (115-116).

(115) Carla, votos de Bom Natal e um 2012 com um **navio de coisas boas**!
Fonte: https://pt-br.facebook.com/permalink.php?story

(116) Sempre em frente que um caminhão, não um **avião de coisas boas** está chegando pra você!!!!!!!!!!!!!!!
Fonte: https://fabi-tudonovodenovo.blogspot.com/

Tais exemplos reiteram a hipótese de Kay acerca da possibilidade de criação de novas expressões por analogia com o que já funciona na língua. No caso, foram selecionados contêineres ainda maiores que o de um caminhão.

Dado o exposto acima, foi proposta a remodelagem da estrutura de quantificação indefinida, através de uma rede de construções, como demonstra a Figura 32.

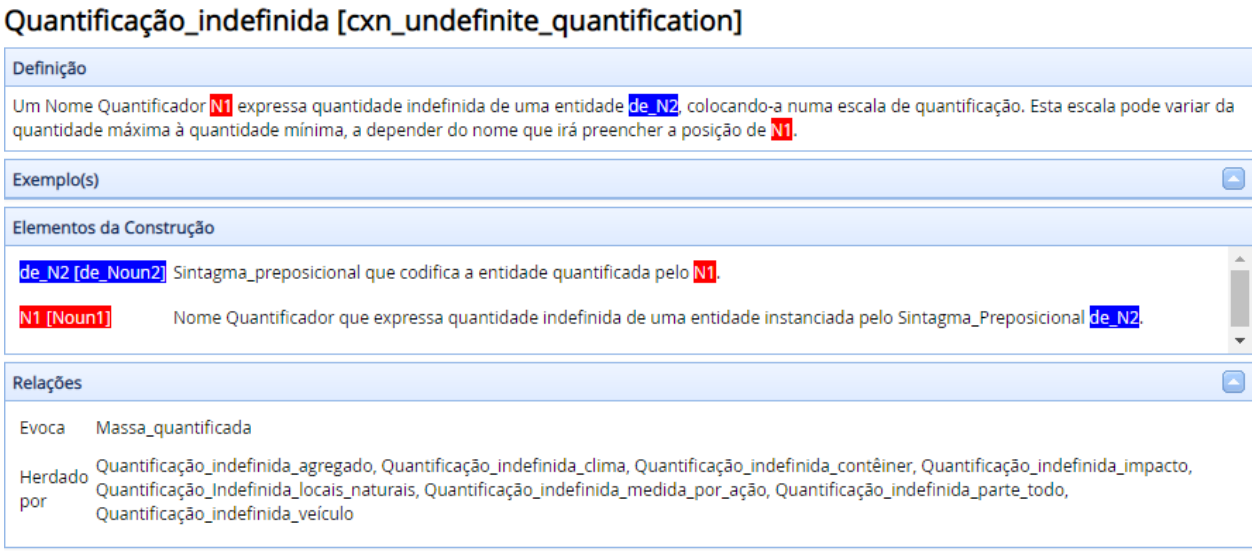


Figura 32: Anotação da construção genérica Quantificação_indefinida

Seguindo os parâmetros de modelagem do Constructicon, denominamos o padrão de cunhagem [N1 [de N2]] como construção genérica de Quantificação_indefinida. Tal construção é composta por dois ECs: N1 e de N2. Aqui apresenta-se a primeira adequação necessária quando da transposição das análises de dados para a modelagem: apesar de reconhecermos, seguindo as propostas anteriores, que a leitura de quantificação se correlaciona a uma reanálise da estrutura da construção para [N1 de [N2]], modelar um EC [N1 de] acarretaria na impossibilidade de apontar para as construções da gramática do português que licenciam cada signo filho da construção de Quantificação_indefinida e seus herdeiros, uma vez que não há, no Constructicon, um construção sintagmática do tipo [N Prep].

A construção genérica é herdada por oito subtipos, a saber:

a) Quantificação_indefinida_agregados: *penca de gente*

b) Quantificação_indefinida_clima: *enxurrada de críticas*

c) Quantificação_indefinida_locais_naturais: *montanha de problemas*

d) Quantificação_indefinida_veículo: *caminhão de filhos*

e) Quantificação_indefinida_impacto: *porrada de dinheiro*

f) Quantificação_indefinida_contêiner: (não lexical)

g) Quantificação_indefinida_medida_por_ação: (não lexical)

h) Quantificação_indefinida_parte_todo: (não lexical)

Os subtipos (f), (g) e (h) também são padrões genéricos que possuem como herdeiros:

f.1) Quantificação_indefinida_contêiner_profundidade: *poço de sabedoria*

f.2) Quantificação_indefinida_contêiner_utensílios: *pá de gente*

g.1) Quantificação_indefinida_medida_por_ação_pequena quantidade: *pitada de blues*

g.2) Quantificação_indefinida_medida_por_ação_grande_quantidade: *bocado de tempo*

h.1) Quantificação_indefinida_parte_todo_dedo: *dedo de prosa*

h.2) Quantificação_indefinida_parte_todo_outros: *fiapo de esperança*

A rede de padrões, ou construções segundo a terminologia do Constructicon, pode ser visualizada a partir do gráfico disponibilizado pela ferramenta, reproduzido na Figura 33.

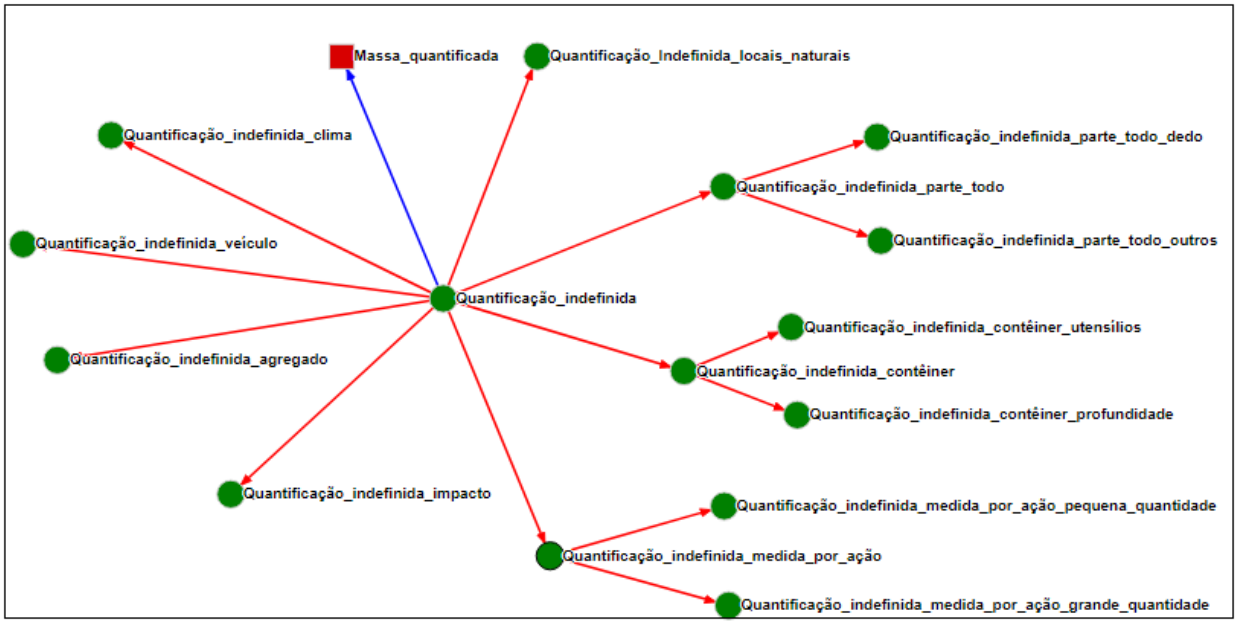


Figura 33: Rede de construções de Quantificação_indefinida no Constructicon

Para cada uma das construções, as restrições de preenchimento levantadas nesta tese e apresentadas no início desta seção foram modeladas, utilizando-se o editor de restrições do Constructicon, o qual é reproduzido na Figura 34.

