DOI: 10.5007/1807-9288.2011v7n2p149

## SEÇÃO ENTREVISTA

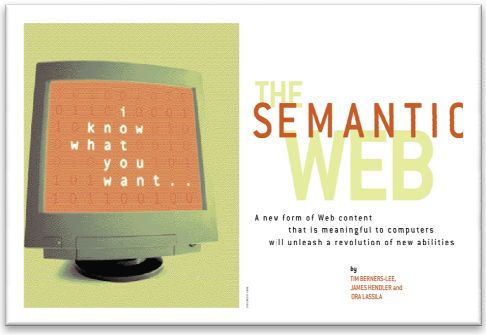
**Everton Vinicius de Santa**
Universidade Federal de Santa Catarina
evertonrep@yahoo.com.br

Os modos de cooperação entre homem e máquina há muito deixaram de estar ligados apenas ao plano da força física e da ideia que elucidava uma relação de interdependência e evolução, como na Revolução Industrial. O que vivenciamos agora, mais do que nunca, é uma relação de cooperação no plano mental em que a máquina pensa pelo homem e com ele, embora (ainda?) dependa deste. É possível que a relação língua–máquina chegue próximo do que a mente humana é capaz de arquitetar, mesmo com a intervenção humana, considerando que essa relação é de dependência e muito pouco autônoma para a máquina? Haveria alguma relação com o que se entende, grosso modo, como pós-humano (e aqui podemos considerar o cenário "apocalíptico" do supercomputador *HALL-9000,* de Kubric) no sentido dessa complexa relação entre mente humana e linguagem? São questões como estas que nos levam à Web Semântica, sobretudo porque têm como finalidade conseguir atribuir um significado (sentido) aos conteúdos publicados na internete, de modo a torná-los perceptíveis tanto pelo humano quanto pelo computador.

Nesta edição da *Texto Digital*, quem poderá nos ajudar a entender melhor esses conceitos e ideias é uma autoridade no assunto: o Prof. Bento Carlos Dias da Silva, doutor em Linguística e Língua Portuguesa /Processamento Automático de Língua Natural, com especialização em Linguística Computacional na Universidad Carnegie Mellon, EUA, professor assistente doutor e pesquisador da Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP), desde 1987, pesquisador do Núcleo Interinstitucional de Linguística Computacional (NILC), do ICMC da Universidade de São Paulo (USP), São

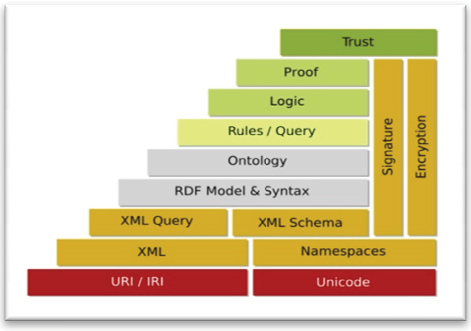
Carlos, desde 1993, e avaliador do SINAES-MEC (BASis), desde 2007. Tem experiência nas áreas de Linguística, Processamento Automático de Línguas Naturais e Letras, com ênfase em Teoria e Análise Linguística, atuando principalmente nos seguintes temas: descrição do português e do inglês, léxico, léxico-gramática, semântica lexical, *wordnets* e ontoléxicos. Ele esteve recentemente na Universidade Federal de Santa Catarina (UFSC), a convite do Núcleo de Pesquisas em Informática, Literatura e Linguística (NUPILL), para ministrar a oficina *Introdução à Web Semântica e ao Processamento Automático de Língua Natural* e nos concedeu esta entrevista.

**TD:** É um prazer poder contar mais uma vez com sua colaboração, professor. Começo com uma questão primordial: O que é ou o que podemos entender como Web Semântica?




