# A Modelagem Computacional do Domínio dos Esportes na FrameNet Brasil

**Alexandre Diniz da Costa[1], Tiago Timponi Torrent[1]**

[1] FrameNet Brasil – Programa de Pós-Graduação em Linguística
Universidade Federal de Juiz de Fora (UFJF)
Rua José Lourenço Kelmer, s/nº, Campus Universitário
36036-900 – Juiz de Fora – Minas Gerais – Brasil

{alexandre.costa,tiago.torrent}@ufjf.edu.br

***Abstract.*** *This paper aims to describe the linguistic-computational modeling of Frames [Fillmore 1982] and Qualia [Pustejovsky 1995] for the Sports domain, carried out in the FrameNet Brasil database. The modeling relied on a domain-specific corpus. By adding Qualia roles to the database, this work promoted the densification of FrameNet Brasil, focusing on its use in tools that deal with Natural Language Processing, such as parsers and machine translators.*

***Resumo.*** *Este trabalho busca descrever a modelagem linguístico-computacional de frames [Fillmore 1982] e relações Qualia [Pustejovsky, 1995], realizada na base de dados da FrameNet Brasil para o domínio dos Esportes. A modelagem se sucedeu a partir de uma pesquisa em corpus. Ao adicionar papéis Qualia à base de dados, este trabalho promoveu o adensamento da FrameNet Brasil, visando a sua utilização em ferramentas que lidam com o Processamento de Língua Natural, tais como parsers e tradutores por máquina.*

## 1. Introdução

A interação entre os seres humanos e máquinas têm se tornado cada vez mais frequente ao longo dos anos. Para que tal interação apresente êxito, se fazem necessárias melhorias nas ferramentas computacionais que envolvem a linguagem. Tarefas comuns no Processamento de Língua Natural incluem a Tradução por Máquina, o Reconhecimento e o Processamento de Fala, Sistemas de Pergunta e Resposta, Parsing, entre outras. No intuito de contribuir para essa melhoria das tarefas de Compreensão de Língua Natual, este trabalho busca descrever a modelagem linguístico-computacional de frames [Fillmore 1982] e relações Qualia [Pustejovsky, 1995], realizada na base de dados da FrameNet Brasil para o domínio dos Esportes. A modelagem se sucedeu a partir de uma pesquisa em corpus constituído no âmbito deste trabalho através da compilação de notícias esportivas, manuais de esportes, além de sites de associações brasileiras de esportes e sites oficiais dos Jogos Olímpicos Rio 2016. Ao adicionar papéis Qualia à base de dados, este trabalho promoveu o adensamento da FrameNet Brasil, visando a sua utilização em ferramentas que lidem com a Compreensão de Língua Natural.

## 2. A FrameNet Brasil

A FrameNet Brasil é uma base de descrição lexicográfica sustentada por dados em corpora para o Português do Brasil fundada a partir dos desenvolvimentos da FrameNet de Berkeley para a língua inglesa. O conceito de frame é essencial para esse projeto, sendo definido como "qualquer sistema de conceitos relacionados de tal modo que, para

entender qualquer um deles, é preciso entender toda a estrutura na qual se enquadram" [Fillmore 1982: 111].

Como um exemplo de frame modelado na FrameNet Brasil referente ao domínio dos esportes, podemos citar o frame Infrações_diretas. Ilustrado na Figura 1, constatamos o nome do frame, seguido de sua definição por extenso, seus Elementos de Frame (EFs) Nucleares e Não-nucleares, as relações que o frame possui com outros frames e as Unidades Lexicais (ULs) que o evocam. O frame em questão, Infrações_diretas, designa "uma infração contra um adversário, considerada imprudente, temerária ou com uso de força excessiva. Essas infrações podem gerar como penalidade o tiro livre direto no futebol, conforme o lugar onde ocorreram". Os Elementos de Frame (EFs) são papéis semânticos definidos especificamente no frame, sendo seus constituintes básicos. Eles oferecem informações semânticas à estrutura da sentença. Os EFs nucleares são definidores e essenciais ao sentido do frame, enquanto que os EFs não-nucleares caracterizam mais de um frame e normalmente apresentam condições comuns a vários frames com ideias que expressam tempo, lugar e maneira, entre outras. Os EFs nucleares do frame de Infrações_diretas são o Adversário e o Infrator e os não-nucleares são Lugar, Tempo e Tipo. As Unidades Lexicais (ULs), que são pareamentos entre lemas e frames, possuem cada uma um significado específico e evocam o frame em que estão modeladas. Algumas ULs que evocam Infrações_diretas são *carrinho.n* e *derrubar.v*. As definições dessas ULs no domínio específico dos esportes são "no futebol, é a infração em que um jogador, com a finalidade de retirar a bola do adversário, atira-se ao chão e desliza com as pernas para frente" e "no futebol, é a infração em que um jogador derruba algum de seus adversários", respectivamente.