O editor de restrições permite adicionar aos Elementos de Construção propriedades que são definidas segundo estruturas de dados disponíveis na base da FrameNet Brasil. As restrições CE > Construction, CE > before e CE > meets foram definidas por Almeida (2016) e têm por objetivo definir propriedades formais dos constituintes da construção. A primeira delas indica que o EC é licenciado, ele próprio, por uma construção já definida no Construticon, a qual

poderá ser identificada durante a anotação. A restrição CE > before determina que um EC deve aparecer antes de outro EC, como ocorre com o EC de_N2 que deve vir antes de N1. A terceira restrição, CE > meets, indica que um EC deve vir antes de outro (portanto esta e a restrição anterior são excludentes) e que não deve haver nenhum material interveniente entre eles. Esta restrição não se aplica ao padrão de Quantificação_indefinida.

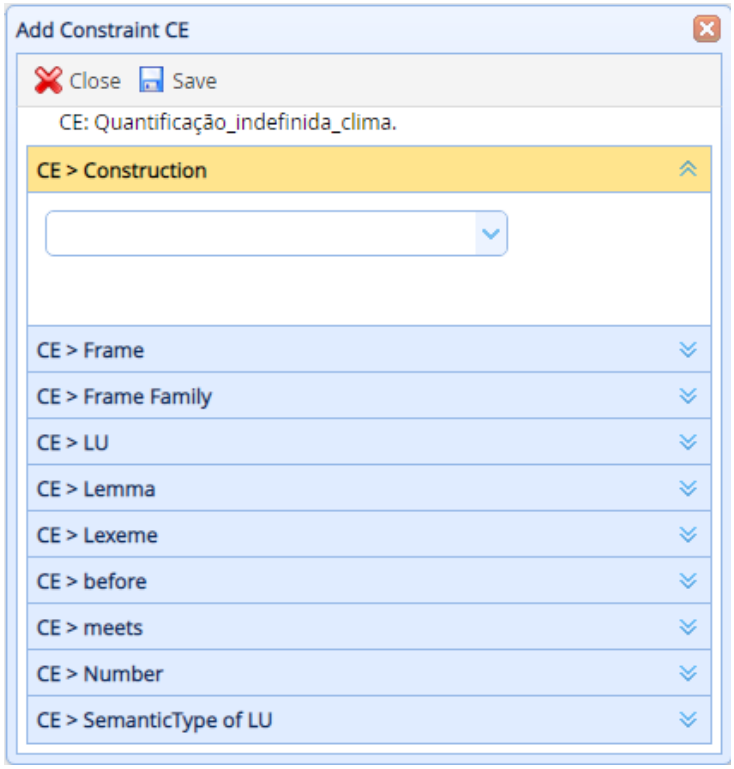


Figura 34: Editor de restrições do Constructicon

Já as restrições CE > Frame, CE > Frame Family, CE > Lexeme e CE > Lemma foram definidas por Lage (2018) e tratam das restrições de preenchimento dos *slots* construcionais. Com a primeira restrição é possível delimitar o preenchimento dos *slots* a partir de determinado frame, assim todas as ULs que evocam tal frame estarão habilitadas a instanciarem a construção. Esta é a maneira mais recorrente de delimitação do preenchimento de N2 no padrão [N1 [de N2]], para o padrão Quantificação_indefinida_dedo, por exemplo, utilizou-se a restrição CE > Frame_Conversa. O preenchimento do *slot* N2 é guiado, na maioria das vezes, por esta restrição, como o subtipo Quantificação_indefinida. A segunda restrição delimita o preenchimento dos *slots* de modo semelhante à primeira, porém a partir de uma família de frames, definida como o conjunto do frame que nomeia a família somado a todos aqueles que herdam dele direta ou indiretamente. Isso ocorre no *slot* N2 de alguns subtipos aqui investigados, como em Quantificação_indefinida_agregados, cuja restrição de preenchimento do *slot* N2 é CE > Frame

Family_Pessoas. Assim, tanto ULs evocadoras de Pessoas – tais como *pessoa.n, gente.n* – quanto aquelas evocadoras de Pessoas_por_vocação – tais como *médico.n* e *professor.n* – podem preencher esse *slot*. A terceira restrição se aplica aos casos em que apenas palavras específicas podem preencher o *slot*, como ocorre com o EC Preposição, que só pode ser preenchido por *de*, enquanto a quarta faz o mesmo para expressões polilexêmicas.

A restrição CE > LU foi implementada no âmbito desta tese e é semelhante às restrições CE > Lexeme e CE > Lemma. Porém, aqui, assume-se que parte do significado original do lexema quantificador (N1) ainda é mantido, ou seja, o frame por ele evocado ainda contribui com nuances de significado para a construção. Também cabe ressaltar que tal restrição permite que outras ULs sejam habilitadas a ocorrerem na construção por analogia e isso pode ser compreendido através do processo de ativação propagada, ou *spreading activation*, realizado pelo CARMA. Tal mecanismo, de modo semelhante ao processamento neural humano, busca conectar as informações através de nós (as unidades cognitivas), e após a ativação de determinado nó ocorre a propagação da ativação, ou seja, ativam-se outros que estejam relacionados ao nó de origem. Logo, a ativação de uma UL como *caminhão*, será propagada e ativará também, ainda que com um nível de ativação menor, ULs como *caminhonete, carreta, navio, avião,* permitindo que o processo de analogia, inerente a este padrão de cunhagem, seja modelado.

Também foram implementadas duas restrições que se mostraram relevantes para casos específicos da construção de quantificação: a restrição CE > Number para os casos em que o *slot* deve ser preenchido por nomes no Singular ou no Plural – isso ocorre nas construções que envolvem pequena quantidade, nas quais o N2 deve ser preenchido por nomes no singular –; e a restrição CE > Semantic Type of LU para os casos em que o preenchimento dos *slot* deve levar em conta o tipo semântico do Nome, se Negativo ou Positivo – como é o caso de Quantificação_indefinida_contêiner_profundidade, que seleciona N2 negativo (*poço de amargura*).

Assim, para a construção de Quantificação_indefinida_agregados, foram propostas as restrições mostradas na Figura 35.

