Universidade de Brasília

Instituto de Letras

Departamento de Línguas Estrangeiras e Tradução

Processamento Automático de Línguas Naturais:

Um estudo sobre a localização do IBM Watson™ para o português do Brasil

Aline Rocha de Sousa

Brasília

2015

Aline Rocha de Sousa

Processamento Automático de Línguas Naturais:

Um estudo sobre a localização do IBM Watson™ para o português do Brasil

Monografia apresentada ao Departamento de Línguas Estrangeiras e Tradução do Instituto de Letras da Universidade de Brasília, para obtenção do título de Bacharel em Línguas Estrangeiras Aplicadas ao Multilinguismo e à Sociedade da Informação.

Orientador: Prof. Dr. Cláudio Gottschalg Duque

Brasília
2015

Aline Rocha de Sousa

Processamento Automático de Línguas Naturais:

Um estudo sobre a localização do IBM Watson™ para o português do Brasil

Monografia apresentada ao Departamento de Línguas Estrangeiras e Tradução do Instituto de Letras da Universidade de Brasília, para obtenção do título de Bacharel em Línguas Estrangeiras Aplicadas ao Multilinguismo e à Sociedade da Informação.

Comissão Examinadora

_____

Prof. Dr. Cláudio Gottschalg Duque (Orientador)
Faculdade de Ciência da Informação - FCI
Universidade de Brasília

_____

Prof. Dr. Cláudio Corrêa e Castro Gonçalves
Departamento de Línguas Estrangeiras e Tradução - LET
Universidade de Brasília

_____

Dr. Milton Shintaku
Instituto Brasileiro de Informação em Ciência e Tecnologia - IBICT

Resumo

O Processamento Automático de Línguas Naturais é um domínio de pesquisa promissor quanto à organização de dados digitais na Sociedade da Informação, ao buscar recuperá-los pela língua de registro. Por outro lado, as especificidades de cada língua exigem tratamentos computacionais diferentes e, se não compreendidas em profundidade, podem constituir entraves à produção tecnológica. Esta pesquisa tem por objetivo investigar as particularidades do processamento automático do português brasileiro, comparado ao inglês americano e ao português europeu, a partir de um estudo de caso sobre a localização do sistema de perguntas e respostas Watson da IBM. Inicia-se com um levantamento teórico sobre o estado-da-arte do Processamento de Línguas Naturais, aborda as especificidades da arquitetura DeepQA no tratamento linguístico do sistema e detalha as particularidades do português do Brasil que devem ser observadas na localização do Watson. Da investigação, possíveis adaptações ao sistema são apresentadas, e a importância da inclusão do português brasileiro e, de modo geral, do Brasil na Sociedade da Informação é posta em reflexão.

Palavras-chave: Processamento Automático de Línguas Naturais, português do Brasil, IBM Watson™.

SOUSA, A. R. **Processamento Automático de Línguas Naturais**: Um estudo sobre a localização do IBM Watson para o português do Brasil. Monografia (Graduação em Línguas Estrangeiras Aplicadas) — Universidade de Brasília, Brasília, 2015.

Abstract

Automatic Natural Language Processing is a promising research domain to the organization of digital data in the Information Society, as it seeks to recover them by the language in which they are registered. Nevertheless, the specificities of each language require different computational treatments and, if not understood in depth, may hinder the technological production. This research aims to investigate the particularities of the automatic processing of Brazilian Portuguese, compared to American English and European Portuguese, from a case study on the localization of IBM Watson question-answering system. It begins with a theoretical survey on the state of the art of Natural Language Processing, addresses the specificities of the DeepQA architecture in processing language for the system and details the particularities of Brazilian Portuguese that have to be regarded in Watson's localization. From the research, possible adaptations to the system are presented, and the importance of including Brazilian Portuguese and, in general, Brazil in the Information Society is brought into consideration.

Keywords: Automatic Natural Language Processing, Brazilian Portuguese, IBM Watson™.

SOUSA, A. R. (2015). *Automatic Natural Language Processing: A study on the IBM Watson's localization to Brazilian Portuguese* (Undergraduate thesis). University of Brasilia, Brasilia, Brazil.

Sumário

1 Introdução

1.1 Dados de identificação da pesquisa

Monografia de autoria de Aline Rocha de Sousa, matrícula 2010/0091041, sob orientação do Prof. Dr. Cláudio Gottschalg Duque, da Faculdade de Ciência da Informação – FCI, da Universidade de Brasília, apresentada como requisito parcial para a obtenção do título de Bacharel em Línguas Estrangeiras Aplicadas ao Multilinguismo e à Sociedade da Informação.

1.2 Tema

1. Processamento Automático de Línguas Naturais[1]
2. Processamento automático do português brasileiro
3. O sistema Watson

1.2.1 Delimitação do tema

Esta pesquisa se restringe ao Processamento Automático de Línguas Naturais aplicado à variedade brasileira da língua portuguesa, comparada ao inglês americano e ao português europeu. As discussões se inserem em um estudo de caso sobre a atual localização[2] do sistema IBM Watson™ ao português do Brasil.

1.3 Formulação do problema

O surgimento da Sociedade da Informação está aliado a uma grande explosão informacional decorrente do acúmulo desordenado de conhecimento, sobretudo no meio digital. Com o intuito de organizar esse conhecimento, pelo menos, em parte, várias tecnologias de Processamento Automático de Línguas Naturais (PLN) procuram acessá-lo

[1] Denominação sinonímica de Processamento de Linguagem Natural, conforme considerações teóricas de Dias-da-Silva (2006) a respeito da diferenciação entre os termos "língua" e "linguagem".
[2] O termo remete ao processo de adaptação de um produto, aplicação ou conteúdo de documento à realidade linguística, cultural e normativa de um país. Na pesquisa, pretendemos limitar-nos aos aspectos linguísticos.

pela língua em que foi registrado, cuja compreensão direta seria impossível ao computador, que apenas reconhece números e códigos exatos. Além disso, tais tecnologias têm facilitado a interação homem-máquina e, portanto, melhorado o acesso ao conhecimento pré-organizado, de maneira que um indivíduo pode pesquisá-lo fazendo uso apenas de sua própria língua.

Em 2011, a International Business Machines (IBM) lançou uma tecnologia de PLN com potência de organização do conhecimento e compreensão da língua humana de maneira tal que seria capaz de ultrapassar as habilidades do ser humano em uma competição de perguntas e respostas sobre conhecimentos gerais. Em janeiro do mesmo ano, o supercomputador Watson participou do programa de televisão Jeopardy!, nos Estados Unidos, e venceu dois dos melhores competidores da história do programa (BORENSTEIN, S; ROBERTSON, 2011). Pouco depois, a empresa anunciou que aplicaria a tecnologia à resolução de problemas reais nas áreas da saúde, das finanças e dos negócios e não levou muito tempo para que o Watson provasse a sua eficiência em pesquisas de oncologia (VOLTOLINI, 2013), nas atividades de marketing e até na descoberta de novas receitas culinárias (STRICKLAND, 2014).

A tecnologia, no entanto, era limitada à língua inglesa e uma preocupação primordial da empresa, além de estender o Watson a outras áreas, era fazê-lo compreender outras línguas. Uma parceria foi firmada com o SoftBank e o Watson começou a *aprender* japonês, sua segunda língua e uma das mais difíceis para aprendizes que têm o inglês como língua nativa (INTERNATIONAL BUSINESS MACHINES [IBM], [2005]b). Em 2014, a IBM firmou também uma parceria com o Bradesco e o Watson passou a *estudar* sua terceira língua: o português brasileiro (MATSU, 2015).

Desde outubro, a IBM Brasil está empenhada em converter o sistema para a língua portuguesa, mas, por se tratar de uma tecnologia de Processamento Automático de Línguas Naturais, uma simples tradução não permitiria ao Watson compreender e organizar informações com registro nessa língua. É necessário, em vez disso, adaptar todo o sistema às especificações que o português possui no que diz respeito ao seu processamento computacional, para que o sistema possa, enfim, ser utilizado em 2016.

