ETIQUETAGEM DE VERBOS PARA O PROCESSAMENTO AUTOMÁTICO DO PORTUGUÊS BRASILEIRO: PROCEDIMENTOS DE CONSTITUIÇÃO DE *CORPUS*[*]

Raquel Meister Ko. Freitag – Universidade Federal de Sergipe
Edna Caroline Alexandria da Cunha – Universidade Federal de Sergipe
Flávia Regina Evangelista – Universidade Federal de Sergipe
Rebeca Rodrigues de Santana – Universidade Federal de Sergipe

RESUMO: Neste trabalho, apresentamos os procedimentos de constituição de um corpus de treino para a etiquetagem de verbos do português brasileiro não somente com valores morfossintáticos, mas contemplando também os valores de tempo, aspecto e modalidade. No âmbito do projeto "Anotação semântico-discursiva de verbos para o processamento de línguas naturais" (FREITAG, 2012). Discutimos os problemas decorrentes do uso de dados reais e explicitamos a importância da etiquetagem semântico-discursiva da categoria verbo para o processamento automático.
PALAVRAS CHAVE: Verbos. Etiquetagem. Processamento automático.

INTRODUÇÃO

Vivemos em um tempo que as relações entre homens e máquinas estão bem próximas, não sendo possível separar, muitas vezes, certas atividades, como pesquisas, produções, traduções e leituras de textos dos computadores. Nem sempre traduções automáticas chegam a se aproximar da língua-alvo desejada e que sistemas de busca muitas vezes trazem informações que não interessam à pesquisa feita.

Para que tais atividades sejam realizadas com mais eficiências é preciso que linguistas descrevam os nossos sistemas linguísticos de modo que estas informações sejam transformadas em uma linguagem que os computadores compreendam, a linguagem de programação, por meio de ferramentas e tecnologias computacionais. Esta é a área de atuação da Linguística Computacional, ciência que vai interpretar as teorias linguísticas em um domínio computacional.

No âmbito do projeto "Anotação semântico-discursiva de verbos para o processamento de línguas naturais" (FREITAG, 2012), que visa desenvolver etiquetas para verbos do português brasileiro não somente com valores morfossintáticos, mas contemplando também os valores de tempo, aspecto e modalidade na nossa língua, neste trabalho, apresentamos os procedimentos de constituição de um corpus de treino para este processo, salientando a tensão entre a língua natural e o processamento automático e explicitamos a importância da etiquetagem da categoria verbo no português.

1 TRADUÇÃO AUTOMÁTICA E O VERBO

[*] X EVIDOSOL e VII CILTEC-Online - junho/2013 - http://evidosol.textolivre.org

Verbos constituem uma classe gramatical complexa, pois envolvem uma gama de valores semântico-discursivos amplos, como a codificação da noção de tempo, de aspecto e de modalidade, o que torna complicado o seu processamento computacional, pois muitas vezes os analisadores automáticos não conseguem captar suas minúcias de significação. Isso se deve a descrições e sistematizações ainda incipientes, tanto no nível intralinguístico como translinguístico. Diferentes projetos têm se dedicado ao tratamento de informação de tempo, aspecto e modalidade, com propostas de anotação (manual e automática) especialmente em *corpora* do inglês e do espanhol (cf. SAURÍ; PUSTJOVSKY (2008), BAKER et alii (2010), SAURÍ; VERHAGEN; PUSTEJOVSKY (2006), MORANTE; DAELEMANS (2012), VÁZQUEZ; MONTRAVETA (2009), entre outros). A anotação de traços de tempo, aspecto e modalidade em *corpora* é tarefa prioritária para o processamento de línguas naturais, na medida em que subsidia estudos contrastivos entre as línguas de modo a permitir o aprimoramento de processos de tradução automática.

Todas as línguas do mundo codificam, de alguma forma, as noções de tempo, aspecto e modalidade (BYBEE; PERKINGS; PAGLIUCA, 1994). A noção de tempo refere-se à ordenação dos eventos, em função do momento da fala; aspecto refere-se ao tempo interno dos eventos; e a modalidade, às atitudes acerca dos eventos. No português, a correlação entre formas e funções não é unívoca, sendo o significado da forma verbal dependente do contexto (cf. FREITAG, 2013).

Uma descrição detalhada do funcionamento das categorias de tempo, aspecto e modalidade no português foi realizada no âmbito do projeto "Variação na expressão do tempo passado: funções e formas concorrentes", que analisou dados de fala e de escrita de uma amostra sociolinguística do português brasileiro.