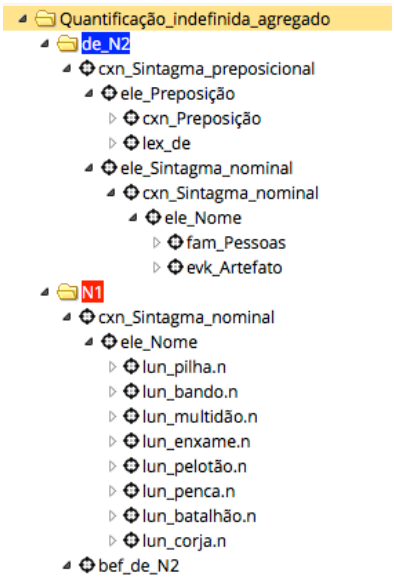


Figura 35: Restrições aplicadas à Quantificação_indefinida_agregados

O EC de_N2 possui a restrição CE > Constructicon Sintagma_preposicional por esta já ser uma construção definida na base de dados. Nela há dois signos filhos: o EC preposição, que possui a restrição CE > Constructicon Preposição, e que só pode ser instanciado por *de,* daí a utilização da restrição CE > Lexeme_de; e o EC Sintagma_nominal, já definido na base (CE > Constructicon_Sintagma_nominal), que possui um EC Nome para o qual são aplicadas as restrições de preenchimento dos *slots* CE > Frame_Family_Pessoas e CE > Frame_Artefato. Quanto ao EC N1 também foi aplicada a restrição CE > Constructicon_Sintagma_nominal, e o signo filho Nome, por sua vez, teve o preenchimento de seu *slot* restringido por CE > LU. A restrição CE > before também foi aplicada a N1, uma vez que este EC deve anteceder de_N2.

A segunda construção modelada, Quantificação_indefinida_clima, apresenta as restrições exibidas na Figura 36.

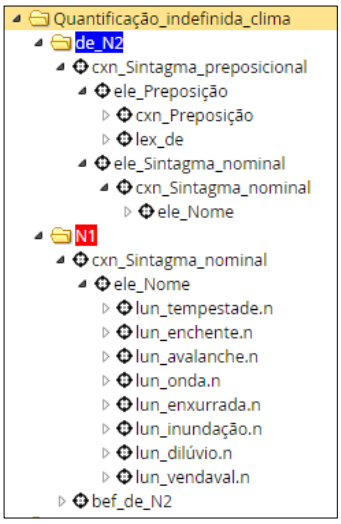


Figura 36: Restrições aplicadas à Quantificação_indefinida_clima

A modelagem desta construção seguiu os mesmos parâmetros da construção anterior, as diferenças encontram-se apenas em relação ao tipo de nome que preenche o EC N1 e ao fato de que o EC de_N2 não tem as restrições de preenchimento do *slot* modeladas, uma vez que as mesmas não foram identificadas.

A próxima construção modelada, Quantificação_indefinida_contêiner, é uma construção genérica, a qual possui duas construções herdeiras, todas apresentadas nas Figuras 37, 38 e 39.

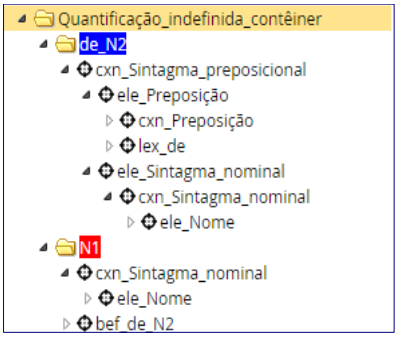


Figura 37: Restrições aplicadas à Quantificação_indefinida_contêiner

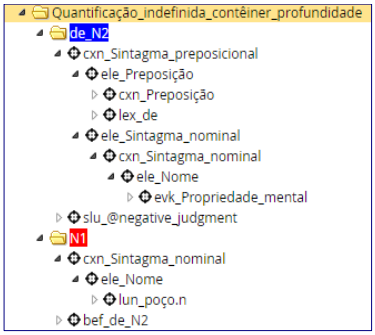


Figura 38: Restrições aplicadas à Quantificação_indefinida_contêiner_profundidade

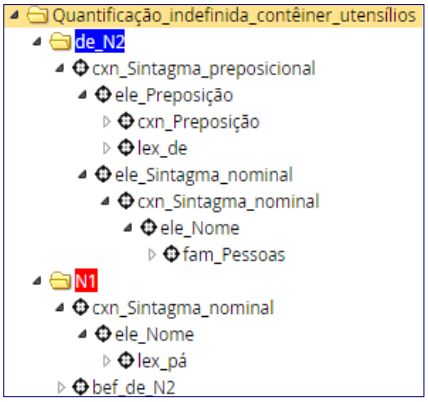


Figura 39: Restrições aplicadas à Quantificação_indefinida_contêiner_profundidade

Os signos filhos compartilham as restrições da construção genérica, mas diferenciam-se em relação ao preenchimento do EC N2 – Quantificação_indefinida_contêiner_profundidade

seleciona N2 que evoca o frame Propriedade_mental, e Quantificação_indefinida_contêiner_utensílios tem o *slot* de_N2 preenchido por nomes que evocam a família de Frames Pessoas. Além disso, a primeira também possui a restrição acerca do Tipo Semântico de N2, o qual, neste caso, envolve uma avaliação negativa, como em *poço de mediocridade, poço de hipocrisia* e *poço de rancor*. Outra ressalva necessária é a de que o EC N1 da segunda construção tem a restrição de preenchimento a partir do lexema, pois apenas *pá* parece instanciar o padrão, não havendo espaço para analogia neste caso.

A construção de Quantificação_indefinida_impacto possui as restrições exibidas pela Figura 40.

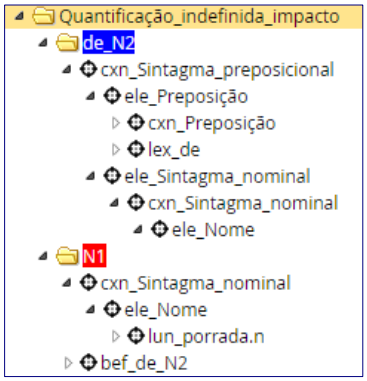


Figura 40: Restrições aplicadas à Quantificação_indefinida_impacto

A construção representada não possui restrições de preenchimento para N2 e tem como restrição para N1 a CE > LU_porrada, possibilitando que outras ULs também evocadoras do frame impacto (nomes como: *cacetada, bordoada, pancada*) possam preencher o *slot* por analogia.

A Figura 41 ilustra a modelagem da construção de Quantificação_indefinida_Locais_naturais.

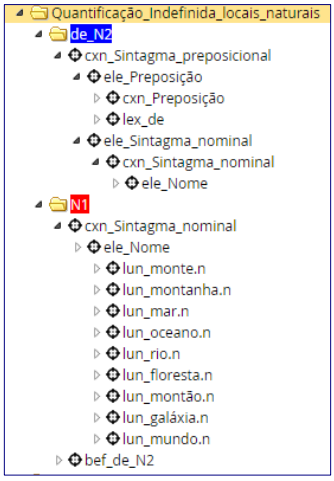


Figura 41: Restrições aplicadas à Quantificação_indefinida_Locais_naturais

Assim como na construção anterior, nesta também não há restrições modeladas para a seleção de N2, uma vez que são inúmeras as possibilidades de preenchimento do *slot*. A construção com N1 evocador de Locais_naturais, assim como Clima, possui alta frequência de type para N1, e embora muitos outros nomes possam ocupar o *slot* por analogia (*arquipélago, cachoeira, lagoa, selva, universo*, etc.) não é toda UL evocadora de Locais_naturais capaz de preenchê-lo adequadamente (*deserto, morro, parque, praia*, dentre outros), reiterando a necessidade de se implementar a restrição a nível da UL.

As Figuras 42, 43 e 44 apresentam a construção de Quantificação_indefinida_Medida_por_ação, genérica e suas construções herdeiras.