Quais seriam então, essas particularidades do processamento automático do português brasileiro e quais adaptações serão necessárias à localização do sistema para posterior comercialização no país? É o que se pretende investigar nesta pesquisa.

1.4 Justificativa

A Sociedade da Informação, não adstrita à explosão informacional no meio digital, resulta de uma verdadeira Revolução da Tecnologia da Informação (CASTELLS, 2005), com efeitos sensíveis na economia, na cultura e na sociedade, de modo geral. Em uma era social em que a informação constitui um elemento de poder, é necessário, além de organizar o conhecimento acumulado por ambos, ciência e senso comum, elaborar políticas e fornecer ferramentas para um respectivo acesso democrático. Nesse sentido, é preciso assegurar que as línguas que veiculam a informação não se sobreponham, mas tenham presença equilibrada nos meios de comunicação, garantindo o futuro da diversidade linguística e o consequente respeito à identidade cultural. Portanto, as tecnologias de processamento linguístico devem contemplar o maior número possível de línguas naturais.

O português, em especial, "é a quinta língua com maior número de falantes no mundo, com cerca de 220 milhões de falantes em quatro continentes – África, América, Ásia e Europa. Das línguas europeias, é a terceira língua com maior número de falantes no mundo. Face aos desafios colocados pela sociedade da informação num mundo globalizado, verifica-se a necessidade premente de se concentrarem mais esforços quer na criação de recursos linguísticos quer na investigação e desenvolvimento de ferramentas e aplicações para o processamento computacional do português" (BRANCO *et al*., 2012).

Nesse contexto, é fundamental que se produza tecnologias especialmente voltadas para a língua portuguesa e que, antes disso, linguistas se predisponham a estudar o português de forma a prepará-lo para o processamento automático. Assim também na conversão de tecnologias já bem-sucedidas em outras línguas para o português, que é o caso do sistema IBM Watson. Considerando que o papel do profissional de Línguas Estrangeiras Aplicadas ao Multilinguismo e à Sociedade da Informação é justamente assegurar a difusão multilíngue de informações no meio digital, cabe a ele, também, a valorização de sua língua-mãe, estando apto a contribuir em projetos de tecnologia do português – sobretudo de conversão para o português – se predisposto a concentrar-se no PLN.

Esta pesquisa se baseia principalmente nesses preceitos, podendo contribuir: a) para a elucidação das particularidades do português brasileiro em comparação com o inglês americano e inserido no contexto do Processamento Automático de Línguas Naturais; b) para a conversão do IBM Watson ao português brasileiro, de forma subsidiária; c) para facilitar futuros trabalhos de conversão; c) para despertar o interesse de linguistas e profissionais de

LEA-MSI no PLN; d) para disseminar os conceitos dessa área interdisciplinar e recente no Brasil.

1.5 Objetivos

1.5.1 Objetivo geral

Analisar as peculiaridades do processamento automático do português brasileiro, comparado ao inglês americano, que poderão constituir entraves à localização do sistema Watson da IBM no Brasil.

1.5.2 Objetivos específicos

- Identificar as referências que abordam o estado da arte de aplicações do Processamento Automático de Línguas Naturais;
- Estudar o funcionamento do Watson via referências;
- Verificar as particularidades envolvidas no processamento automático do português do Brasil;
- Discutir as prováveis adaptações do sistema frente às particularidades do processamento automático do português brasileiro.

1.6 Embasamento teórico

"Segundo Dias-da-Silva (2006), a possibilidade de interação homem/máquina por meio da língua dos homens e o surgimento dos primeiros sistemas de tradução automática impulsionaram os estudos ou investigações que receberam o nome *Processamento Automático de Línguas Naturais* (do inglês, *Automatic Natural Language Processing* ou *Natural Language Processing*)" (FELIPPO; DIAS-DA-SILVA, 2008). Hoje com diversas outras aplicações, que vão desde ferramentas para o estudo dos fenômenos linguísticos até sistemas especializados na resolução de problemas por meio de acesso à informação, o PLN constitui um domínio de pesquisa interdisciplinar, com contribuições teórico-metodológicas

da Linguística, da Linguística Computacional, das Ciências da Computação, da Inteligência Artificial, da Matemática, da Lógica, da Filosofia e da Psciologia (DIAS-DA-SILVA, 2006). O objetivo geral é permitir ao computador "compreender" e gerar sentenças na língua do usuário, para tanto, traduzindo a língua em formalismos inteligíveis por máquina e fazendo uso de modelos estatísticos para analisar padrões do comportamento linguístico.

O IBM Watson, por sua vez, é um sistema de perguntas e respostas (*question-answering system* ou QA,), ou seja, uma aplicação do PLN capaz de manter um diálogo com seu usuário em língua natural. Em seu processamento, faz uso da tecnologia de computação cognitiva, que simula o processamento linguístico realizado pela mente humana e permite ao sistema *aprender* de forma semelhante ao ser humano, em uma tentativa de aproximá-los em termos de linguagem e *pensamento*. A arquitetura do sistema é composta por 5 fases: 1. na **Análise da Pergunta** (*Question Analysis*), é realizada automaticamente uma análise sintática e semântica da própria questão do usuário, utilizando, para isso, uma série de tecnologias do PLN (um *parser* estatístico e componentes para o reconhecimento de entidades nomeadas, resolução de anáforas e extração de relações); 2. a fase de **Geração de Hipóteses** (*Hypothesis Generation*) produz todas as possíveis respostas para a pergunta, a partir de corpora com dados estruturados ou não; 3. na **Recuperação de Evidências** (*Supporting Evidence Retrieval*), o sistema busca novos dados que evidenciem cada uma das hipóteses geradas na fase anterior; 4. a fase de **Pontuação de Hipóteses e Evidências** (*Hypothesis and Evidence Scoring*) utiliza vários algoritmos quantificadores para medir a relevância de cada hipótese como resposta à pergunta; a **Compilação e Classificação Final** (*Final Merging and Ranking*), faz uma coletânea das hipóteses e de seus respectivos escores e aplica um modelo de Aprendizagem de Máquina para classificá-las hierarquicamente (CORTIS *et al.,* 2014). O sistema original, no entanto, é baseado em tecnologias voltadas para o processamento automático do inglês norte-americano – especialmente no *parser* da gramática formal English Slot Grammar (ESG) – e é necessário adaptá-lo ao processamento automático do português brasileiro com a inserção de ferramentas locais.

No que diz respeito às particularidades da língua portuguesa, há de se considerar que "o português é uma língua românica, pelo que a maioria do seu léxico deriva do Latim" [7], enquanto o inglês pertence ao ramo das línguas germânicas e, consequentemente, as propriedades do seu léxico são bastante distintas. "A ordem básica das palavras em português [e em inglês] é dita ser SVO − Sujeito Verbo Objeto (*ele leu o livro*)", mas, enquanto no português ela varia mais por questões estilísticas e de ênfase ("*o livro, ele não leu*", OSV), no

inglês, há casos de uso comum da língua em que a mudança de ordem é obrigatória, como nos verbos auxiliares de perguntas (*Do you like it?*, Aux-S, inversão da forma canônica SV). Por outro lado, "o português é uma língua que permite sujeitos nulos, isto é, o sujeito de uma dada frase pode não estar realizado foneticamente (_ *li o livro*). [...] Esta é uma das características do português que representa um desafio acrescido para a análise sintática automática dos textos e da fala." Além disso, "o paradigma flexional do português é muito mais rico que o de línguas como o inglês, em particular no que diz respeito aos verbos" e "posição dos pronomes clíticos na frase [(*dá-lo-ei*)] é outra característica que coloca desafios específicos ao processamento automático da língua portuguesa". O português do Brasil, especificamente, possui uma grande variação linguística em virtude de sua extensão geográfica e sua diversidade cultural, o que pode constituir um desafio a mais na localização de um sistema de PLN original da língua inglesa.