**Figura 1. Frame Infrações_diretas na FrameNet Brasil.**

Essa representação da língua em termos de frames se faz útil na estruturação linguística para fins computacionais e de tarefas de Processamento de Língua Natural (PLN). A proposta de adensamento da base de dados da FrameNet Brasil surge com o

propósito de viabilizar e otimizar certos tipos de tarefas que envolvem a língua e a máquina. Entretanto, para certas tarefas de processamento como a Tradução por Máquina, Compreensão de Língua Natural e Desambiguação, se faz necessária a existência de relações mais locais e específicas entre as ULs, visto que os frames são mais gerais. Passemos então às relações criadas e modeladas entre ULs na FrameNet Brasil através dos papéis Qualia.

## 3. Papéis Qualia

A Teoria do Léxico Gerativo (TLG), proposta por Pustejovsky (1995), surge como uma abordagem que lida com a semântica das palavras, como elas se combinam, o que denotam, além de mecanismos peculiares como a polissemia e a coerção de tipos. O avanço da teoria se deve a uma insatisfação de muitos linguistas teóricos e computacionais com a caracterização do léxico como um conjunto fechado e estático de traços sintáticos, morfológicos e semânticos. Os papéis ou relações Qualia são aspectos essenciais do significado das palavras. O autor busca lidar com o aspecto criativo de combinações lexicais e a representação semântica a partir da descrição de uma série de componentes primitivos do significado. Essa abordagem sofre fortes influências de fatores situacionais e contextuais.

O quale constitutivo propõe a relação entre um objeto e suas partes ou materiais constituintes. O quale formal é o que distingue um objeto dentro de um domínio maior. Ele inclui características como a orientação, forma, dimensões, cor, posição, tamanho etc. Cada atributo pode ser preenchido por um valor. O quale télico se relaciona à função ou propósito da entidade, sendo ela um objeto, uma pessoa ou um lugar. E por fim, o quale agentivo se coloca nos fatores envolvidos na origem ou no vir a existir do objeto. Características inclusas nessa relação são o criador, o artefato, o tipo natural e uma corrente causal.

Inicialmente, no domínio dos esportes dentro da FrameNet Brasil, foram modeladas as relações Télico_de e Constitutivo_de, dado que essas relações contribuíam para tarefas de desambiguação e extração de representações semânticas. Em um momento posterior, prevê-se a modelagem das outras duas relações qualia, quais sejam Agentivo_de e Formal_de. Vejamos um exemplo da modelagem da relação Qualia Télico_de na Figura 2.


**Figura 2. Frame Atletas_por_esporte e a modelagem da relação Télico_de (tlc_).**

Na Figura 2, temos ilustrado o frame de Atletas_por_esporte com sua definição, seus EFs nucleares e as ULs que evocam este frame. Ao lado esquerdo da imagem, temos a representação do frame em uma barra vertical com as ULs contidas nesse frame tais como *jogador de futebol.n*. Na UL *jogador de futebol.n* temos modelada a relação Télico_de (tlc_) apontando para algumas ULs dos frames de Jogadas e de Infrações, visto que essa relação é estabelecida para mostrar o propósito ou a função de uma entidade, sendo ela o jogador de futebol nesse exemplo. Portanto, o ato de cometer *falha.n*, *falta.n*, *infração.n*, *carrinho.n*, *derrubar.v* e *entrada.n* seriam télicos de jogador, ou seja, sua função ou propósito em uma partida de futebol. A ideia original criada por Pustejovsky (1995) atribui o papel télico às funções ou propósitos da entidade. No âmbito desta pesquisa, pretende-se ampliar o conceito desse papel, levando-o a cobrir todos os movimentos e ações realizadas especificamente pela entidade, sendo ela um objeto, uma pessoa ou um lugar. Partindo desse princípio, a relação télico_de estabelecida para a UL *jogador de futebol.n* conecta essa UL não apenas às jogadas pertinentes a ela, mas também às infrações que são ações cometidas por esse atleta em especial.