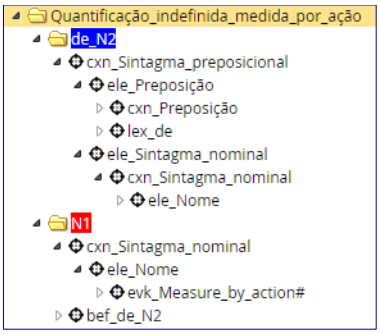


Figura 42: Restrições aplicadas à Quantificação_indefinida_Medida_por_ação

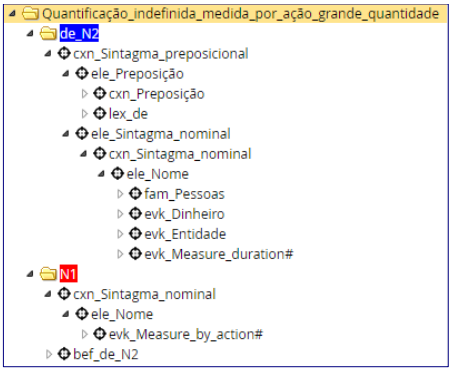


Figura 43: Restrições aplicadas à Quantificação_indefinida_Medida_por_ação_grande_quantidade

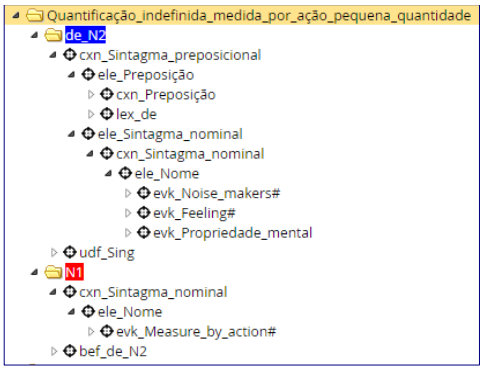


Figura 44: Restrições aplicadas à Quantificação_indefinida_Medida_por_ação_pequena_quantidade

As construções herdeiras compartilham as propriedades formais apresentadas na construção genérica e diferem-se quanto ao preenchimento dos *slots*. A construção de grande quantidade tem a restrição de preenchimento de N2 a partir dos frames: Pessoas, Dinheiro, Entidade e Medida_de_duração; já os frames evocados por N2 da construção de pequena quantidade são: Noise_makers, Sentimento, Propriedade_mental. Uma peculiaridade destas construções é o fato de a restrição sobre N1 ser feita a nível do frame, e não da UL, como havia sendo demonstrado para as outras construções. Isso se justifica pelo fato de o frame Medida_por_ação apresentar apenas ULs que instanciam regularmente o padrão (*bocado, punhado* e *pitada*), não sendo necessário especificar tais ULs, descartando também o processo de analogia, que não parece ser possível neste caso.

A Figura 45 apresenta a construção genérica Quantificação_indefinida_Parte_todo e as Figuras 46 e 47 as construções herdeiras, Quantificação_indefinida_Parte_todo_outros e Quantificação_indefinida_Parte_todo_dedo.

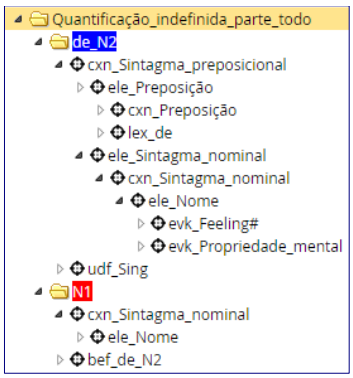


Figura 45: Restrições aplicadas à Quantificação_indefinida_parte_todo

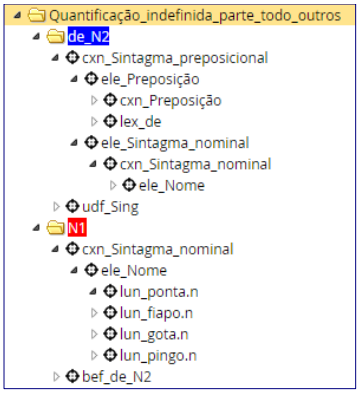


Figura 46: Restrições aplicadas à Quantificação_indefinida_parte_todo_outros

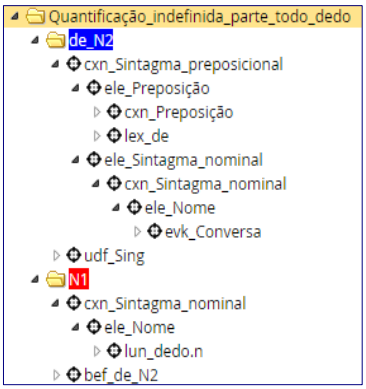


Figura 47: Restrições aplicadas à Quantificação_indefinida_parte_todo_dedo

A construção genérica já especifica os frames evocados por N2 – CE > Frame_Feeling e CE > Frame_ Propriedade_mental –, pois as construções herdeiras compartilham tais restrições, a construção com N1 *dedo*, entretanto, também possui N2 que evoca o frame Conversa, por isso a separação da construção de quantificação_indefinida_parte_todo em dois padrões distintos. Todas as construções compartilham a restrição CE > Número, tendo a exigência de N2 ser um nome Singular.

A última construção modelada é apresentada na Figura 48.

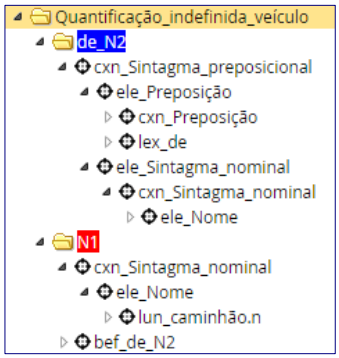


Figura 48: Restrições aplicadas à Quantificação_indefinida_veículo

A construção com N1 evocador de veículo não apresenta restrição de preenchimento pelos mesmos motivos de algumas construções anteriores, isto é, por não ser possível identificar padrões de seleção de N2. A restrição CE > LU_caminhão.n, por sua vez, possibilita que outros veículos possam ocorrer no padrão.

A modelagem de um padrão tão heterogêneo como o de quantificação indefinida levou em conta as preferências observadas por subtipo acerca das restrições de preenchimento dos *slots,* mas cabe destacar que tal modelo não esgota todas as possibilidades de preenchimento. Esta é uma dificuldade imposta pelos padrões de cunhagem, determinar com precisão aquilo que

é possível, sem a necessidade de se criar uma lista dos casos que funcionam. A estrutura de quantificação não nos permite determinar as regras de preenchimento de N1 e muito menos de N2, ela oferece algumas pistas, as quais foram modeladas como restrições *soft*, ou evidências adicionais, e que não devem, portanto, ser entendidas como restrições de fato.

O quadro a seguir apresenta um resumo das propostas deste capítulo.

Resumo do capítulo	
Análise das Correlações entre Frames de N1 e N2.	• A análise da correlação entre frames de N1 e N2 revelou a heterogeneidade do padrão [N1 [de N2]], uma vez que identificaram-se subtipos mais produtivos e subtipos mais restritos em termos da seleção do frame de N2. • Apesar de se identificar a persistência da imagem conceptual de N1 sobre o significado da construção, influenciando, de certa forma, a seleção de N2 – que precisa se ajustar à conceptualização de N1 –, também se observa a expansão da categoria e de combinações não esperadas ou harmônicas entre N1 e N2.
Resultados do Teste de Julgamento de Aceitabilidade	• O teste de julgamento por falantes nativos revelou a limitação da expansão do padrão de quantificação na língua, através da baixa taxa de aceitação de novos usos.
Extensão metafórica entre as expressões de quantificação indefinida	• Postulou-se uma Construção de Medida Subjetiva, da qual herdam as construções de Quantificação Indefinida Literal e Quantificação Indefinida Figurada. • Existe um continuum de interpretações entre a construção literal e a construção figurada;
A continuidade entre microconstruções e a construção de quantificação indefinida	• Postulou-se a continuidade entre a construção e seus elementos constituintes; • Os nomes que preenchem N1 apontam para o frame de Quantidade Indefinida, ainda que em uma configuração sintática distinta, como na Construção de Modificação.

