FOLHA INTERNACIONAL: ANOTAÇÃO AMR EM UM CORPUS DE NOTÍCIAS EM INGLÊS

ANA BEATRIZ JOAQUINA DA SILVA[1], CLÁUDIA DIAS DE BARROS[2].

[1] Graduando em Letras – Português/Inglês, Bolsista PIBIFSP, IFSP, Câmpus Sertãozinho, Ana.b.j.s@hotmail.com
[2] Docente em Letras – Português/Inglês, IFSP, Câmpus Sertãozinho, Claudiabarros@ifsp.edu.br

Área de conhecimento (Tabela CNPq): 8.01.01.00-3 - Teoria e Análise Linguística

Apresentado no 10° Congresso de Inovação, Ciência e Tecnologia do IFSP – 2019
27 e 28 de novembro de 2019 – Sorocaba – SP, Brasil

RESUMO: É de suma importância um corpus anotado para algumas aplicações de Processamento de Língua Natural (PLN), como Sumarização Automática, Tradução Automática, Análise de Sentimentos, entre outras. Essa anotação pode ser em várias camadas linguísticas, como a anotação semântica. Como exemplo de anotação semântica, pode-se citar a AMR (Abstract Meaning Representation), a qual possui uma estrutura simples, que estabelece relações/conexões entre nós/conceitos e foca na estrutura predicado-argumento presente no PropBank. Portanto, o presente projeto tem como objetivo realizar a anotação de sentenças presentes em um corpus em inglês, formado por notícias da Folha Internacional, com a finalidade de construir as camadas de anotação, observando e comparando as possíveis diferenças entre a anotação de um corpus jornalístico e uma anotação já feita de um texto literário (Pequeno Príncipe).

PALAVRAS-CHAVE: Abstract Meaning Representation (AMR); Processamento de Língua Natural (PLN); Anotação Semântica.

FOLHA INTERNACIONAL PAPER: AMR ANNOTATION IN A ENGLISH NEWS CORPUS

ABSTRACT: A corpus annotation is extremely important for some applications of Natural Language Processing (NLP), such as Automatic Summarization, Machine Translation, Sentiment Analysis, etc. The annotation can be a semantic annotation, as AMR (Abstract Meaning Representation), which has a simple structure and establishes relationship/connections between nodes/concepts and focuses on the predicate-argument structure present in PropBank. Thus, the present project aims to make AMR annotation of sentences in a corpus formed by news from Folha Internacional paper, in order to build the annotation layers and to observe and compare the possible differences between a news corpus and a literary corpus (The Little Prince).

KEYWORDS: Abstract Meaning Representation (AMR); Natural Language Processing (NLP); Semantic Annotation.

INTRODUÇÃO

As tarefas de Processamento de Língua Natural (PLN) como Sumarização Automática, Tradução Automática, Análise de Sentimentos, entre outras, necessitam utilizar, muitas vezes, como um recurso importante, um corpus anotado com várias camadas, dentre as quais pode-se destacar a anotação semântica.

Um tipo de representação semântica que pode ser utilizada em um corpus é a AMR (Abstract Meaning Representation) (BANARESKU et al. 2013), que pode ser realizada como um grafo, com nós (conceitos) etiquetados e arestas (relações) entre eles em uma sentença. Os conceitos AMR podem ser palavras, framesets (conjunto de sentidos dos verbos) do PropBAnk (KINGSBURY e PALMER, 2002) (que são verbos ligados à lista de possíveis argumentos e seus papéis semânticos) ou palavras-chave especiais, como “date-entity”, “distance-quatity”, entre outras. A Figura 1 ilustra o grafo das sentenças

"*The girl made adjustment to the machine", "The girl adjusted the machine", "The machine was adjusted by the girl",* mostrando que, apesar das sentenças serem dispostas de formas diferentes, os seus sentidos se mantêm.

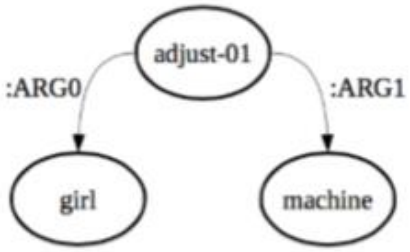

Figura 1: Representação por grafo (extraído de Anchieta e Pardo (2018))

O presente projeto tem como objetivo geral realizar a anotação da representação semântica AMR de sentenças presentes na parte em inglês de um corpus paralelo, formado por notícias da Folha Internacional, com a finalidade de construir as camadas de anotações do presente corpus, para que seja possível observar e comparar as possíveis diferenças entre a anotação de um corpus jornalístico e uma anotação já feita de um texto literário (Pequeno Príncipe). Sendo assim, seus objetivos específicos são:

- Realização da anotação AMR em sentenças em inglês do corpus formado por notícias presentes na Folha Internacional;
- Observação da anotação AMR que fora realizada em um corpus em inglês do livro Pequeno Príncipe;
- Comparação da anotação realizada no corpus literário e no jornalístico, identificando suas possíveis diferenças de anotação.