**Prof. Bento:** A Web Semântica, tal como idealizada por Tim Berners-Lee, James Hendler e Ora Lassila, em artigo publicado em 2001 na revista *Scientific American*, é a extensão semântica da Web Sintática que conhecemos hoje. Idealizada por Berners-Lee em 1989, a Web Sintática é desprovida de uma codificação do conteúdo dos documentos (dados). Tece apenas a estrutura dos documentos e as relações entre eles por meio das conhecidas anotações especificadas nas linguagens de marcação/formatação de (hiper)texto (cuja evolução é acentuadamente marcada pelas siglas Tex > Latex > GML > SGML > HTML > XML). Sem descartar as tecnologias já desenvolvidas e amplamente empregadas na "versão sintática" da Web (HTTP, UNICODE, HTML, CSSs, Javascripts, APIs, URIs, URLs, entre outras), as tecnologias que constituem a Web Semântica, incentivadas pelo *World Wide Web Consortium* (o W3C, criado em 1994), desenvolvem padrões para a criação e interpretação de conteúdos da Web.

A ideia é incrementar a Web atual, através de novas tecnologias que modelam e exploram a dimensão lógico-semântica que está "dormente" nos dados e possibilitam a troca de informações semanticamente plenas tanto entre os agentes artificiais que circulam na Web quanto o "diálogo" entre esses agentes e os usuários, nas tarefas que envolvem o acesso, a interpretação, o uso e o compartilhamento dessa massa de dados que, hoje, calcula-se estar distribuída em 8 bilhões de páginas atualmente indexadas na Web.




Como um "bolo em camadas", as tecnologias que deverão consolidar a Web Semântica entrelaçam-se: na base, as tecnologias de codificação do material escrito (UNICODE), de localização (URL) e de identificação (URI) dos dados; mais acima, as formas de especificação em termos das linguagens sucessivamente cada vez mais lógico-semantizadas: a linguagem RDF (fornece os esquemas básicos que correspondem a predicações semelhantes a uma estrutura oracional do tipo "Sujeito – Predicado – Objeto"), a linguagem das ontologias (fornece as bases de conhecimento em termos de relações de significado entre os dados e metadados) e Lógica (possibilita a extração de novas informações a partir de informações dadas); mais no topo e finalizando o "bolo", há as camadas que correspondem às tarefas de avaliação, que consiste na sanção do valor de verdade, da fidedignidade e do grau de confiança das informações manipuladas na Web.

**TD:** E quanto ao Processamento Automático de Língua Natural, o PLN? Como isso funciona?

**Prof. Bento:** O estudo do PLN, do ponto de vista histórico, foi inaugurado, simultaneamente ao encrudescimento da Guerra Fria entre os EUA e a então URSS, no final da década de 40, com as primeiras propostas de construção de sistemas de tradução automática envolvendo o russo e o inglês. Desde então, e buscando subsídios teórico-metodológicos na linguística pura e computacional, nas ciências da computação e na inteligência artificial, as investigações nesse campo se expandiram e, hoje, elas abraçam o estudo descritivo de fenômenos da linguagem com vistas à construção de uma ampla gama de sofisticadas **tecnologias linguísticas**, ou seja, tecnologias que são capazes de modelar aspectos da competência e do desempenho linguísticos humanos como, por exemplo, corrigir, ler e traduzir textos, responder e fazer perguntas, dar instruções ou selecionar e organizar informações extraídas de textos.

Essas tecnologias, por sua vez, podem ser classificadas em três grandes categorias:

- **As aplicações de PLN**, cuja parte significativa já se encontra disponível e comercializada em diferentes **tipos de programas que automatizam, total ou parcialmente, tarefas que envolvem material de natureza linguística**, como a correção ortográfica e/ou gramatical, a leitura, a tradução e a sumarização de textos, a recuperação e a extração de textos e informações, a produção de textos (orais e escritos) e a participação em diálogos;
- **Os aplicativos de PLN**, que formam o núcleo interno das aplicações e não são "visíveis" ao usuário, são **os tokenizadores** (Do inglês: *tokenizer*), programas específicos que normalizam a informação textual a ser processada pela máquina, marcando todo tipo de informação extra-gramatical, como títulos, subtítulos, limites de parágrafos, de palavras, de morfemas, letras maiúsculas, sinais de pontuação e diacríticos, travessões,

aspas, entre outros; **os *parsers*** (Do latim: *pars orationis* = "partes do discurso"), programas que realizam as tarefas de análise léxico-gramatical das frases que integram os textos e que podem estar associados a programas de desambiguização léxico-gramatical; **os etiquetadores**, programas que rotulam as palavras de um texto segundo diferentes tipos de informação, como, por exemplo, informação sintática (substantivo, verbo, adjetivo, pronome, numeral, etc.), informação semântica (agente, paciente, experienciador, beneficiário, locativo, etc.) e informação pragmática (marcadores discursivos, relações retóricas, trocas de turno, participantes do discurso, tempo, espaço, etc.); **os planejadores** de texto/discurso, entre outros;