Remodelagem do padrão de Quantificação Indefinida	• Postularam-se diferentes níveis de produtividade em relação ao type de N2, devido ao continuum entre padrões mais esquemáticos e padrões mais idiomatizados; • O processo de analogia atua sobre N1 e é responsável pela expansão do padrão na língua; • Modelaram-se 8 padrões de cunhagem de quantificação indefinida e suas restrições sintático-semânticas. • Implementaram-se as restrições de preenchimento dos *slots:* a) a restrição CE> LU possibilitou a modelagem do processo de analogia por ativar indiretamente o frame lexical, assim como outras ULs evocadoras deste frame; b) CE> Number é responsável por modelar a restrição de que N1s de pequena quantidade selecionam N2 singular; c) e a restrição CE > Semantic Type dá conta da seleção de N1 por N2 com julgamento negativo.

6 IMPLICAÇÕES PARA A TRADUÇÃO POR MÁQUINA

6.1 O papel de *Frame*s e Construções na Tradução por Máquina

Uma das principais razões para se empregarem os princípios da Semântica de Frames no desenvolvimento de recursos que envolvam Processamento de Línguas Naturais, como a Tradução Automática, é o fato de ela colocar o significado no centro de sua teorização.

Em relação à tradução feita por tradutores humanos, os *frame*s são o *background* necessário ao tradutor para que este escolha a melhor tradução do texto-fonte. O processo de desambiguação de sentido então é feito através do próprio conhecimento do tradutor acerca dos *frame*s evocados pelo texto. Este processo pode ser relativamente mais simples, se envolver os chamados *frame*s universais, ou mais complexo, ao compreender *frame*s específicos de uma cultura. O *frame* de Relacionamentos_pessoais, por exemplo, é estruturado de modo diferente no Inglês e no Alemão e, por isso, uma expressão inglesa como *sugar daddy* não apresenta uma contraparte exata no Alemão, sendo necessário que a tradução seja feita por uma longa paráfrase como: *spendabler älterer Mann, der ein junges Machenaushalt* 'homem mais velho generoso que sustenta garota jovem' (BOAS, 2013, p. 150).

Outro conceito fundamental para a compreensão do processo de tradução é o de *Construção*. Como vimos, os *frames* guiam o tradutor na escolha do significado do que está sendo traduzido, mas não é só o polo do sentido que está sendo considerado neste processo, na verdade, uma tradução eficiente deve levar em conta a relação entre informação gramatical (forma) e significado. Isso, pois o tradutor frequentemente se depara com expressões que não podem ser traduzidas de modo composicional ou considerando-se apenas a forma. A expressão *chuva de reclamações*, por exemplo, se tratada composicionalmente, será tomada como um fenômeno da natureza inexistente, mas, se tratada como uma construção, será traduzida adequadamente como uma expressão de grande quantidade. Nesse sentido, a tradução deve levar em conta o caráter simbólico da linguagem, ou seja, o pareamento forma-sentido/função que dá origem às construções de uma língua. O tradutor, por sua vez, precisa ter conhecimento das construções que compõem as línguas envolvidas na tradução.

O processamento cognitivo envolvido numa tradução feita por humanos é diretamente responsável pela qualidade dos resultados oferecidos, por isso, tais bases de conhecimento (*frame*s e construções) também são indispensáveis num sistema de tradução automática. Mas como simular tal processamento cognitivo numa máquina? A implementação computacional da Semântica de *Frame*s, a FrameNet, é um recurso que oferece representações computacionais das

estruturas cognitivas essenciais na construção do sentido – os *frames* –, auxiliando assim o processamento de línguas naturais pelo computador. Outro desdobramento desta teoria é o Constructicon, recurso que visa ao tratamento computacional das construções de uma língua, servindo também como base de conhecimento para ferramentas que envolvem PLN.

Matos (2014), investigando a desambiguação de itens lexicais, afirma que um computador é capaz de realizar uma tarefa complexa como a desambiguação lexical, desde que sustentado por uma base de dados como a FrameNet. Nossa hipótese é a de que tais bases de conhecimento (*frame*s e construções) possam alimentar sistemas de tradução automática e interferir nos resultados de sistemas estatísticos através de um modelo híbrido, contribuindo para a construção de equivalentes de tradução.

A FrameNet também pode funcionar como um recurso multilíngue e alguns trabalhos (FILLMORE & ATKINS, 2000; PETRUCK & BOAS, 2003; BOAS, 2002, 2003, 2005a, 2013) vêm demonstrado essa funcionalidade da ferramenta. Boas (2013) sugere que a Semântica de *Frames* possa ser utilizada para a análise e descrição de outras línguas, tendo como base a informação já descrita para o inglês, na FrameNet de Berkeley. Assim como a teoria baseada em *frame*s é imprescindível para a tradução feita por humanos, ela pode ser de grande utilidade para os casos de tradução automática.

O pesquisador apresenta como se dá a correlação entre fragmentos do léxico de línguas como o inglês e o alemão. O processo inicia-se pela identificação da lista de ULs do inglês evocadoras de determinado *frame* e, em seguida, busca-se pelas traduções equivalentes no alemão. Assim, uma UL do inglês como *argue* evoca o frame Communication_conversation, o qual descreve uma situação na qual as partes envolvidas trocam informações sobre determinado tópico, e tem como participantes (EFs) centrais o Interlocutor e o Tópico. A partir de dicionários e *corpora* eletrônicos é possível então identificar o verbo equivalente de *argue* no alemão, o verbo *streiten.* Tendo identificado os equivalentes, é necessário que se busque por sentenças no *corpus* que atestem o uso de cada *frame* sintático associado a uma UL do alemão.

Mas encontrar equivalentes ainda não tornará a tradução eficiente se não forem criadas também entradas lexicais das ULs do alemão que sejam paralelas a sua contraparte no inglês. Segundo Boas (2013), tais entradas são relevantes neste processo, pois contêm listas exaustivas de combinações de propriedades sintáticas e semânticas. E é a partir dessa quantidade de informação que se consegue correlacionar adequadamente a entrada lexical da língua-fonte com a língua-alvo. As Figuras 49 e 50 são exemplos preliminares das entradas produzidas para o alemão a partir do inglês, com as possibilidades combinatórias de cada entrada lexical.

Interlocutors	TARGET	Topic
NP.Ext	argue.v	INI
NP.Ext	argue.v	PP_over.Comp
NP.Ext	argue.v	PP_about.Comp
NP.Ext	argue.v	PPing_about.Comp
NP.Ext	argue.v	Swhether.Comp

Figura 49: Entrada lexical parcial do verbo *argue* (BOAS, 2013, p. 133)

Interlocutors	TARGET	Topic
NP.Ext	streiten.v	INI
NP.Ext	streiten.v	PP_um.Comp
NP.Ext	streiten.v	PP_über.Comp

Figura 50: Entrada lexical parcial do verbo *streiten* (BOAS, 2013, p. 133)

Diferentemente de dicionários cujas entradas lexicais são organizadas por ordem alfabética, a FrameNet organiza a informação lexical baseada em *frames* de maneira sistemática e, ao funcionar como recurso multilíngue, pode facilitar muito o próprio trabalho humano de tradução.