1.7 Metodologia

Esta pesquisa se trata de um Trabalho de Conclusão de Curso (TCC) na modalidade monografia. É uma pesquisa exploratória (GIL, 2002 *apud* CAJUEIRO, 2013), bibliográfica, com estudo de caso interpretativo e caráter qualitativo.

Seguindo o método de abordagem dedutivo, foi executada nos seguintes passos (e respectivas durações):

- Subtema "PLN": pesquisa bibliográfica, leituras e redação;
- Subtema "Watson": pesquisa bibliográfica, leituras e redação;
- Subtema "Watson em pt-BR": pesquisa bibliográfica, leituras e redação;

Para o subtema "PLN", foram utilizadas publicações de autores brasileiros renomados e obras estrangeiras de cunho clássico ou que sejam referenciadas em MOOCs atuais sobre o domínio do PLN. Em caso de divergências conceituais, foram observados os entendimentos do estudo epistemológico de Bento Dias da Silva (2006). Para o subtema "Watson", foi dada prioridade às publicações do DeepQA Research Team, grupo de pesquisa responsável internacionalmente pelo IBM Watson, aos vídeos oficiais da empresa e aos *press releases* da IBM Brasil. Para o subtema "Watson em pt-BR", relativo aos desafios que o Watson terá de enfrentar diante das particularidades do processamento automático do

português do Brasil, foram utilizadas gramáticas e outras obras específicas do português brasileiro.

1.9 Organização da obra

Esta obra está organizada segundo sua metodologia de pesquisa e objetivos específicos. No capítulo 2, "Processamento Automático de Línguas Naturais", após esta introdução, aborda-se o PLN sob a perspectiva de níveis linguísticos. São discutidos aspectos relativos à área, a dificuldade do processamento da língua pelo computador, a metodologia envolvida no tratamento linguístico-computacional, os níveis em que a língua é dividida para acesso pela máquina a suas especificidades, os tipos básicos de arquitetura de um sistema de PLN e suas respectivas aplicações. Nesse capítulo, os exemplos são dados em inglês americano, língua de partida da localização.

Em seguida, no capítulo 3, "IBM Watson" discorre-se sobre o sistema: o desafio da IBM em produzir um computador para competir com humanos em um quiz de alto nível; a arquitetura do sistema resultante e cada uma de suas fases; e diversas aplicações do Watson no mercado, após o sucesso no Jeopardy!. Ainda aqui, os exemplos estão em inglês e contextualizam-se com o programa de televisão a que o sistema participou.

O capítulo 4, "Watson em português brasileiro", traz um estudo prévio sobre os projetos de internacionalização e localização do sistema — Watson Multilíngue e Watson em outros idiomas —, passa para as particularidades do português brasileiro frente ao inglês americano e ao português europeu e, então, apresenta possíveis adaptações que o sistema terá de incluir quando localizado para o português do Brasil.

O capítulo 5, "Discussões finais" traz uma reflexão sobre a situação do Brasil e da variedade brasileira do português frente ao PLN e à Sociedade da Informação. Finalmente, o capítulo 6 lista as referências bibliográficas citadas ao longo do texto.

2. Processamento Automático de Línguas Naturais

2.1 Conceito

O Processamento Automático de Línguas Naturais (PLN) é uma área interdisciplinar que se propõe a estudar e elaborar técnicas para o tratamento dos diversos idiomas por meio do uso do computador. Pela forma como é constituído, o computador é incapaz de compreender comandos ambíguos, como se caracteriza a comunicação humana, sendo trabalho do PLN organizar a língua em modelos exatos que possam ser inteligíveis por máquinas. Feito isso, é possível construir uma série de ferramentas que facilitam a interação do computador com o usuário, otimizam atividades de edição e tradução de texto ou permitem aos estudiosos investigar com maior precisão os fenômenos linguísticos.

Com início na década de 1950, o estudo do tratamento de línguas pelo computador possui raízes em várias áreas, nas quais recebe abordagem e título diferenciados: Linguística Computacional, em Linguística; Processamento de Linguagem Natural, em Ciência da Computação; Reconhecimento de Fala, em Engenharia Elétrica; e Psicolinguística Computacional, em Psicologia (JURAFSKY; MARTIN, 2008). Atualmente, pelo avanço das pesquisas e pela complementaridade que tais abordagens têm, esse estudo vem se consolidando como uma área à parte, sobretudo sob o título de Processamento de Linguagem Natural. Bento Carlos Dias da Silva (2006, p. 115), no entanto, atenta para o fato de que "linguagem" recebe conotação diferente em Linguística:

> O termo 'linguagem', por ser de aplicação mais geral que o termo 'língua', é licitamente usado para denotar os sistemas de comunicação em geral, naturais e artificiais, entre seres humanos ou não: as linguagens de programação, a linguagem das abelhas, a linguagem corporal humana, a linguagem do trânsito, etc.

Enquanto isso, "línguas" denomina apenas os sistema de comunicação humana que utilizam palavras, a exemplo do inglês, do espanhol e do francês. São elas o objeto do PLN, afinal. Além disso, o título original em inglês *Natural Language Processing* é também usado em Ciência Cognitiva para denominar o estudo do processamento linguístico realizado pela mente, ao passo que *Automatic Natural Language Processing* é exclusivo para computadores. O autor sugere, então, a tradução "Processamento Automático de Línguas Naturais", embora também aceite o título "Processamento de Línguas Naturais" (como em DIAS-DA-SILVA *et al.*, 2007). Ambas as denominações foram adotadas ao longo deste trabalho.

2.2 Breve história

O interesse em fazer com que máquinas compreendam a língua humana data do próprio surgimento do computador, na década de 1940 (DIAS-DA-SILVA *et al.*, 2007). Para que o computador pudesse "entender" instruções e, portanto, executar as tarefas que lhe incumbiam, foram criadas as linguagens de programação, compatíveis com seu funcionamento lógico. No entanto, as linguagens de programação possuem, ainda hoje, demasiada rigidez e, para tornar os computadores mais acessíveis ao público geral, os cientistas visualizaram duas soluções: construir interfaces gráficas que mascarassem a codificação computacional ou criar programas capazes de interpretar comandos em línguas naturais. A primeira opção era indiscutivelmente mais simples e foi, de fato, a adotada, mas logo a necessidade de elaborar programas especiais que também utilizassem a língua como material de entrada, como tradutores automáticos e sistemas de perguntas e respostas, deu origem a estudos paralelos sobre o Processamento Automático de Línguas Naturais.

Conforme Dias-da-Silva *et al.* (2007, p. 6), "as primeiras investigações institucionalizadas sobre o PLN começaram a ser desenvolvidas no início da década de 50", após a distribuição de uma carta da Fundação Rockfeller convidando universidades e instituições de pesquisa a desenvolverem trabalhos sobre tradução automática (do inglês *machine translation*). À época, os Estados Unidos participavam da Guerra Fria com a União Soviética e traduzir documentos russos com maior rapidez facilitaria avanços tecnológicos frente à adversária. Assim, em 1952, foi realizada a primeira conferência científica sobre tradução automática no Instituto de Tecnologia de Massachussets (MIT) e, em 1954, a primeira demonstração de um sistema capaz de traduzir do russo para o inglês, na Universidade de Georgetown. As produções subsequentes, no entanto, ficaram muito aquém do esperado pelas instituições financiadoras e o consequente corte de incentivos desaqueceu as pesquisas da área.

Na década de 60, segundo Madeleine Bates (1995), o foco se voltou para a produção de sistemas de perguntas e respostas. Buscava-se fazer com que o computador respondesse a perguntas digitadas pelo usuário e, para tanto, o conhecimento disponível sobre determinado assunto era previamente codificado e armazenado em um banco de dados. Cada vez que o usuário fazia uma pergunta ao sistema, respeitados os limites de tema e formato da oração, a pergunta era, então, analisada de forma profunda e a interpretação resultante permitia ao computador recuperar a resposta ideal no banco. Além disso, nessa época, foram criados os

primeiros *corpora* online, isto é, amostras de textos reais para investigação da língua pelo computador: o Brown Corpus do inglês americano, em 1964, e o dicionário de dialetos chineses DOC (*Dictionary on Computer*), em 1967 (JURAFSKY; MARTIN, 2008).