Como se pode ainda observar na coluna vertical lateral à esquerda, na modelagem das ULs que evocam o frame de Atletas_por_posição, optou-se pela criação de ULs polilexêmicas tais como *jogador de futebol.n*, *jogador de badminton.n*, *jogador de basquete.n*, entre outras, na tentativa de cobrir mais dados peculiares a cada esporte em si, facilitando o reconhecimento por máquina das diferenças entre cada especificação dos tipos de jogador. A UL *jogador.n*, mais genérica, foi modelada no frame mais genérico de Atletas, a partir do qual foi herdado Atletas_por_esporte. Essa UL *jogador.n* apresenta a relação télico_de mapeada a todos os diversos tipos de ações, movimentos ou jogadas realizadas por todas especificações de cada esporte que utiliza a UL. Na seção 4, detalharemos como essa relação Qualia local entre ULs pode contribuir para o adensamento da base da FrameNet Brasil e para tarefas de PLN.

## 4. A Modelagem Computacional do Domínio dos Esportes na FrameNet Brasil

Os Jogos Olímpicos são eventos em que pessoas de países e culturas distintos se encontram em um mesmo local. Essa situação viabiliza o contato e a troca em um ambiente multilíngue que não envolve apenas esportes, mas também comidas, bebidas, acomodação, línguas e culturas em contato. A modelagem computacional do domínio dos Esportes na FrameNet Brasil surge como uma extensão ao domínio do Turismo, visto que o evento base que incitou essa pesquisa, os Jogos Olímpicos Rio 2016, lidavam com turistas em busca de informações relacionadas ao Turismo em geral e aos Esportes Olímpicos. Consequentemente, essa área fornece um domínio linguístico específico parcialmente controlado em que tecnologias de comunicação vem sendo desenvolvidas com o intuito de minimizar as diversas diferenças linguísticas existentes.

Em 2014, a FrameNet Brasil (http://www.framenetbr.ufjf.br) desenvolveu um dicionário trilíngue baseado em frames para a Copa do Mundo realizada no Brasil. Já em 2016, começou a desenvolver o m.knob (Multilingual Knowledge Base), sendo uma aplicação computacional que funciona como guia turístico multilíngue e intérprete pessoal, com um algoritmo de tradução enriquecido semanticamente com frames e papéis qualia. Uma das funções básicas da aplicação é um sistema de recomendação baseado em técnicas de geolocalização que pode ser utilizado por turistas para se localizarem e sugere a eles locais específicos conforme suas necessidades e sua interação com o aplicativo.

Uma outra função do aplicativo ainda em desenvolvimento relaciona-se a um tradutor semanticamente melhorado baseado em frames, papéis qualia e ontologias, abarcando os idiomas português, inglês e espanhol. Pretende-se, com esse tradutor que fará uso de redes neurais, gerar melhores equivalentes de tradução a partir de uma base de dados semanticamente enriquecida, tornando-se diferente de algoritmos de tradução que utilizam apenas estatística e sintaxe.

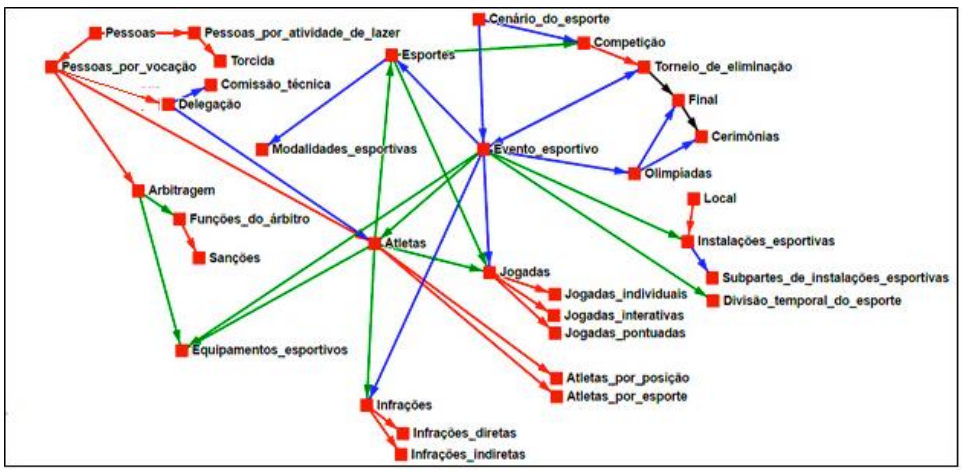


**Figura 3. Rede de Frames que compõem o Cenário_do_Esporte.**

Para a modelagem do domínio, inicialmente constituiu-se um corpus de domínio específico dos esportes, contendo textos de manuais esportivos, websites de associações brasileiras de esportes, notícias esportivas, além de websites oficiais dos Jogos Olímpicos Rio 2016. A partir do corpus e de um agrupamento prévio das diversas modalidades esportivas, suas peculiaridades e características em comum, realizado pelos linguistas da

FrameNet Brasil, foi proposta a modelagem de uma rede de frames para o Cenário_do_Esporte, observada na Figura 3.