A próxima seção lidará especificamente com as implicações de se adotar a teoria dos padrões de cunhagem para uma abordagem computacional, com enfoque na Tradução Automática.

6.2 Padrões de cunhagem e sistemas híbridos de tradução

A qualidade da Tradução por Máquina tem melhorado bastante graças às pesquisas em Tradução por Máquina Estatística baseada em sintagma, assim como a disponibilidade e diversidade de *corpora* paralelos, os quais são necessários para o treinamento de modelos estatísticos (ARCAN, 2017). Como já mencionado no capítulo 1, tais modelos operam com sequências de palavras ou sintagmas, reduzindo o espaço amostral e consequentemente a ambiguidade na tradução, uma vez que sequências de palavras juntas podem oferecer um contexto maior para a análise probabilística. Contudo, o processo de decodificação estatístico pode levar o sistema a um erro de busca (KOEHN, 2010), quando o tradutor falha ao buscar pela melhor tradução, o que geralmente ocorre quando a tradução escolhida (a mais frequente) não corresponde à tradução mais adequada. O que este trabalho vem afirmando é que a tradução menos frequente também precisa ser considerada, ao menos quando se tem um padrão de

cunhagem como o de quantificação indefinida, o qual não tem sido traduzido adequadamente na maioria das vezes.

O sistema falha quando alguns fenômenos linguísticos surgem na língua, mas não possuem alta frequência de ocorrência. Como observa Bybee (2010), os *prefabs* não precisam ser altamente frequentes para se tornarem convencionais, bastando que tenhamos contato com a nova forma poucas vezes para que ela se integre ao nosso conhecimento linguístico. Daí expressões possíveis, mas raras, como *poço de nervos* dificilmente seriam tratadas pelo sistema de tradução adequadamente, ou seja, como uma expressão de quantificação indefinida, pois o tradutor lida melhor com aquilo que é estatisticamente relevante (*pilha de nervos*), ainda que com dificuldades, conforme mostram as Figuras 51 e 52.

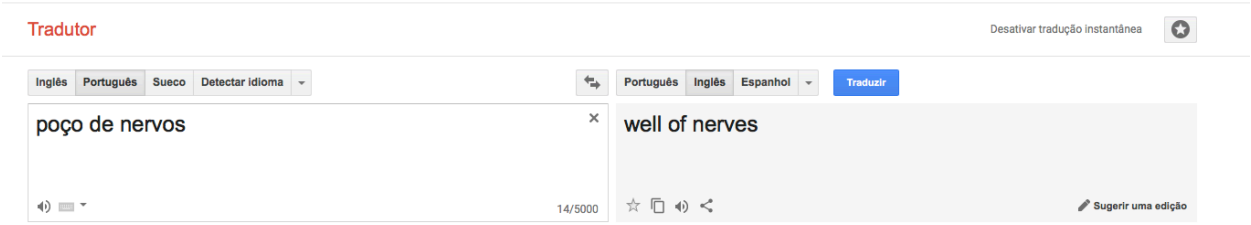


Figura 51: Tradução de “poço de nervos” no Google Tradutor

Figura 52: Tradução de “pilha de nervos” no Google Tradutor

O resultado obtido com a tradução de *poço de nervos* na Figura 53 apresenta a tradução literal da expressão, uma expressão não existente no inglês para a expressão de grande quantidade. Na Figura 54, a probabilidade, peça chave na tradução por máquina, ranqueia a tradução correta de *pilha de nervos* (nervous wreck) como a segunda opção, dada a ambiguidade de cell em inglês, que pode se referir a uma célula biológica (que co-ocorre muito frequentemente como nervo) ou a uma célula de energia, cujo sinônimo em Pt-Br é pilha, também ambígua.

Observa-se, então, que uma construção produtiva, ou genérica o suficiente para gerar inúmeras expressões na língua, possivelmente será melhor tratada pelos sistemas atuais de tradução, uma vez que suas instâncias têm maiores chances de serem frequentes nos corpora de

treinamento; enquanto expressões geradas por processos analógicos, como o que se identificou com os padrões de cunhagem, não possuem a mesma frequência de ocorrência, o que dificulta seu tratamento por estes sistemas.

A Tradução por Máquina Estatística (SMT) procura resolver este impasse criando sistemas híbridos, ou seja, incorporando informações acerca do modelo de língua baseado em regras, reconhecendo os fenômenos que têm comportamento próprio e buscando compor uma lista de correspondências (regras), como ocorre com o sistema de data, cujo formato pode sofrer muitas variações conforme a língua. Assim, as regras converteriam um sistema de data do português para o sistema de data do inglês, por exemplo, em vez de operar simplesmente com o modelo estatístico.

Essa também parece ser a solução para o objeto desta tese, uma vez que o mesmo não apresenta comportamento típico de construções gramaticais, ou seja, não possui alta frequência de *type* – logo não é produtivo –, é um padrão de difícil regularização e não permite a criação de novas expressões livremente. Sendo assim, a estrutura não pode ser derivada estatisticamente do *corpus*, diante das restrições que não seguem um padrão capaz de ser generalizado a partir da análise e das descobertas feitas neste estudo. Ela também não será derivada por esse modelo porque nos textos paralelos utilizados como referência para o tradutor nem sempre o que é mais frequente é a tradução mais adequada. O exemplo (117) e as traduções (117a-b) também ilustram tal dificuldade

(117) caminhão de notícias (muitas notícias)
(117a) *bunch of / a heap of / lots of news* (muitas notícias)
(117b) *news truck* (tipo de veículo utilizado especificamente por empresas de comunicação?).

Observa-se que a tradução (117a) era a esperada para a expressão, porém o que se obtém pelos sistemas atuais de tradução, como o *Google Translate*, são traduções como a de (117b). Parece razoável que expressões como *a bunch of, a heap of* e *lots of* sejam bastante frequentes no *corpus* de treinamento, afinal são expressões de quantificação bastante usuais e genéricas da língua inglesa, porém quando se trata da tradução de *caminhão de* pode-se inferir que tal expressão é recorrentemente traduzida por *truck* e não pelas expressões de quantificação mencionadas, justificando a escolha do tradutor por *truck news* em vez de *lots of news*, por exemplo. Por essa razão, a construção aqui investigada não pode ser derivada estatisticamente,

tendo em vista a dificuldade do sistema em reconhecer o item pouco frequente (ou até mesmo inexistente nos *corpora* paralelos) como a tradução mais adequada.

Tendo em vista a fragilidade do sistema de tradução estatística, reafirma-se a necessidade de se tratar a estrutura de quantificação como padrão de cunhagem, o qual não se localiza entre as construções genéricas e nem entre as construções lexicais, mas entre os fenômenos que se encontram no meio do caminho, isto porque não são completamente produtivos, mas também não são itens lexicais que precisam ser aprendidos individualmente e memorizados. O módulo híbrido, ou seja, aquele que integra ao modelo estatístico um modelo baseado em regras, parece ser adequado para este fenômeno ao interferir no resultado do tradutor, apontando para a tradução apropriada. Tal resultado também pode ser obtido através da injeção terminológica, que consiste em sugerir ao tradutor o equivalente de tradução, quando este não é derivado do sistema.

A proposta de solução vislumbrada, mas não implementada por razões de prazo nesta tese, é fazer uso do modelo que aqui foi proposto em um sistema de hibridização da tradução estatística. Porém, ao invés de se usarem regras fechadas ou uma lista de correspondências, poder-se-ia explorar justamente a rede de frames e construções da FN-Br.