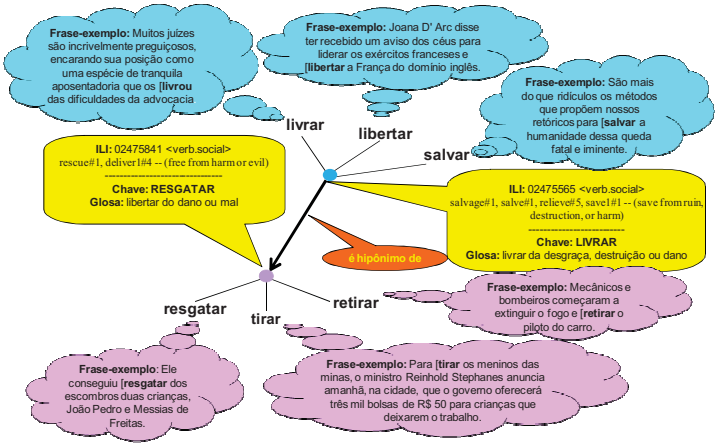
- **Os recursos do PLN**, que constituem **o *lingware***, isto é, os recursos linguísticos computacionalmente codificados e construídos para integrar aplicações e aplicativos, e as bases de conhecimento léxico-gramatical e conceitual dos sistemas de PLN, são os recursos como dicionários, glossários, enciclopédias, *thesaurus*, léxicos, ontologias, ontoléxicos, gramáticas e *corpora* textuais.

**TD:** A sua equipe está envolvida em projetos ligados ao PLN e um desses projetos envolve a construção de uma base lexical para a montagem de uma *wordnet* para o português brasileiro. O que é uma rede *wordnet*? Esse projeto é pioneiro no Brasil?

**Prof. Bento:** No contexto do PLN, uma rede *wordnet* ("rede de palavras") é um recurso do tipo ontoléxico. Trata-se de um tipo particular de modelagem computacional do conhecimento que os falantes possuem das palavras da sua língua, e, portanto, dos conceitos ou das entidades do mundo por ela evocados, e das relações de sentido (abstratas) que se estabelecem entre elas. Os falantes do português brasileiro, por exemplo, conhecem as palavras *zagueiro*, *beque*, *homem*,

*mulher*, *rosa*, *flor*, *andar, correr*, *trotar*, *nariz*, *orelha*, *cabeça*, *roncar*, *dormir*, *quebrar* e são capazes de perceber relações de sentido entres essas palavras, aqui indicadas com os **termos técnicos**: *zagueiro* e *beque* são **sinônimas**, porque têm o mesmo sentido; *homem* e *mulher* são antônimas, porque têm sentidos opostos; *rosa* é **hipônima** de *flor*, porque a rosa é um tipo de flor; *correr* e *trotar* são **tropônimas** de *andar*, porque *correr* e *trotar* são modos específicos de andar: andar com velocidade e andar a trote, respectivamente; *nariz* e *orelha* são merônimas de *cabeça*, porque o nariz e as orelhas são partes da cabeça; *roncar* **acarreta/implica** *dormir*, porque não é possível roncar sem se estar dormindo; *matar* causa *morrer*, porque a consequência necessária de matar é morrer.

A iniciativa de construir uma rede de palavras para modelar computacionalmente essas relações de sentido coube a pesquisadores da Universidade de Princeton, EUA, que cunharam o termo *WordNet*, hoje sinônimo de "rede semântica" (<http://wordnet.princeton.edu/>) e um modelo computacional do que se pode considerar "léxico mental". Outras iniciativas se espalharam por todos os cantos do globo. Para o português, há as iniciativas pioneiras de pesquisadores em Portugal, a rede *WordNet.PT* (<http://www.clul.ul.pt/clg/wordnetpt/index.html>), e no Brasil, a rede *WordNet.Br* (<http://www.nilc.icmc.usp.br/~carol/wn.html>).