A base de dados, no domínio do Turismo, incluindo frames genéricos com relações aos frames do mesmo domínio, conta com aproximadamente 52 frames e 425 unidades lexicais previamente modeladas conforme se pode conferir em Torrent et. al. (2014b), Gamonal (2013), Gomes (2014) e Souza (2014). No âmbito dessa pesquisa, o domínio específico dos Esportes foi modelado até o momento através de 32 frames e 640 ULs. Com exceção dos frames de Pessoas, Pessoas_por_vocação, Local e Competição, todos os demais frames foram criados no âmbito deste trabalho, a partir da observação do comportamento linguístico dos itens vocabulares do domínio dos Esportes nos corpora.

O processo de criação dos frames, EFs e ULs no domínio dos Esportes partiu de uma pesquisa em corpora através da ferramenta SketchEngine. Após a compilação dos dados, analisaram-se as ocorrências de possíveis candidatas a ULs. Através dos padrões de valência apresentados pelas ULs e a frequência de ocorrência em corpus específico, criaram-se os frames e estabeleceram-se frames filhos com certas especificidades. Como exemplo desse processo de criação, podemos mencionar os frames de Jogadas, Jogadas_individuais, Jogadas_interativas e Jogadas_pontuadas. Ao observar o comportamento sintático-semântico das jogadas e movimentos relacionados aos atletas, o frame Jogadas ficou sendo mais genérico e possuindo os três frames filhos ligados a ele por relação de herança. No frame de Jogadas_individuais, o perfilamento se deu na movimentação individual do atleta. Já no frame Jogadas_interativas, a perspectiva se coloca na interação de um atleta com outro na realização do movimento. Já o último, Jogadas_pontuadas, a pontuação gerada pelo movimento incitou a criação desse frame específico. O mesmo processo se deu ao longo da criação dos demais frames dos Esportes, seus EFs e as ULs que os evocam. Passemos agora à seção 5, na qual discutiremos as implicações de uma modelagem computacional enriquecida semanticamente para a realização de tarefas de Compreensão de Língua Natural.

## 5. Implicações de uma modelagem computacional enriquecida no uso de ferramentas que lidam com a linguagem

No PLN, a tradução por máquina apresenta-se como uma tarefa que busca cada vez mais melhorias na geração de melhores equivalentes de tradução. Dentro das tarefas de Processamento e Compreensão de Língua Natural, a tradução por máquina é uma das áreas que se constitui como uma das mais desenvolvidas no momento.

Os estudos tradutórios que envolvem algoritmos de tradução passaram por diversas etapas. Inicialmente, analisava-se palavra por palavra, passou-se aos agrupamentos de sintagmas, considerando posteriormente a colocação sintática dos elementos e a transposição entre eles nas línguas traduzidas. Ainda mais à frente, começa-se a trabalhar com ocorrências em corpora e a tratar de n-grams. Nessa abordagem, grandes quantidades de textos traduzidos são analisados observando estatisticamente os padrões de ocorrência de palavras e suas fronteiras, que elementos as acompanham. Individualmente, tais palavras foram tratadas como unigrams. A colocação estatística de duas palavras que ocorrem com frequência juntas são chamadas bigrams, e assim por diante [Koehn 2010].

Atualmente, os sistemas e algoritmos de tradução têm sido melhorados substancialmente com a utilização de redes neurais. As redes neurais tentam emular o

funcionamento do cérebro humano e levam em consideração sua característica de processamento em que as informações são tomadas de forma descentralizada. As redes neurais trabalham com a capacidade de compreensão de padrões e contexto. Uma grande quantidade de textos é submetida à inteligência que os processa, e se retroalimenta dos mesmos, gerando melhores equivalentes de tradução [Bahdanau et al. 2014]. Isso tem ocorrido de forma satisfatória. Entretanto, para certas instâncias de domínios específicos como o Turismo e os Esportes, ainda existem algumas inconsistências que buscamos solucionar através de um algoritmo que apresente relações semânticas adensadas entre frames, EFs e ULs, a fim de que uma rede densa com relações qualia e ligada a dados abertos de ontologias possa melhorar ainda mais a tarefa de geração de equivalentes de tradução.

Demonstrada a modelagem dos frames para o domínio específico dos Esportes e os papéis qualia que adensam e estreitam a relação entre as ULs, analisemos a Figura 4 que propõe uma tradução de uma sentença do português para o inglês no domínio dos esportes em um algoritmo conhecido de tradução.