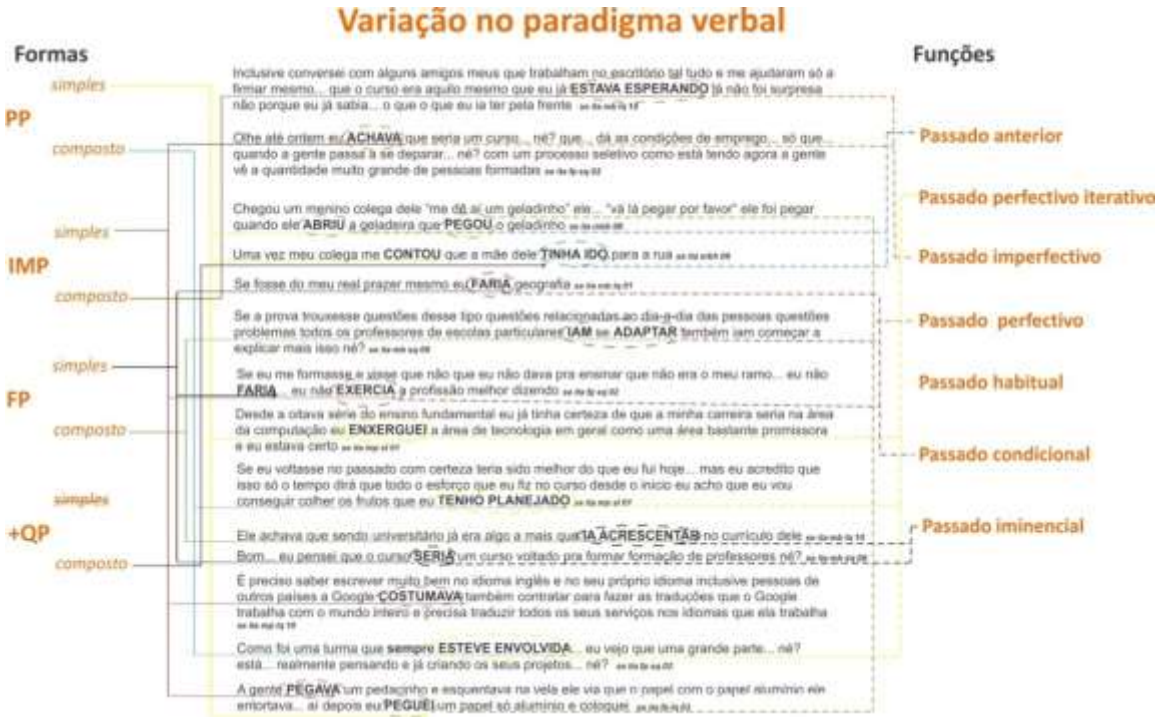

Figura 1: Sistematização da correção entre formas e funções dos verbos do português (FREITAG, 2012, p. 390).

No domínio do passado, modo indicativo, foram identificadas as seguintes formas verbais:

- Pretérito Perfeito simples (PP simples)
- Pretérito Perfeito composto (PP composto)
- Pretérito Imperfeito simples (IMP simples)
- Pretérito Imperfeito composto (IMP composto)
- Futuro do Pretérito simples (FP simples)
- Futuro do Pretérito composto (FP composto)
- Pretérito Mais que Perfeito composto (+QP composto)

Estas formas são responsáveis pela codificação dos seguintes valores: Passado anterior (um evento passado anterior a outro evento, também passado); passado perfectivo iterativo (um evento passado que ocorre sistematicamente do passado ao presente), passado imperfectivo (um evento passado em referência a outro evento, também passado); passado perceptivo (um evento passado cuja referência é o momento da fala); passado habitual (um evento passado com recorrência irregular); passado condicional (um evento decorrente de outro evento passado); e passado iminencial (um evento apresentado antes de sua ocorrência).

Etiquetas semântico-discursivas de verbos são do interesse da comunidade acadêmica que desenvolve pesquisas na área de processamento de línguas naturais, particularmente com os desenvolvedores de softwares *taggers* e *parsers.* Para tanto, as etiquetas precisam ser testadas em ambientes autênticos e que propiciem a anotação automática.