Em linhas muito gerais, visualizamos uma rede *wordnet* como uma rede cujos nós são os *synsets* (os conjuntos de sinônimos que representam os conceitos simbolizados pelas palavras sinônimas, podendo ser

também um conjunto unitário, quando não se detectam palavras sinônimas) e as ligações entre os nós são as relações de sentido (hiponímia, meronímia, acarretamento e causa) que se estabelecem entre os *synsets*. A relação de antonímia, excepcionalmente, estabelece-se entre palavras específicas de *synsets,* e não entre *synsets*.

A melhor referência para a visualização da intrincada malha de relações de uma rede *wordnet* é o VISUALTHESAURUS[THINKMAP], uma aplicação que implementa a rede *WordNet de Princeton* numa rede 3D (<http://www.visualthesaurus.com/>).

**TD:** Um outro projeto de pesquisa desenvolvido sob sua coordenação tem relação com a elaboração de um *Thesaurus Eletrônico para o Português do Brasil* (o *TeP*). O que é um *thesaurus*? Há alguma relação com a *WordNet.Br*?

**Prof. Bento:** No *TeP* (<http://www.nilc.icmc.usp.br/tep2/index.htm>), a ideia foi construir um pequeno recurso lexical, tanto para consulta pelo usuário da língua quando para uso pelo projetista de sistemas de PLN. Nele, cerca de 44.000 palavras do português (11.000 verbos, 17.000 substantivos, 15.000 adjetivos e 1.000 advérbios) estão organizadas em termos de sinônimos e de antônimos, incluindo frases que contextualizam o uso de cada palavra, ou seja, como em um *thesaurus*, que pode ser concisamente definido como um acervo (livro ou bando de dados lexicais) que contêm uma listagem hierarquizada de sinônimos. Essa base lexical foi o ponto de partida da construção da rede *WordNet.Br*, pois forneceu o acervo inicial dos *synsets* (os conjuntos de sinônimos do *TeP*) que compõe o núcleo da rede.

**TD:** De que modo a construção ou organização de ontologias e de ontoléxicos se beneficia do estudo do PLN?

**Prof. Bento:** Como assinalei nas considerações sobre a Web Semântica, o estudo do PLN desempenha papel essencial no

desenvolvimento de qualquer tecnologia que envolva ou esteja associada à modelagem de aspectos do conhecimento e do desempenho linguísticos. Sua importância é igualmente decisiva no desenvolvimento de um ontoléxico, pois este se trata de um constructo computacionalmente tratável que toma a forma de uma base de dados léxico-conceituais constituída de um ou mais léxicos (acervo das unidades lexicais de uma língua) e de uma ontologia única cujos conceitos são inventariados a partir das unidades lexicais que os simbolizam. Essa tarefa envolve, de um lado, o levantamento de unidades léxicas e a especificação dos conceitos a elas associados e, de outro, a sistematização dos conceitos em uma ontologia, atividades que integram aspectos linguísticos e computacionais, portanto, próprios do estudo do PLN. Especificamente, as técnicas propostas no estudo do PLN podem auxiliar no desenvolvimento e aprendizagem de ontologias que, como já se disse, são instrumentos de representação do conhecimento decisivos para a viabilização da Web Semântica.

**TD**: Há um projeto em desenvolvimento no NUPILL de uma ontologia de termos literários em que já se buscam alternativas de utilização em aulas de literatura, assim como há outros trabalhos que focam em estatísticas textuais e estudos literários mais específicos. De que modo os estudos ligados à literatura podem ser favorecidos pelo estudo do PLN?

**Prof. Bento:** Como o universo da Literatura, valendo-se das linguagens e, sobretudo, das línguas, pode, sem dúvida, ser modelado em domínios e subdomínios de dados, de conhecimentos e de atividades, cujos agentes e objetos constituintes interagem entre si, agrupam-se e interligam-se de modos complexos e diversos, essa tarefa de modelagem computacional pode também se beneficiar com o uso das tecnologias do PLN e da Web Semântica que resumimos nesta entrevista. Em particular, destaco o uso das tecnologias que são desenvolvidas para a construção e reaproveitamento de léxicos, ontologias e ontoléxicos, elementos essenciais para a consolidação de

uma “web semântica”, em que agentes humanos e artificiais atuam para a disseminação e o compartilhamento de dados, conhecimentos, informações e descobertas.

*