MATERIAL E MÉTODOS

Os processos e etapas metodológicas que foram definidos para que seja realizado este projeto se encontram nas seguintes atividades:

1) Construção de um corpus de notícia em inglês extraído da Folha Internacional Online;
2) Anotação semântica utilizando a ferramenta AMR Editor (HERMJAKOB, 2013), como mostra a Figura 2;
3) Comparação das sentenças anotadas do corpus jornalístico em inglês com o corpus já anotado do texto literário Pequeno Príncipe.

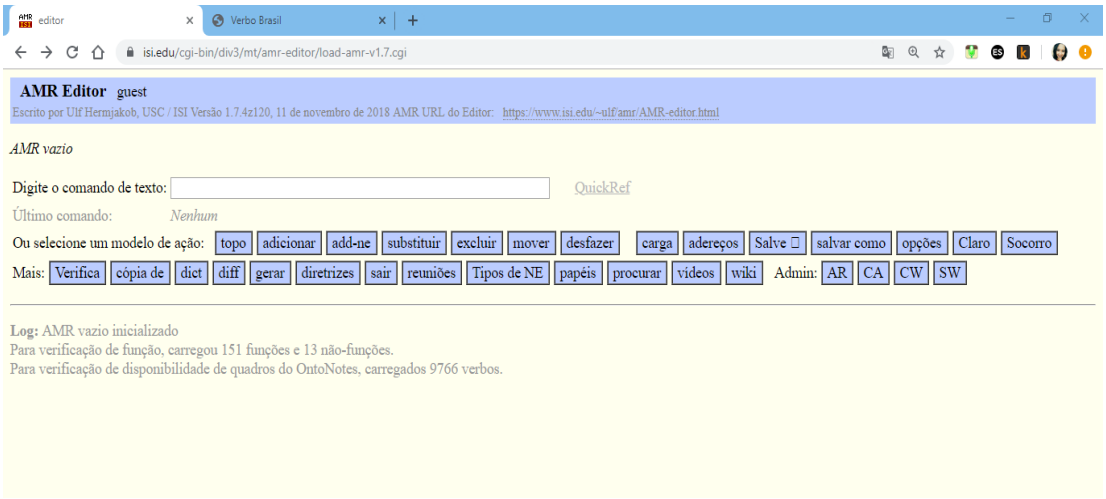

FIGURA 2. Tela da ferramenta AMR Editor

RESULTADOS E DISCUSSÃO

O presente projeto encontra-se em andamento, sendo que, até o presente momento foi realizada a definição do escopo do problema, leitura de artigos sobre o tipo de anotação AMR,

anotação de sentenças em português para familiarização com a ferramenta e construção do corpus de notícias em inglês que fora retirado da Folha Internacional Online, sendo formado por 30 títulos de manchetes. A próxima etapa do projeto será a anotação AMR do corpus, para que seja possível fazer a comparação com a anotação já existente do texto literário.

CONCLUSÕES

Espera-se, com a realização deste trabalho, promover o conhecimento sobre Processamento de Língua Natural (PLN), mais especificamente a área de anotação de corpus, a qual se apresenta como uma ferramenta muito importante para auxiliar em tarefas posteriores, como Tradução Automática, Sumarização Automática, Análise de Sentimentos, entre outras. Além da construção de um corpus em inglês anotado com a representação semântica AMR, espera-se, com este trabalho, contribuir para a área de análise linguística, devido às análises comparativas das anotações de sentenças em inglês em gêneros textuais diferentes, como o jornalístico e o literário.

AGRADECIMENTOS

Agradecemos ao IFSP pelo financiamento para o Programa Institucional de Bolsas de Iniciação Científica e Tecnológica.

REFERÊNCIAS

Anchiêta, R. T.; Pardo, T. A. S. Towards AMR-BR: a SemBank for Brazilian Portuguese Language. In: Proceeding of the 11th Internacional Conference on Language Resources And Evaluation (LREC, 2018). Miyazaki, Japan, 2018.

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Grif-Fitt, K. Hermjakob, U., Knight, K., Palmer, M., and Schneider, N. Abstract meaning representation for sembanking.In Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Disc ourse, pages 178–186, 2013.

Hermjakob, U. Amr editor: A tool to build abstract meaning representations, 2013.

Kingsbury, P. and Palmer, M. From treebank to propbank. Ind Proceedings of the 3rd International Con-ference on Language Resources and Evaluation, pages 1989–1993., 2002.