Ainda segundo Bates (1995), o interesse por aplicações além de interfaces de bancos de dados cresceu na década de 70, mas o foco se manteve em sistemas de comunicação verbal com o usuário. Um dos destaques do período foi o SHRDLU ou "Mundo dos Blocos", de Terry Winograd. O sistema consistia na representação gráfica do braço de um robô que manipulava blocos sobre uma mesa conforme comandos digitados no teclado e, por sua eficiência, corroborou a hipótese de que a interação homem-máquina por meio da língua era possível (DIAS-DA-SILVA *et al.*, 2007; JURAFSKY; MARTIN, 2008). As aplicações da época, todavia, dependiam de uma compreensão completa dos comandos ou perguntas de entrada: cada simples palavra podia alterar a interpretação, o que gerava resultados ou perfeitos, ou totalmente falhos.

Na década seguinte, as aplicações de PLN elaboradas até então começaram a ser comercializadas. Na pesquisa, o objetivo agora era compreender trechos escritos cada vez maiores, não limitados a comandos e perguntas digitadas ao computador (BATES, 1995). Para tanto, os pesquisadores passaram a produzir sistemas que, em vez de analisar palavra por palavra do material de entrada, buscavam extrair o significado do texto em geral, com os quais poderiam examinar textos maiores e obter resultados melhores, se avaliados de forma absoluta. Nessa tarefa, os métodos baseados em probabilidade, sobretudo investigados pelo Centro de Pesquisa Thomas J. Watson da IBM, exerceram papel fundamental (JURAFSKY; MARTIN, 2008). Nessa época, houve, ainda, esforços para gerar automaticamente sentenças em língua natural, ademais de compreendê-las.

Na década de 90, as técnicas se mesclaram e uma infinidade de aplicações foram produzidas: sistemas de fala, que combinavam reconhecimento de sentenças faladas pelo usuário com interpretação linguística; sistemas de geração de língua natural, para a emissão de sentenças pelo computador; sistemas de classificação automática de textos conforme seu conteúdo; interfaces de língua natural mais interativas; entre outros tantos. Mesmo a Tradução Automática voltou a ser pesquisada, com aplicações de texto e fala. Diversos *corpora* foram compilados e disponibilizados para utilização nesses sistemas, o que permitiu aos cientistas dispensar os antigos exemplos manipulados para trabalhar com a língua em sua forma mais natural (BATES, 1995). Além disso, com a expansão da *World Wide Web*, houve necessidade

de recuperar as informações ali difundidas com o uso de ferramentas baseadas em texto (JURAFSKY; MARTIN, 2008).

Já no início dos anos 2000, a principal novidade foi a aprendizagem de máquina (*machine learning*), um método que, aplicado ao PLN, é usado para "ensinar" o computador a reconhecer padrões na língua e, com isso, conseguir prever classificações gramaticais e atribuições de ordem semântica automaticamente, poupando a intervenção de agentes humanos na elaboração de aplicações. Para "treinar" os sistemas, *corpora* como o PropBank e o Penn Discourse Treebank, com anotações sintáticas, semânticas e pragmáticas, logo foram disponibilizados e os sistemas de alto desempenho produzidos no período facilitaram o armazenamento e a compilação das informações. Em outra categoria do método (aprendizagem não supervisionada), estudos começaram a ser realizados para eliminar, inclusive, a necessidade de *corpora* anotados, reduzindo ainda mais as intervenções (JURAFSKY; MARTIN, 2008).

Atualmente, apesar da multiplicidade de técnicas de PLN, persiste o interesse em aprimorar os sistemas para que sejam mais independentes do homem, embora mais próximos dele em questão de processamento cognitivo. Com a necessidade de organizar a imensa quantidade de dados presentes na Web, em formatos diferentes, técnicas como a aprendizagem não supervisionada automatizam o processo e tem sido exploradas em larga escala. Além disso, a abordagem conexionista da Inteligência Artificial começa a exercer maior influência nos estudos linguístico-computacionais graças a seus modelos biologicamente motivados (CAMBRIA; WHITE, 2014). A tendência é que, nos próximos anos, sistemas de Processamento de Línguas Naturais simulem o cérebro com eficiência, sendo capazes de assimilar os conceitos do mundo à sua volta, compreender o ser humano emocional e culturalmente e, por incluir seus próprios traços orgânicos, atendê-lo melhor.

2.3 A complexidade do processamento linguístico

Fazer com que o computador compreenda a língua humana não é uma tarefa fácil. Do latim *computare* (calcular), o computador é simplesmente uma máquina eletrônica que recebe instruções para processar dados por meio de cálculos. Os circuitos do microprocessador, considerado o núcleo do dispositivo, contam com chaves (transistores) que permitem ou impedem a passagem da corrente elétrica, a depender do sinal que lhes é

transmitido. Se o sinal for 0, a chave é desligada; se for 1, a chave é ligada e a corrente pode passar. Nesse contexto, em que apenas duas situações são possíveis, a combinação dos algarismos[3] é que faz a representação dos dados que se quer processar. Para a codificação dos caracteres do inglês, por exemplo, é utilizado o padrão ASCII (American Standard Code for Information Interchange), que combina os algarismos em cadeias de sete (VAHID, 2008):

Tabela 1. Padrão ASCII

Símbolo	**Codificação**	**Símbolo**	**Codificação**
R	1010010	r	1110010
S	1010011	s	1110011
T	1010100	t	1110100
L	1001100	l	1101100
N	1001110	n	1101110
E	1000101	e	1100101
0	0110000	9	0111001
.	0101110	!	0100001
<parágrafo>	0001001	<espaço>	0100000

Exemplos de codificações em ASCII. Fonte: VAHID (2008).

Na prática, transmitir manualmente cada código numérico se faz inviável e, por isso, as instruções são dadas ao computador por meio de linguagens de programação. O vocábulo "ABBA", por exemplo, seria codificado como "1000001 1000010 1000010 1000001", uma cadeia para cada caractere. No caso do padrão Unicode, utilizado para representar caracteres de diversos idiomas do mundo, o número de algarismos por cadeia sobe para 16 ou até 32 para línguas ideográficas, dificultando ainda mais o processo (VAHID, 2008). Dessa maneira, conveio criar linguagens de programação para facilitar o comando do computador e a manipulação dos dados nele armazenados. As instruções são, portanto, escritas em linguagens como Java, Lua e Python e, só então, codificadas para o sistema numérico binário com o auxílio de programas compiladores ou interpretadores.

Embora facilitem a comunicação humana com o computador, tais linguagens de programação diferem radicalmente das línguas naturais. Ambas possuem regras para a formação de sentenças (sintaxe) e significados atrelados à sua estrutura (semântica), mas,

[3] Também chamados *bits*, abreviação de *binary digits*. Computadores modernos de uso pessoal são capazes de processar 32 ou 64 bits ao mesmo tempo.

enquanto as linguagens de programação são projetadas para uma interpretação unívoca pelo computador, assegurando um comando preciso, as línguas humanas surgiram e fluem naturalmente, caracterizadas por inúmeros casos de ambiguidade. Para a oração em inglês *I made her duck*, por exemplo, é possível perceber ao menos 5 interpretações (JURAFSKY; MARTIN, 2008):

(2.1) Eu cozinhei um pato para ela;

(2.2) Eu cozinhei o pato dela;

(2.3) Eu construí o pato artificial que ela tem;

(2.4) Eu a fiz desviar-se;

(2.5) Eu a transformei em um pato.