2 PREPARAÇÃO DO CORPUS

"*Corpus* é um corpo de linguagem natural (autêntica) que pode ser usado como base para pesquisa linguística" (SINCLAIR, 1995 apud SARDINHA, 2004, p.17). Para o desenvolvimento de etiquetas para os verbos do português, tomamos como corpus uma amostra chamada "Banco de Dados de Escrita – textos narrativos e opinativos", constituída para subsidiar a análise da variedade culta da língua, e a estratificação das tipologias abordadas (opinião/narração), descrevendo os contextos de uso de tais estratégias linguísticas (ARAUJO, PEIXOTO, FREITAG, 2012). Este *corpus* eletrônico, em formato eletrônico, possui 10.000 palavras (é um micro corpus) e foi constituído levando em consideração as variáveis *sexo* e *escolaridade* (Ensino Médio e Ensino Superior). Dos oitenta textos que compõem o corpus, 40 são opinativos/argumentativos e 40, narrativos. Os 40 textos de opinião/argumentação foram subdivididos da seguinte forma: 10 do primeiro ano do Ensino Médio, 10 do segundo ano, 10 do terceiro ano e 10 produzidos por alunos de nível superior. Cada subgrupo de 10 produções textuais é composto por 05 textos elaborados por estudantes do sexo masculino e 05 por estudantes do sexo feminino. O mesmo se dá no agrupamento dos 40 textos narrativos.

Como nosso objetivo é etiquetar tal corpus em uma ferramenta eletrônica, percebemos a necessidade em fazer a revisão e normatização ortográfica destas produções textuais, pois o software utilizado para a etiquetagem reconheceria mais de uma forma para a mesma palavra, o que geraria ambiguidade nas análises.

Realizamos, também, a conversão dos arquivos, um a um, para o formato *.txt, por se tratar de uma extensão que pode ser facilmente lida ou aberta por qualquer programa que lê texto e que, por essa razão, é considerada universal. Em seguida, cada arquivo *.txt foi submetido ao software Dexter Converter, transformando-os para o formato *.xml, de

forma a permitir a adição de metadados cujas especificações identifica o informante, como número do falante, sexo e faixa etária.

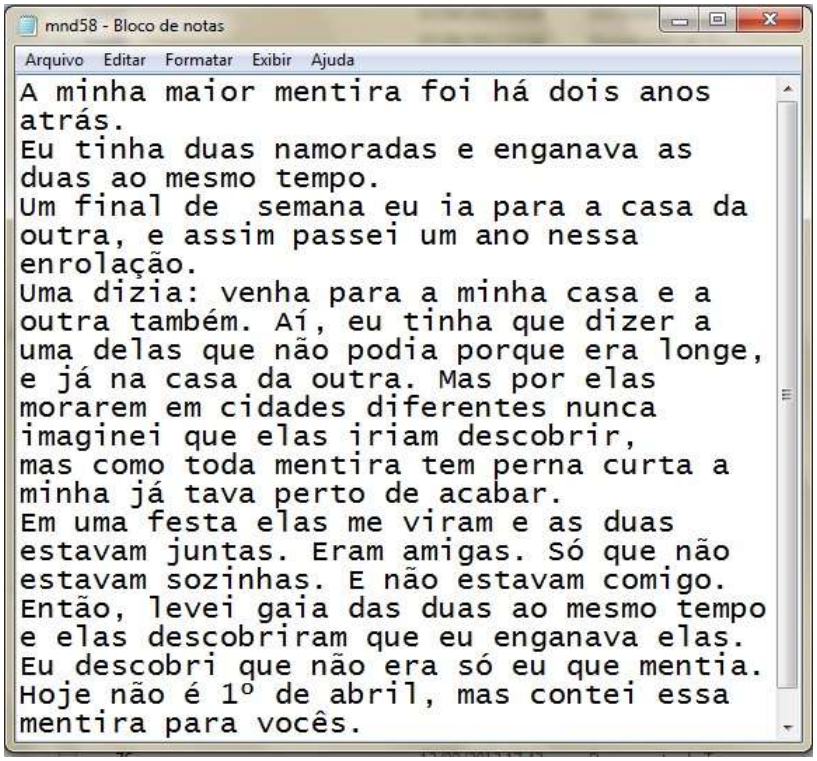

Figura 2: Amostra de texto do Banco de Dados de Escrita – Textos Narrativos e Opinativos (com ortografia revisada e em formato txt.)