O CARMA (MATOS ET AL., 2017), *parser* que opera a partir da base da dados da FrameNet Brasil, é um analisador construcional e por isso é capaz de reconhecer as construções modeladas no Constructicon. Ao se deparar com uma expressão como *monte de filhos,* o *parser* processaria a informação de que *monte* pertence à estrutura de quantificação modelada – assim, o frame evocado será o de Massa_quantificada –, e não ao *monte* evocador de Locais_naturais. Desse modo, no processo de desambiguação da tradução, o analisador será responsável por realizar a injeção do termo equivalente em inglês – *lots of kids*.

Enquanto a injeção terminológica simples, que consiste na substituição de termos a partir de uma lista fechada – um dicionário terminológico – não tem a capacidade de ser genérico, nem flexível, o modelo proposto neste trabalho avança no sentido de que é uma rede capaz de reconhecer um padrão polilexêmico como o [N1 [de N2]], trabalhar com restrições *soft* e abrir espaço para a analogia, modelando, assim, um importante princípio da Linguística Cognitiva (BYBEE, 2010).

Atualmente, a equipe de TI da FN-Br está desenvolvendo o sistema de hibridização que faz uso da base de dados lexicais e construcionais criada para melhorar os resultados de tradutores estatísticos. Em futuros trabalhos, pretende-se avaliar, segundo as métricas padrão da área, a contribuição do modelo aqui proposto para a tradução.

7 CONCLUSÃO

Na Introdução deste trabalho, foram apresentados os problemas motivadores, a hipótese e seus dois objetivos principais, os quais são reformulados através das perguntas (i) e (ii).

(i) A abordagem estrita de gramática apresentada por Kay (2005, 2013), acera dos padrões de cunhagem, se aplica às expressões de quantificação indefinida do tipo *enxurrada de protestos, avalanche de notícias, porrada de gente*? E qual é a implicação desta teoria para o tratamento computacional que, por vezes, tem objetivos distintos de uma gramática das construções?

A revisão da literatura acerca da Gramática das Construções reiterou a importância de se adotar os princípios da continuidade entre gramática e léxico e a postulação de uma rede de construções, interligadas via links de herança. Tais princípios governaram em muitas das vezes as análises empreendidas nesta tese. Contudo, uma abordagem bastante estrita da gramática também foi considerada nesta revisão, a abordagem de Kay (2005, 2013), segundo a qual deve-se considerar como construção apenas a quantidade mínima de informação que o falante precisa ter para que seja capaz de entender e produzir sentenças da língua. Isto é, assume-se que as construções de uma língua sejam apenas estes padrões mais gerais e produtivos.

A concepção de que o conhecimento linguístico é formado por um *continuum* entre léxico e gramática, que abarca desde estruturas como palavras até as regras mais gerais da gramática, não é incompatível com a proposta de Kay sobre a separação daquilo que é ou não é construção gramatical. **Neste trabalho, embora se adote o termo padrão de cunhagem, admite-se que tal estrutura localiza-se nesse *continuum*, e no final das contas seja também uma construção, mas que ora se aproxima do léxico (subtipos menos produtivos) ora se aproxima das construções gramaticais, daí a necessidade de se adotar uma terminologia apropriada e um tratamento diferenciado para a mesma**.

Nesse sentido, a estrutura de quantificação indefinida foi tratada como um padrão de cunhagem, pelos motivos apresentados em (a) e (b):

a) N1 não é um *slot* aberto que permite a instanciação de qualquer Nome do português. O preenchimento do *slot* por novos nomes não se dá por um processo produtivo e, sim, analógico;

b) os subtipos que herdam da construção genérica de quantificação indefinida são bastante heterogêneos entre si e, mais importante, as idiossincrasias também estão presentes dentro dos próprios subtipos;

Nesse ponto, evidencia-se a relevância da abordagem de Kay para o propósito da modelagem computacional de construções. Uma vez que não foi possível postular generalizações ao padrão [N1 de [N2]] que recobrissem todas as particularidades do mesmo e seus subtipos, procedeu-se à remodelagem da estrutura, que por não ser produtiva, não pode ser modelada da mesma maneira que uma construção gramatical.

(ii) Por que a utilização da base de dados de um Constructicon pode ser interessante para um modelo de tradução automática? E qual a implicação da teoria dos padrões de cunhagem para tais sistemas?

Os sistemas automáticos de tradução são frágeis ao lidar com expressões pouco frequentes nos textos treinados para tradução. Além disso, seu modelo baseado em regras opera através de regras relativamente simples, quando comparado a um Constructicon. Alguns casos específicos, que não são tratados apenas via probabilidade, são resolvidos através de listas de correspondências entre as línguas e, novamente, trata-se de um mecanismo bastante simples que não é capaz de lidar com operações mais complexas e estruturas como os padrões de cunhagem.

O modelo proposto neste trabalho, por sua vez, mostrou-se potencialmente mais eficaz, no sentido de que é capaz de oferecer regras mais sofisticadas, as quais se baseiam nas construções da língua, modeladas num Constructicon e analisadas por um *parser* construcional, o CARMA.

Assim, esta tese contribui tanto no sentido de trazer para o domínio da modelagem linguístico-computacional de construções importante debate teórico acerca do lugar dos padrões de cunhagem no conhecimento linguístico, propondo soluções computacionais (restrições *soft*) inovadoras para seu tratamento; quanto no sentido de apontar para uma aplicação real de tal modelo, qual seja um novo paradigma de hibridização em tradução por máquina.

REFERÊNCIAS

ALMEIDA, V. G. **Identificação Automática de Construções de Estrutura Argumental**: um experimento a partir da modelagem linguístico-computacional das construções Transitiva Direta Ativa, Ergativa e de Argumento Cindido. (Dissertação de Mestrado em Linguística). Universidade Federal de Juiz de Fora. Juiz de Fora, 2016.

ALONSO, K. S. B. **Construções Binominais Quantitativas e Construção de Modificação de Grau: Uma abordagem baseada no uso**. 2010**. (**Tese de Doutorado em Linguística). PPG em Linguística, Universidade Federal do Rio de Janeiro, 2010.

ARCAN, M. **Machine translation of domain-specific expressions within ontologies and documents.** (Tese de doutorado). Insight Centre for Data Analytics National University of Ireland, Galway. 2017.

BOAS, H. C. (ed.) Contrastive Studies in Construction Grammar. Amsterdam/Philadelphia: John Benjamins, 2010.a.

_____. Frame Semantics and Translation. In: ROJO, A; IBARRETXE-ANTUÑANO, I. (Eds.). **Cognitive Linguistics and Translation:** Advances in some theoretical models and applications. Berlin: Mouton de Gruyter, 2013. p. 125-158.

BICK, E. **The Parsing System Palavras** - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Famework, Aarhus: Aarhus University Press, 2000.

BRODBECK, R. C. M. S. Um monte de problemas gera uma chuva de respostas: um estudo de caso de desencontro na quantificação nominal em português. 2010. (Tese de doutorado) Faculdade de Letras, Universidade Federal de Juiz de Fora, 2010.

BYBEE, J. **Language, usage and cognition**. Cambridge: Cambridge University Press, 2010.