Isso faz com que a língua seja árdua para o processamento automático. Apenas o contexto poderia evidenciar qual das interpretações acima é a correta e, mesmo com acesso a informações complementares sobre a situação, seria necessário fornecer ao computador conhecimentos sobre o vocabulário, a estrutura e o universo conceitual da língua inglesa, além de fazê-lo capaz de aproveitar esses conhecimentos de forma sistemática e autônoma. Em outras palavras, seria preciso prover o computador da mesma base linguística e extralinguística de um falante nativo e de semelhante capacidade cognitiva de processamento.

A complexidade é tal, que "para muitos, a habilidade de computadores processarem a língua tão bem quanto humanos significará a chegada de máquinas realmente inteligentes" (JURAFSKY; MARTIN, 2008, p. 6). É o caso de Alan Turing, que, diante da imprecisão do que seria "pensar", na década de 50, propôs um teste empírico baseado no uso computacional da língua para determinar se uma máquina era inteligente: em um ambiente separado, sem que o juiz pudesse ver, a máquina teria de dialogar com ele e conseguir passar-se por humana. Conquanto tivesse sofrido críticas ao longo dos anos, como as do filósofo da mente John Searle, "o Teste de Turing é ainda hoje considerado como parâmetro para a avaliação de certos programas de inteligência artificial" (OTHERO; MENUZZI, 2005, p. 34).

Dadas as divergências entre a língua humana e a dinâmica de funcionamento do computador, cabe aos estudiosos do Processamento de Línguas Naturais atuarem nas duas frentes: modelar a língua em formatos lógico-matemáticos, de maneira a eliminar ou reduzir ambiguidades, e criar mecanismos para que o computador possa tratá-la. Nessa tarefa, utilizam uma ampla metodologia, sobre a qual se discorre a seguir.

2.4 Metodologia

O Processamento Automático de Línguas Naturais, por sua natureza interdisciplinar, compreende técnicas e estudos conceituais de diversas áreas do conhecimento. Para Bento Carlos Dias da Silva (2006, p. 133), esse empréstimo se faz necessário na "complexa tarefa de criar um simulacro computacional da competência e do desempenho linguísticos humanos". Classificados por disciplinas matrizes, ele sistematiza os principais recursos teórico-metodológicos de que o PLN dispõe na figura abaixo[4]:

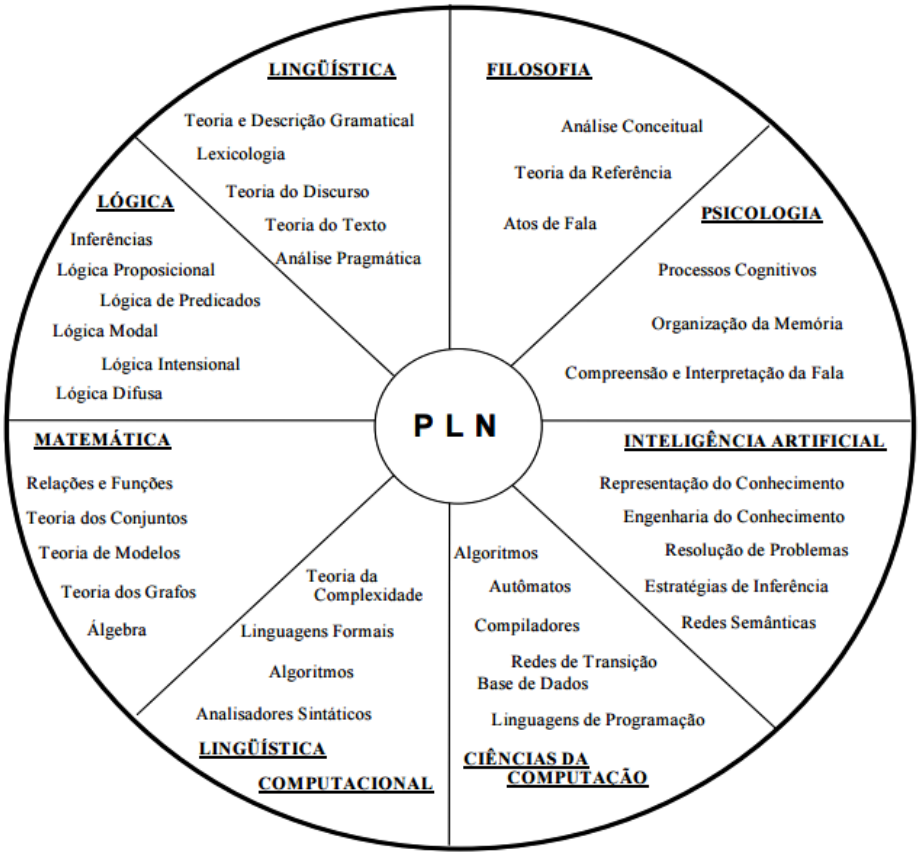


Figura 1. Recursos teórico-metodológicos do PLN. Fonte: DIAS-DA-SILVA (2006).

A adoção de alguns desses recursos em detrimento de outros costuma caracterizar a pesquisa conforme quatro abordagens distintas: simbólica, estatística, conexionista e híbrida (LIDDY, 2001). A abordagem simbólica do Processamento de Línguas Naturais baseia-se em técnicas para transcrever sentenças linguísticas em predicados lógicos, dos quais se pode tirar inferências automaticamente. A abordagem estatística reconhece a língua como um processo aleatório (LUGER, 2013), cujos casos de ambiguidade são previsíveis com o emprego de

[4] Pode-se acrescentar, ainda, o Processamento Digital de Sinais (PDS) da Engenharia Elétrica, responsável pelo estudo da fala humana em sua acústica.

métodos probabilísticos. Além disso, é possível aplicar fórmulas matemáticas a amostras de texto para medir a ocorrência, a classificação gramatical e a combinação das palavras, em geral. De modo semelhante, a abordagem conexionista utiliza cálculos para elaborar modelos generalistas, mas o faz combinado a métodos de representação cognitiva do conhecimento. Finalmente, a abordagem híbrida une as anteriores para conferir melhor desempenho aos sistemas de aplicação.

2.5 Níveis de processamento

A língua possui várias facetas e cada uma das quais é tratada de modo diferente pelo computador. Dias da Silva *et al.* (2007) citam cinco níveis de processamento linguístico, correspondentes aos vários enfoques dados à língua no estudo da Linguística. São eles: i) fonético-fonológico; ii) morfológico; iii) sintático; iv) semântico; v) e pragmático-discursivo. Apesar das peculiaridades, Jurafsky e Martin (2008) atentam para o trabalho comum que esses níveis têm em resolver ambiguidades, presentes desde os traços superficiais aos mais profundos da língua. Nos tópicos a seguir, examina-se o papel de cada nível no tratamento linguístico-computacional e as respectivas tarefas de modelagem e desambiguação linguística.

2.5.1 Nível fonético-fonológico

Na Teoria Linguística, Fonética é o estudo dos sons da fala humana per si, dissociados de qualquer significado. Sua preocupação é analisar o aparelho fonador e as especificidades de sua produção acústica, realizada nos chamados "fones". A Fonologia, por outro lado, estuda os sons particulares a uma língua, influenciados por ela em aspectos socioculturais e, portanto, dotados de significado. Seu objetivo é verificar como os "fonemas" se unem para formar sílabas e como as palavras são pronunciadas e acentuadas pelos falantes. Inseridas no PLN, ambas, Fonética e Fonologia, caracterizam um mesmo nível de processamento linguístico — o dos sons — e contribuem para o reconhecimento automático e a síntese da fala humana.

O reconhecimento de fala é a atividade computacional de transcrição dos sons emitidos pela voz em palavras. Segundo Liddy (2001), o computador analisa as ondas sonoras e faz um registro em sinal digitalizado para posterior aplicação de regras e comparação com modelos da língua em questão. Nesse processo, algumas ambiguidades são inevitáveis, como

é o caso do artigo inglês *the*, que pode ser pronunciado como "thee" ou "thuh", a depender da palavra que o acompanha. Uma regra determinando a pronúncia "thee" antes de palavras que comecem com som de vogal e "thuh" antes de consoantes facilitaria o reconhecimento do vocábulo, nesse caso. Em outros, de resolução mais complexa, o emprego de modelos probabilísticos baseados no teorema de Thomas Bayes, por exemplo, seria satisfatório. Quanto à tipologia, o reconhecimento de fala pode ser por palavras, pronunciadas individualmente, ou de fala contínua, com trechos maiores que devem ser segmentados (JURAFSKY; MARTIN, 2008).