Após a normatização, conversão e etiquetagem, o *corpus* serve para dados de treino em programas como *taggers* e *passers*. O corpus de treino é constituído por

> palavras etiquetadas previamente que o etiquetador lê para 'aprender' vários aspectos importantes para o seu funcionamento, como as etiquetas existentes no corpus, as sequências de etiquetas mais adequadas, qual a etiqueta mais provável caso uma palavra desconhecida seja encontrada num texto etc. (SARDINHA, 2005, p.13)

Ao final, disponibilizaremos as etiquetas desenvolvidas à comunidade acadêmica – particularmente, aos desenvolvedores de softwares para o processamento de línguas naturais – e usuários da língua portuguesa e interessados em conhecê-la e/ou aperfeiçoar-se nela, pois, observamos que na Era Digital e diante de um mundo globalizado, a língua portuguesa adquiriu maior representatividade e interesse entre as nações (REHM; USZKOREIT, 2012).

CONCLUSÃO

Nosso objeto de estudo alinha-se a uma área que, apesar de não ser nova, ainda é pouco explorada no âmbito da língua portuguesa. Reconhecemos a necessidade de um investimento maior em relação às produções e pesquisas nas descrições e formalizações das estruturas linguísticas do português para uso em novas tecnologias. Por conseguinte, pesquisas desenvolvidas neste ramo contribuem para elevar a qualidade da produção e

inovação científica nas áreas temáticas relacionadas ao Processamento de Línguas Naturais e à Linguística Computacional. Afinal, linguistas capazes de descrever formas linguísticas para o processamento automático computacional do português brasileiro podem se integrar em empresas tecnológicas com atividades no ramo de PLN. Portanto, vemos como crucial o diálogo intrínseco em torno da produção científica nas áreas de linguística e informática.

Sob esse cenário, e, levando-se em consideração os desafios colocados pela sociedade da informação num mundo globalizado, verifica-se, pois, a necessidade de se concentrar mais esforços para fomento nos recursos linguísticos, visando o desenvolvimento de ferramentas e aplicações para o processamento computacional do português. Assim, o incremento de tecnologia da linguagem para a língua portuguesa é de fundamental importância para a consolidação do português como uma língua de comunicação internacional e com projeção global.

REFERÊNCIAS

ARAUJO, A. S.; PEIXOTO, J. C.; FREITAG, R. M. **Banco de Dados de Escrita - Textos Narrativos e Opinativos**. In: JORNADA DE PESQUISA CIÊNTÍFICA DO GEMPS/CNPq, 2, 2012, Aracaju. Artigo, Anais, 2013. p.10.

BYBEE, J.; PERKINGS, R.; PAGLIUCA, W. **The evolution of grammar:** tense, aspect, and modality in the language of the world. Chicago: The University of Chicago Press, 1994.

FREITAG, R. M. K. . Past tense in Brazilian Portuguese: set of tense-aspect-modality features. In: **Proceedings of the VIIth GSCP International Conference Speech and Linguistic Analysis**. Firenze: Firenze University Press, 2012. p. 388-392.

MORANTE, R.; DAELEMANS, W. Annotating Modality and Negation for a Machine Reading Evaluation. In: **CLEF 2012**. Disponível em http://www.clef-initiative.eu/documents/71612/463956e9-2b22-4e68-aa39-b711302c97b1

OTHERO, Gabriel; MENUZZI, Sérgio. **Linguística Computacional**: **teoria e prática**. São Paulo: Parábola Editorial, 2005.

REHM, Georg; USZKOREIT, Hans (Org.). **A língua portuguesa na era digital**. Coleção livros brancos – Meta Net. Springer: Berlim, 2012.

SARDINHA, Tony Berber. **Linguística de Corpus**. Barueri: Manole, 2004.

SAURÍ, R.; PUSTEJOVSKY, J.. From Structure to Interpretation: A Double-layered Annotation for Event Factuality. In: **Prooceedings of the 2nd Linguistic Annotation Workshop**. The Sixth International Conference on Language Resources and Evaluation, LREC, 2008, p. 1-8.

SAURÍ, R.; VERHAGEN, M.; PUSTEJOVSKY, J. Annotating and Recognizing Event Modality in Text. In: **Proceedings of the 19th International FLAIRS Conference, FLAIRS 2006**. Melbourne Beach, Florida, 2006, p. 333.338.

VÁZQUEZ, G.; MONTRAVETA, A. Ampliación del Banco de Datos de Verbos del español SenSem. In: CANTOS, P.; SÁNCHEZ, A. (eds.), **A Survey on Corpus-based Research.** Panorama de investigaciones basadas en corpus. Murcia: Universidad Murcia, 2009, p. 957-969.