FILLMORE, C. J. The Case for Case Reopened In: Peter Cole (ed.), **Grammatical Relations**. Syntax and Semantics 8. New York: Academic Press, 1977, p. 59–81,

_____. Frame semantics. In: **Linguistics in the Morning Calm**, Seoul, Hanshin Publishing Co., 111-137, 1982.

_____. Frames and the semantics of understanding. In: **Quaderni di Semantica**, Vol. 6.2: 222-254, 1985.

_____. **Border Conflicts:** FrameNet Meets Construction Grammar. In: EURALEX, 13, 2008, Barcelona. *Anais*... Barcelona: Universitat Barcelona Fabra, 2008.

_____. Berkeley Construction Grammar. In: HOFFMANN, T.; TROUSDALE, G. **Oxford Handbook of Construction Grammar (**Eds.). Oxford University Press, 2013

_____; ATKINS, B. T. S. Towards a frame-based lexicon: The semantics of RISK and its neighbors. In: LEHRER, A.; KITTAY, A. (Eds.) Frames, Fields, and Contrast: **New Essays in Semantics and Lexical Organization**. Hillsdale: Lawrence Erlbaum Associates, 1992. p. 75-102.

_____. LEE-GOLDMAN, RUSSELL& RHOMIEUX, RUSSELL. The FrameNet Constructicon. In: Ivan A. Sag and Hans C. Boas (eds.), **Sign-Based Construction Grammar.** Stanford, CA: CSLI Publications, 283–99, 2012.

_____ & KAY, P. **Construction Grammar**. Ms. University of California, Berkeley, 1995.

FRIED, MIRJAM, ÖSTMAN "Construction Grammar: A Thumbnail Sketch" In: Mirjam Fried and Jan-Ola Östman (eds.), *Construction Grammar in a Cross-language Perspective*. Amsterdam: John Benjamins, 11–86, 2004 *a*.

GOOGLE TRADUTOR. Disponível em <https://translate.google.com.br/> Acesso em: 27 de fevereiro de 2018.

GOLDBERG, A. **Constructions:** A Construction Grammar Approach to Argument Structure. Chicago: The University of Chicago Press, 1995.

_____. Constructions at Work: The nature of generalization in language. Oxford: Oxford University Press, 2006.

GOLDBERG, A. Constructionist Approaches. In: HOFFMANN, T.; TROUSDALE, G. **The Oxford Handbook of Construction Grammar (**Eds.). Oxford University Press, 2013.

JACKENDOFF. a. Constructions in the Parallel Architecture. In Thomas Hoffman and Graeme Trousdale (eds.), **The Oxford Handbook of Construction Grammar,** 70-92. Oxford: Oxford University Press, 2013.

JONES, K. S. **Computational Linguistics:** What About the Linguistics? Association for Computational Linguistics. Vol. 3, n. 33, 2007.

KAMRAN, A. *Hybrid Machine Translation*. (Projeto de Doutorado) Charles University in Prague. Faculty of Mathematics and Physics. Institute of Formal and Applied Linguistics, 2013.

KAY, P. Argument Structure Constructions and the Argument-adjunct Distinction. In: Mirjam Fried and Hans C. Boas (eds.), **Grammatical Constructions:** Back to the Roots. Amsterdam: John Benjamins, 2005, p.71–100.

_____. The Limits of (Construction) Grammar. In: Hoffmann, T.; Trousdale, G. (eds). **The Oxford Handbook of Construction Grammar**. Online publications, 2013.

KOHEN, P. Introduction. In: **Statistical Machine Translation.** New York: Cambridge University Press, 2010. p.3-31.

_____. Words, Sentences, Corpora. In: **Statistical Machine Translation.** New York: Cambridge University Press, 2010. p.33-62.

LANGACKER, R. W. **Foundations of cognitive grammar,** vol. 1: Theoretical prerequisites. Stanford, CA: Stanford University Press, 1987.

LAGE, L. M. **Frames e construções:** A implementação do constructicon na FrameNet Brasil. Dissertação (Mestrado em Linguística). Universidade Federal de Juiz de Fora, Juiz de Fora, 2013.

LAGE, L. M. Modelagem Linguístico-Computacional das Relações entre Construções e Frames no Constructicon da Framenet Brasil. (Doutorado em Linguística). Universidade Federal de Juiz de Fora, Juiz de Fora, 2018.

LAKOFF, G. The Contemporary Theory of Metaphor. In: ORTONY, A. **Metaphor and Thought**. Cambridge: Cambridge University Press. 1992.

_____. & JOHNSON, M. **Metaphors We Live By**. Chicago: Chicago University Press. 1980

_____. & JOHNSON, M. **Philosophy in the Flesh**: The embodied mind and its challenge to the Western thought. 1a ed., New York, Cambridge University Press, 1999.

LAVIOLA, A. B. **Frames e Construções em Contraste:** uma análise comparativa Português – Inglês no tangente à implementação de Constructicons. Dissertação de Mestrado em Linguística. Universidade Federal de Juiz de Fora. Juiz de Fora, 2015

MANNING, C., SCHUTZE, H. **Foundations of Statistical Natural Language Processing.** MIT Press. Cambridge, MA: May, 1999.

MATOS, E. E. S. **LUDI:** Um framework para desambiguação lexical com base no enriquecimento da Semântica de Frames. 2014. (Tese de Doutorado em Linguística) – Faculdade de Letras, Universidade Federal de Juiz de Fora.

_____., TORRENT, T. T., ALMEIDA, V. G., SILVA, A. B. L., LAGE, L. M., MARÇÃO, N. D., TAVARES, T. S. Constructional Analysis Using Constrained Spreading Activation in a FrameNet- Based Structured Connectionist Model In: **The AAAI 2017 Spring Symposium on Computational Construction Grammar and Natural Language Understanding Technical Report** SS-17-02. Palo Alto, CA: AAAI Publications, v.17, 2017, p.222-229.

NAVIGLI, R. **ACM Computing Surveys***,* Vol. 41, No. 2, Article 10, February 2009.

PETRUCK, M. Frame Semantics In: Jef Verschueren, Jan-Ola Östman, Jan Blommaert, and Chris Bulcaen (eds.), **Handbook of Pragmatics**. Amsterdam: John Benjamins, 1–11, 1996.

PUSTEJOVSKY, J. & STUBBS, A. **Natural Language Annotation forMachine Learning:** A guide to corpus-building for applications. New York: O'ReillyMedia, 2012.

RUPPENHOFER, J.; ELLSWORTH, M.; PETRUCK, M.; JOHNSON, C.; SCHEFFCZYK, J. **FrameNet II:** Extended Theory and Practice. Versão 14 set. 2010. Disponível em: http://framenet.icsi.berkeley.edu/ Acesso em 20 out. 2010.

SAWAF, H. et al. **Patent Application Publication**. 2010. Disponível em: <https://docs.google.com/viewer?url=patentimages.storage.googleapis.com/pdfs/US20100179803.pdf> Acesso em: 02 de março de 2015.

SILVA, J. R. **Motivações semântico-cognitivas e discursivo-pragmáticas nos processos de intensificação**. Tese (Doutorado em Letras) – Programa de Pós Graduação em Estudos da Linguagem (PPGEL), Universidade Federal do Rio Grande do Norte (UFRN), 2008.

TAVARES, T. S. **Construção Binominal de Quantificação Indefinida do Português** – Uma Abordagem Construcionista. (Dissertação de Mestrado). Faculdade de Letras, Universidade Federal de Juiz de Fora, 2014.

VERVECKKEN, K. Towards a constructional account of high and low frequency binominal quantifiers in Spanish, 2012. **Cognitive Linguistics**, p. 421 – 478.