**Figura 4. Tradução de uma sentença do domínio dos esportes do Português para o Inglês em uma ferramenta online.**

Na Figura 4, em "O jogador deu um carrinho no meia Ederson", temos uma UL que apresenta ambiguidade, a palavra *carrinho.n*. A mesma forma lexical carrinho.n poderia remeter a ULs diferentes, como, por exemplo, "um carrinho de bebê", "um brinquedo em forma miniatura de um carro", ou ainda "um objeto utilizado no transporte de mercadorias em um supermercado". Para cada sentido diverso da UL *carrinho.n*, um frame diferente seria evocado. Percebemos que, no domínio específico dos esportes, o frame Infrações é evocado pela UL *carrinho.n*, sendo atribuído ao jogador o EF Infrator e ao meia o EF Adversário.

É proposta uma sentença traduzida em inglês: "The player gave a stroller to midfielder Ederson". Temos como sugestão de tradução para *carrinho.n* a UL *stroller.n* em inglês. A palavra s*troller* designa um carrinho de bebê e não uma infração cometida no futebol. A expressão que seria o equivalente mais adequado dentro do domínio dos esportes seria *slide tackle*. Uma tradução da sentença apropriada seria, portanto, "the player made a sliding tackle to midfielder Ederson".

Considerando-se a rede proposta de modelagem dos Esportes, através da modelagem da relação local Télico_de entre a UL *jogador.n* do frame Atletas ou da UL *jogador de futebol.n* do frame Atletas_por_Esporte e da UL *carrinho.n* do frame Infrações_diretas, é estabelecida uma relação in loco entre as ULs contribuindo para uma melhor representação semântica de nomes comuns de entidade como *carrinho.n*. A partir dessa análise baseada em uma rede de frames, os frames que evocariam os outros sentidos possíveis de carrinho estariam mais distantes na rede do que a relação estabelecida entre as ULs dos frames de Atletas, Atletas_por_Esporte e Atletas_por_posição com os frames relacionados a Jogadas e Infrações. Essa relação local modelada com os papéis qualia traria ao algoritmo de tradução informações semânticas que outros sistemas baseados apenas em estatística ou sintaxe não conseguiram ainda fornecer. Portanto, com este

trabalho, oferecemos uma alternativa linguístico-computacional para a geração de melhores equivalentes de tradução conforme o frame, as relações semânticas e o contexto de domínio específico em que estão inseridos.

## 6. Considerações Finais

Este trabalho apresentou uma proposta de enriquecimento da base de dados lexicais da FrameNet Brasil com a modelagem de relações Qualia entre ULs. Como aplicação sugerida, discute-se de que maneira a tradução por máquina poderia se beneficiar da referida base de dados, em especial para melhorar escolhas lexicais em léxicos de domínio específico, tais como o Turismo e os Esportes.

## Referências


Bahdanau, D., Cho, K., & Bengio, Y. (2014). "Neural machine translation by jointly learning to align and translate". *arXiv preprint arXiv:1409.0473*.

Fillmore, C. J. (1982). "Frame semantics" In: Linguistics in the Morning Calm. Seoul, South Korea, Hanshin Publishing Co., p. 111-137.

Gamonal, M. A. (2013). "COPA 2014 FrameNet Brasil: Diretrizes para a Constituição de um Dicionário Eletrônico Trilíngue a partir da Análise de Frames da Experiência Turística." Dissertação de Mestrado em Linguística. Universidade Federal de Juiz de Fora. Juiz de Fora.

Gomes, D. S. (2014). "Frames do Turismo Esportivo no Dicionário Copa 2014_FrameNet Brasil." Dissertação de Mestrado em Linguística. Universidade Federal de Juiz de Fora. Juiz de Fora.

Koehn, P. (2010). "Statistical Machine Translation", Cambridge, Cambridge University Press.

Pustejovsky, J. (1995). "The Generative Lexicon", Cambridge, USA, MIT Press.

Souza, B. C. P. (2014). "Frames de turismo como negócio no Dicionário Copa 2014_FrameNet Brasil." Dissertação de Mestrado em Linguística. Universidade Federal de Juiz de Fora. Juiz de Fora.

Torrent, T.T; Salomão, M. M.; Campos, F. A.; Braga, R. M; Matos, E. E.; Gamonal, M. A.; Gonçalves, J.; Souza, B. C.; Gomes, D. S. & Peron-Correa, S. R. (2014b). "Copa 2014 FrameNet Brasil". Proceedings of COLING 2014, p. 10 -14.