Já a síntese de fala é a reprodução automática de voz humana para a leitura de sentenças geradas pelo computador. Dito de outra forma, é a atividade de fala da própria máquina para o usuário. Conforme Jurafsky e Martin (2008), o processo consiste em dois passos: construir uma representação fonêmica interna, que combine a transcrição das sentenças em alfabético fonético, isto é, em símbolos empregados por linguistas na descrição dos sons da fala humana, a indicativos de ordem prosódica, relacionados com a entonação; e converter o resultado em ondas sonoras, geralmente o comparando a amostras de fala armazenadas no sistema. Nas figuras abaixo, é possível ver a representação fonêmica dos autores para a sentença *PG&E will file schedules on April 20.*, seguida de uma ilustração da onda sonora produzida.

			*			*			* L-L%
P	G	AND	E	WILL	FILE	SCHEDULES	ON	APRIL	TWENTIETH
p iy	jh iy	ae n d	iy	w ih l	f ay l	s k eh jh ax l z	aa n	ey p r ih l	t w eh n t iy ax th

Figura 2. Representação fonêmica interna da sentença *PG&E will file schedules on April 20*. A oração é escrita por extenso e dividida em grupos intermediários, que se destacam como unidades durante a pronúncia. Abaixo das palavras e das letras *P*, *G* e *E*, articuladas individualmente, segue a transcrição fonética no ARPAbet, um alfabeto elaborado para a descrição de fones do inglês americano em caracteres cobertos pelo padrão ASCII. Acima, os asteriscos indicam as palavras realçadas na fala e *L-L%* denota uma queda do tom ao final da sentença, característica de orações declarativas do idioma. Fonte: JURAFSKY; MARTIN (2008).

Figura 3. Reprodução ilustrativa da onda emitida pelo sistema ao pronunciar *PG&E will file schedules on April 20*. Fonte: JURAFSKY; MARTIN (2008).

2.5.2 Nível morfológico

Morfologia é o estudo das palavras que compõem a língua, seja pelas unidades que lhes formam, chamadas morfemas, seja pelas categorias gramaticais nas quais se classificam. A relação entre morfemas e classes de palavras, aliás, é bastante próxima. Se *person*, em inglês, significa "pessoa" e é um substantivo, *personal* é um adjetivo referente a pessoas, *personally* é advérbio e *personality* é um substantivo dele derivado. Ainda que não se conheça o significado dessas palavras, é possível, pois, qualificá-las gramaticalmente pelos morfemas de suas terminações. No Processamento Automático de Línguas Naturais, o objetivo da Morfologia é justamente contribuir com a análise da estrutura das palavras, restando a níveis posteriores determinar o seu real sentido.

Embora não seja o foco do processamento a nível morfológico decidir o significado de cada uma das palavras ou expressões de uma sentença, listar as suas possíveis acepções é importante desde já. É nessa etapa que são elaborados dicionários das línguas naturais que possam ser legíveis pelo computador. Os "léxicos" geralmente são organizados por morfemas (radicais e afixos) acompanhados de informações gramaticais e regras de combinação, a fim de compreender um número maior de palavras em menor volume do que um simples catálogo de itens lexicais. Ademais, em línguas aglutinantes, cujas palavras são formadas pela união de quaisquer morfemas, prever todos os vocábulos seria praticamente impossível e o repositório se tornaria ineficiente. Jurafsky e Martin (2008) exemplificam a palavra turca abaixo:

(2.6a) *uygarlaştıramadıklarımızdanmışsınızcasına*

(2.6b) uygar +laş +tır +ama +dık +lar +ımız +dan +mış +sınız +casına

(2.6c) "(comportando-se) como se vocês fizessem parte daqueles que nós não pudemos civilizar"

Outra tarefa da análise morfológica é a etiquetagem automática ou *part-of-speech* (POS) *tagging*. Após recebido o texto de entrada, as palavras são segmentadas — em um processo chamado tokenização — e passam por uma "normalização", pela qual são padronizadas. Os formatos podem seguir dois critérios: na lematização, as palavras são substituídas por suas formas canônicas, isto é, substantivos e adjetivos flexionados são postos no masculino singular e verbos conjugados são exibidos no infinitivo, tal qual os verbetes de um dicionário tradicional; já no *stemming*, as palavras são reduzidas aos seus radicais, para compatibilização dos dicionários de morfemas. Feito isso, é possível prosseguir com a etiquetagem, na qual as palavras são classificadas segundo sua natureza gramatical. No caso

do inglês, alguns dos rótulos mais usados fazem referência ao Penn Treebank, como NNPS, para substantivo próprio no plural, e DT para artigos. Símbolos e sinais de pontuação são igualmente etiquetados, por interferirem na subsequente reconstrução das sentenças.

Também na etiquetagem há casos que exigem desambiguação. *Can*, por exemplo, pode ser tanto um verbo auxiliar com sentido de "poder", quanto o substantivo comum "lata" ou o verbo "enlatar" em inglês. Para determinar a classificação gramatical correta de suas acepções, é necessário, portanto, o uso de algumas técnicas, como citam Jurafsky e Martin (2008): regras manuscritas; métodos estatísticos, como os modelos ocultos de Andrei Markov e o modelo de entropia máxima; métodos baseados em transformação; e métodos baseados em memória. Tais técnicas observam sobretudo as classes gramaticais das palavras adjacentes (como artigos, que acompanham substantivos), para averiguar qual classe responderia melhor à lacuna e atribuí-la à palavra ambígua.

2.5.3 Nível sintático

Do grego *sýntaxis* (*sýn*: junto + *taxis*: ordenar), a Sintaxe é o ramo da Linguística que investiga a maneira como as palavras se combinam para formar frases. Assim como a Morfologia, ela se limita ao estudo da língua em seu arcabouço, favorecendo, mas não compenetrando, a interpretação plena das orações. No âmbito do Processamento de Línguas Naturais, a Sintaxe "determina o papel de cada uma das palavras de uma sentença e, assim, permite ao sistema convertê-la em estruturas mais facilmente manipuláveis" (COPPIN, 2004, p. 573). Nessa tarefa, utiliza principalmente dois recursos: uma gramática e um *parser* (LIDDY, 2001).

As gramáticas do PLN podem ser definidas como sistemas de regras formais para a descrição linguística, mais especificamente para a descrição das possíveis sentenças constituídas em uma língua. Segundo Dias-da-Silva *et al.* (2007), a mais célebre tentativa de formalização do conhecimento linguístico estrutural se deve a Noam Chomsky, autor da noção de gramáticas de constituintes imediatos ou *phrase-structure grammars* (PSGs). Tais gramáticas têm, todas, relação direta com os elementos constituintes das sentenças, mas poderem variar em quatro tipos, a depender do grau de complexidade e expressividade que suas regras têm: gramáticas regulares, do tipo 3, representam apenas estruturas bem objetivas; gramáticas livres de contexto, do tipo 2, fornecem regras para um número maior de possibilidades; gramáticas sensíveis ao contexto, do tipo 1, podem descrever os casos dos

tipos anteriores e outros mais; e as gramáticas irrestritas, do tipo 0, possuem regras sem limites de descrição. As gramáticas mais utilizadas na formalização sintática das línguas naturais são as do tipo 2 — livres de contexto —, geralmente com regras de reescrita segundo o formalismo de Backus-Naur (John Backus e Peter Naur).

