

Relatório de actividades desenvolvidas no âmbito da Linguateca através do contrato de prestação de serviços com a FCCN

Julho-Dezembro de 2009

Cláudia Freitas

Durante os seis meses em que fui contratada da FCCN, a um regime de 30%, minhas actividades concentraram-se principalmente na anotação de relações semânticas entre entidades mencionadas (anotação de relações semânticas em todos os documentos da Coleção Dourada do Segundo HAREM) e na análise de relações semânticas extraídas pelo PAPEL (<http://www.linguateca.pt/PAPEL/papel.html>). A seguir detalho cada uma das actividades, tomando por base as actividades propostas no meu contrato:

1. Preparação do Terceiro HAREM

a) Revisão das categorias e opções lingüísticas do Segundo HAREM para uma eventual proposta para o Terceiro HAREM.

Como não haverá Terceiro HAREM, nada foi feito aqui.

b) Ampliação da marcação das relações entre EM, no sentido de vir a cobrir completamente a CD do HAREM.

No Segundo HAREM foi proposta uma tarefa piloto, o ReReEM, com o objectivo avaliar sistemas capazes de identificar relações semânticas entre entidades mencionadas (EM). No entanto, enquanto tarefa piloto, apenas 12 textos foram alvo da análise dos sistemas. Embora a anotação desses textos tenha sido proveitosa para os sistemas participantes e para o próprio desenvolvimento da tarefa, consideramos que a extensão da anotação seria de grande valia não apenas para o refinamento e validação das relações inicialmente propostas, mas principalmente em termos da qualidade do recurso disponibilizado pela Linguateca – um conjunto de textos ricamente anotado com diversos tipos de relação semântica é de grande utilidade para sistemas que realizam extração de informação e outros tipos de tarefas que envolvem uma compreensão mais profunda do texto. Vale notar a inexistência de recursos com tamanha informação semântica para a língua portuguesa.

Os 129 documentos da CD do Segundo HAREM foram anotados com as relações semânticas do ReReEM, totalizando a anotação de relações semânticas em mais de 2000 EM.

Como mencionado, um dos objetivos da anotação era a validação das relações inicialmente propostas. Com a análise de mais textos, foi possível generalizar algumas relações, refinar e ainda criar outras que nos pareceram relevantes.

Durante o processo de anotação, não foram poucas as dúvidas e discussões. Todo esse processo de anotação e discussão está documentado em <http://sites.google.com/site/anotacaodorerelem/>

c) Eventual co-supervisão de uma bolsa de pesquisa na PUC, nível de graduação, orientada pela prof. Violeta Quental, relacionada com o HAREM e o ReReEM, se essa bolsa for aprovada.

Com a aprovação da bolsa, o trabalho da bolsista consistiu, em um primeiro momento, na familiarização com o HAREM e o ReReEM (leituras e familiarização com o material das Coleções Douradas e com o programa Etiket(h)arem). Em uma segunda fase do trabalho, ainda visando uma maior familiarização com as relações semânticas, foi feita uma comparação inicial entre as relações do ReReEM e as relações semânticas presentes na WordNet.PT. Tal comparação teve como objectivo verificar em que medida as relações entre EM poderiam ser encampadas por relações semânticas mais gerais, tradicionalmente consideradas em recursos lexicais usados em tarefas que envolvem o processamento computacional da língua portuguesa.

2. Avaliação e uso dos corpos da Linguateca em ambiente de ensino

a) Em coordenação com as aulas de sintaxe na PUC-Rio, criação de exercícios e sua resolução com base na Floresta e no AC/DC e melhoria/compilação da lista de "Perguntas já respondidas sobre a utilização dos recursos da Linguateca"

Foram trabalhadas questões envolvendo a coordenação de elementos no interior da frase – coordenação tanto no nível oracional quanto no nível do sintagma. Foi possível ainda apresentar o uso de corpos – e, especificamente, recursos e ferramentas que lidam com corpos anotados (como o AC/DC e a Floresta) – como auxiliares do professor de língua portuguesa na medida em que possibilitam a apresentação de fenómenos da língua em ambiente natural e facilitam a busca – por parte do professor – por esses mesmos fenómenos tendo em vista a elaboração de exercícios.

3. Escrita de artigos e apresentações quando tal se justificar

- Colaboração na escrita do artigo "Detection of relations between named entities: report of a shared task" (Freitas et al., 2009), apresentado no NAACL HLT Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009).
- Colaboração na preparação do material "Floresta Sintá(c)tica: apresentação e exercícios" para apresentação na Escola de Verão Belinda Maia (EdV 2009).
- Colaboração na escrita e apresentação do artigo "O papel das relações semânticas em português: comparando o TeP, o MWN.PT e o o PAPEL" (Santos et al., 2009) apresentado no XXV Encontro Nacional da Associação Portuguesa de Linguística.
- Colaboração na escrita do resumo "Second HAREM: advancing the state of the art of named entity recognition in Portuguese" para a conferência Language Resources and Evaluation (LREC 2010)

- Elaboração e apresentação da palestra “Linguística Computacional, corpus, pesquisa e ensino de língua portuguesa” para o departamento de Letras da PUC-Rio.
- Elaboração e apresentação da palestra “Apresentação da Linguatca com ênfase nos recursos” para alunos de pós-graduação do departamento de Letras da PUC-Rio.
- Elaboração e apresentação da palestra “Floresta Sintática” para alunos de pós-graduação do departamento de Letras da PUC-Rio.

Adicionalmente:

4. Análise e validação de relações semânticas do PAPEL

As relações do PAPEL são extraídas sem intervenção humana, o que torna o processo de revisão e análise das relações extraídas fundamental para a melhoria do recurso disponibilizado. Com relação ao PAPEL, minhas atividades foram:

Revisão e análise de mais de 200 relações

Colaboração no desenvolvimento da interface do **VARRA** – **Validação, Avaliação e Revisão de Relações semânticas no AC/DC**, usando as relações do PAPEL.

Escrita do “Manual de utilização do VARRA”.