A partir da noção de PSGs, vieram os *parsers*, ferramentas computacionais para a análise sintática propriamente dita. Em primeira instância, a sentença é divida em blocos chamados sintagmas ou constituintes, os quais podem ser classificados conforme a função que assumem na frase (OTHERO; MENUZZI, 2005). Então, os itens são submetidos às regras da gramática. Uma sentença formada por um sintagma nominal e um sintagma verbal, por exemplo, deve ser capaz de reescrever a regra <S> ::= <SN> <SV> de Backus-Naur. Se aprovada pelas regras, a sentença é considerada gramatical e pode ser reestruturada em representações sintáticas do *parser*; caso contrário, é dita agramatical e rejeitada pelo sistema, que a considera inválida do ponto de vista linguístico. A gramaticalidade, todavia, é avaliada apenas em termos de boa formação da frase, pouco importando o sentido gerado, como é típico do processamento sintático. Chomsky esclarece essa questão com a sentença *Colorless green ideas sleep furiously*, perfeitamente organizada sob a ótica da Sintaxe, mas claramente irreal (MANNING; SHÜTZE, 1999).

A representação sintática gerada pelo *parser*, após a avaliação positiva da sentença, possibilita ao computador acessar a estrutura das línguas naturais e lidar com ela a seu modo. A forma mais comum é a representação arbórea, na qual a estrutura das sentenças é ilustrada em uma árvore invertida: a sentença, no topo, simboliza a raiz; os constituintes frasais são reproduzidos em nós de ramos, níveis intermediários; e as palavras aparecem na base, nível terminal da análise sintática (DIAS-DA-SILVA *et al.*, 2007). Nesse processo, além do uso de uma gramática, é essencial o acesso a um léxico ou que as palavras já venham etiquetadas da análise morfológica, para que suas classes gramaticais sejam reorganizadas em sintagmas pelo *parser*. Um sintagma preposicional, por exemplo, deve ser formado por uma preposição e um sintagma nominal, que, a seu turno, pode ser constituído por um artigo e um substantivo. Além das árvores sintáticas, uma maneira de representar a estrutura frasal são os colchetes rotulados, que exibem a distribuição hierárquica do *parsing* de modo compacto. Os exemplos abaixo ilustram a estrutura da sentença *I prefer a morning flight* em ambas as formas de representação (JURAFSKY; MARTIN, 2008). Esse tipo de análise, que parte da sentença para os itens lexicais, é chamado *top-down*. Há também *parsers* com análise *bottom-up*, em sentido inverso, e *chart parsers* e *left-corner parsers*, com análises alternativas.

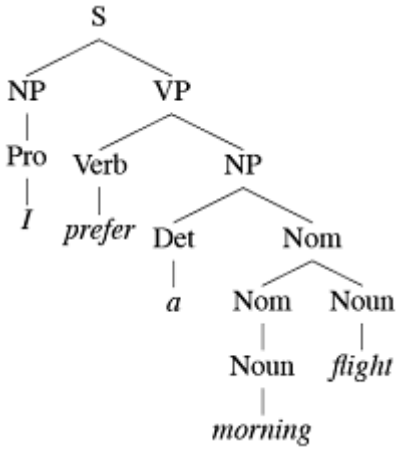


Figura 4. Árvore sintática da sentença *I prefer a morning flight*, em análise *top-down*. *S*: sentença; *NP*: sintagma nominal; *VP*: sintagma verbal; *Pro*: pronome; *Verb*: verbo; *Det*: determinante (artigo); *Nom*: grupo nominal (substantivo adjetivado e substantivo); *Noun*: substantivo. Fonte: JURAFSKY; MARTIN (2008).

(2.7) [$_{S}$ [$_{NP}$ [$_{Pro}$ I]] [$_{VP}$ [$_{V}$ prefer] [$_{NP}$ [$_{Det}$ a] [$_{NOM}$ [$_{N}$ morning] [$_{NOM}$ [$_{N}$ flight]]]]]]

A ideia de que a estrutura da língua é formada por blocos frasais segue o princípio sintático da constituência, mais explorado pela abordagem simbólica do Processamento de Línguas Naturais. A abordagem estatística, por outro lado, dá maior ênfase ao princípio sintático da dependência, segundo o qual as palavras são unidades que, associando-se, dependem umas das outras. O objetivo do *parser*, nesse caso, é descobrir quais palavras exercem funções superiores ou inferiores na frase e como elas se relacionam entre si. O resultado é ilustrado como na figura 5, com arcos que partem das palavras subordinadas e são direcionados àquelas dominantes (MANNING; SCHÜTZE, 1999).

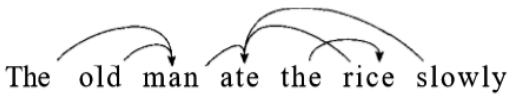


Figura 5. *Dependency parsing* da sentença *The old man ate the rice slowly*. O verbo principal *ate* comanda o sujeito *man*, o objeto *rice* e o adjunto adverbial *slowly*, dos quais os outros itens lexicais são subalternos. Fonte: MANNING; SCHÜTZE (1999).

2.5.4 Nível semântico

Semântica é a área da Linguística que se ocupa do estudo dos significados. Seu objetivo, no PLN, é dar sentido à estrutura frasal já reconstituída em formatos inteligíveis por máquinas, para tanto, aproveitando aspectos de níveis anteriores e adicionando-lhes informações de cunho conceitual. Pode abranger tanto a interpretação das palavras — quando é chamada "Semântica Lexical" — quanto a compreensão de unidades maiores, como é o caso

das colocações *white hair* e *white wine* e da expressão idiomática *to kick the bucket* da língua inglesa, nas quais o sentido é pouco ou não previsível por suas palavras dissociadas (MANNING; SCHÜTZE, 1999). Para interpretá-las, primeiro elabora modelos de representação formal do conhecimento humano e, então, realiza tarefas para relacionar significados adequados à estrutura significante.

Conforme Jurafsky e Martin (2008), algumas das maneiras de representar semanticamente uma sentença são: a lógica de primeira ordem; as redes semânticas; os diagramas de dependência conceitual e as representações baseadas em *frames*. Todas têm em comum o fato de simbolizarem objetos, as suas propriedades e as relações que eles mantêm entre si no mundo extralinguístico, podendo codificar o significado de uma sentença em particular ou representar os conceitos que formam a compreensão humana da realidade, de modo geral. A figura abaixo reproduz os exemplos dos autores para a representação semântica de *I have a car*, nos quais fica clara a correspondência dos símbolos às entidades *speaker* (falante) e *car* (carro) e à relação de posse que o primeiro tem sobre o segundo. Relações semelhantes podem ser aplicadas a outros objetos da realidade, permitindo ao computador acessar a percepção de mundo que o homem tem e, a partir disso, fazer as suas próprias interpretações de materiais com registro linguístico.

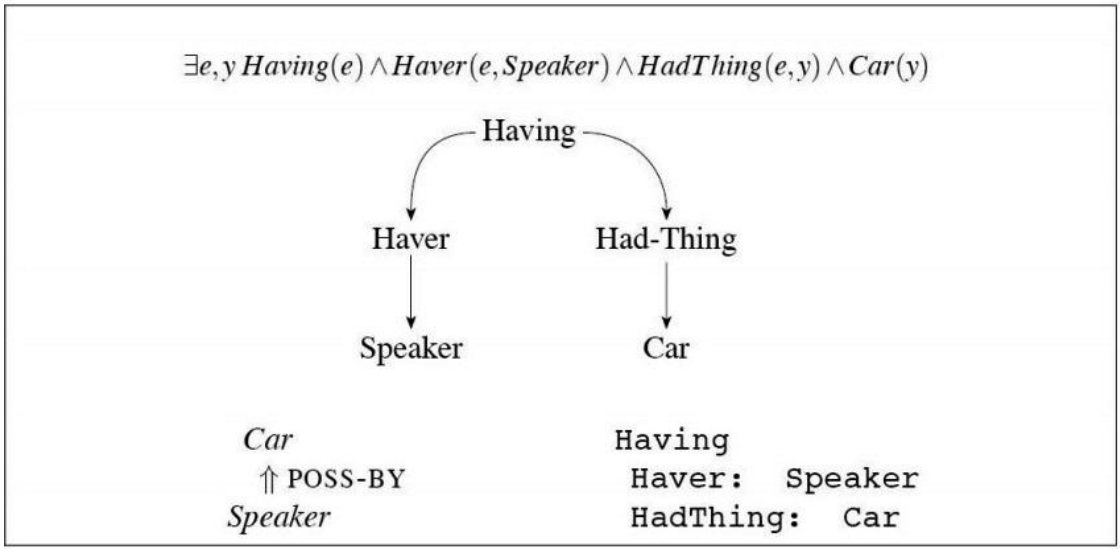


Figura 6 - Exemplos de representação do significado de *I have a car*: uma sentença em lógica de primeira ordem (no topo); uma rede semântica (ao centro); um diagrama de dependência conceitual (canto inferior esquerdo); e um registro baseado em *frames* (canto inferior direito). Fonte: JURAFSKY; MARTIN (2008).

Para automatizar a análise semântica, representações gerais do conhecimento costumam ser pré-delineadas e armazenadas em léxicos, tesauros e ontologias. Os léxicos têm participação importante nos níveis morfológico e sintático, por conterem informações

gramaticais sobre as palavras, e podem contribuir também com a interpretação do sentido linguístico, pelas diversas definições que carregam junto ao vocabulário. Os tesauros, por sua vez, são repositórios de palavras organizadas não por morfemas ou pela ordem alfabética, como nos léxicos, mas pelas relações de significado que elas apresentam quando confrontadas. Tais relações podem ser, por exemplo, a sinonímia, em que duas ou mais palavras possuem significados semelhantes; a antonímia, quando seus traços semânticos são opostos; e a homonímia, quando palavras que assumem a mesma forma possuem significados diferentes. Um recurso que alia as vantagens do léxico às do tesauro é o WordNet, um banco de dados lexicais com substantivos, verbos, adjetivos e advérbios anotados com sentidos e associados entre si semanticamente. No tocante às ontologias, as entidades são organizadas hierarquicamente segundo a especificidade de seus conceitos, geralmente pertencendo à terminologia de um domínio particular do conhecimento.

Certas tarefas são essenciais para que a representação dos conceitos seja corretamente mapeada na estrutura das sentenças, tendo em vista que uma mesma palavra ou expressão pode assumir sentidos diversos. No plano das palavras, a desambiguação de sentido lexical (*word sense disambiguation*), por exemplo, é responsável por pesquisar os repositórios de representações e, dentre várias acepções para um mesmo item lexical, descobrir aquela mais adequada. Outra tarefa comum é computar a similaridade lexical (*word similarity*), isto é, realizar cálculos para determinar a proximidade das relações de um item lexical para outro, com base na distância que apresentam na hierarquia de uma ontologia ou tesauro. O significado de uma sentença não é baseado apenas nas palavras que a compõem, mas também na ordem, no agrupamento e nas relações que elas ali manifestam (JURAFSKY; MARTIN, 2008). Por isso, há também tarefas no plano geral das sentenças, como a rotulagem de papéis semânticos (*semantic role labeling*), que consiste em verificar quais constituintes são argumentos semânticos e atribuir-lhes papéis para a compreensão da frase. Recursos complementares, nesse caso, podem ser acessados, como o *Proposition Bank* (PropBank), que analisa os argumentos a partir do verbo — mais ou menos como um *dependency parser* o faz, a nível sintático — e o FrameNet, que analisa o papel semântico das palavras em blocos e individualmente.

2.5.5 Nível pragmático-discursivo

A Pragmática é o estudo das línguas naturais aplicadas às situações de uso na comunicação. É o ramo da Linguística que, para além da Semântica, observa o significado associado ao contexto, haja vista o fato de que a estrutura linguística pode assumir sentidos conotativos, diferentes do usual. A frase *Can you pass the salt?* exemplifica a necessidade desse tipo de análise. Se dita à mesa, é mais provável que não seja uma pergunta sobre a capacidade do ouvinte passar o sal, considerando o sentido literal do verbo auxiliar *can*, mas um pedido para que o entregue ao emissor da sentença. Segundo Dias da Silva *et al.* (2007, p. 22), a Pragmática se concentra, sobretudo, nos elementos contexto e intenção, ao buscar respostas para as perguntas: "quem são os sujeitos envolvidos na situação discursiva? O que querem dizer esses sujeitos? Qual é o contexto da enunciação?" Para tanto, utiliza abstrações como a Teoria dos Atos de Fala, de John Austin e John Searle, que classifica as mensagens linguísticas conforme a intenção dos interlocutores — como expressar emoções ou firmar um compromisso.

A Análise do Discurso, por sua vez, se preocupa com um contexto mais amplo no qual a língua está inserida: o contexto social. A produção verbal dos falantes varia sempre conforme as condições históricas, econômicas e políticas às quais estão submetidos e o objetivo dessa área é depreender da estrutura linguística a postura ideológica que eles adotam diante da sociedade. Em outras palavras, procura-se compreender a opinião do emissor. Como o discurso costuma ser maior que uma sentença, a Análise do Discurso frequentemente está atrelada à Linguística Textual no PLN, envolvendo, além do contexto, os marcadores discursivos responsáveis por garantir coerência e coesão aos textos (DIAS-DA-SILVA *et al.*, 2007). Para Liddy (2001), isso é importante porque a esse nível o PLN trabalha com o texto como um todo, fazendo conexões entre as sentenças.

O nível pragmático-discursivo pode ser entendido, portanto, como o estágio de análise do material linguístico sob uma visão mais global: o contexto situacional e social confere acuidade à interpretação semântica e a avaliação do texto na íntegra permite deduzir aspectos não observáveis em sentenças isoladas. O contexto é tratado automaticamente com a elaboração de um modelo do usuário que preveja suas preferências, intenções e grau de compreensão do domínio abordado pelo sistema, os quais podem influenciar o tipo de linguagem adotada na comunicação (DIAS-DA-SILVA *et al.*, 2007). Além disso, como na Semântica, faz-se uso de bases de conhecimento e módulos de inferência. Quanto ao texto

discursivo, duas tarefas são especialmente comuns: a resolução de anáforas (*anaphora resolution*) e o reconhecimento de estruturas do discurso (*discourse structure recognition*) (LIDDY, 2001). A resolução de anáforas identifica itens como pronomes e os substitui pelas palavras que eles referenciam. Nos trechos abaixo, por exemplo, o sistema deve ser capaz de perceber que *Peter* e *He* em (2.8a) e *the other passenger* e *The man* em (2.8b) fazem referência à mesma pessoa (MANNING; SCHÜTZE, 1999):

(2.8a) *Mary helped* Peter *get out of the cab.* He *thanked her.*

(2.8b) *Mary helped* the other passenger *out of the cab.* The man *had asked her to help him because of his foot injury.*

Já o reconhecimento de estruturas do discurso determina a função das sentenças no texto, assumindo um papel importante na compreensão da produção linguística e na elaboração de textos pelo computador. Um texto jornalístico do gênero notícia deve ser composto, por exemplo, de: manchete, *lead*, evento principal, contexto, eventos anteriores, história, consequências/reações, expectativa e avaliação (LIDDY, 2001; VAN DIJK, 2004). Da mesma forma, artigos científicos, correspondências e redações administrativas seguem uma estrutura que deve ser observada pelo sistema ao interpretar ou gerar materiais escritos.

2.6 Arquitetura de sistemas genéricos

Os sistemas de Processamento Automático de Línguas Naturais podem ser divididos em dois tipos, conforme a atividade que executam com a língua: interpretação ou geração de línguas naturais. Os sistemas de interpretação de línguas naturais objetivam reconhecer a estrutura linguística e compreender o significado que lhe está atrelado. Os sistemas de geração de línguas naturais, por outro lado, visam a produzir sentenças ou textos maiores automaticamente. Em ambos os casos, os níveis de processamento detalhados acima passam a constituir "módulos linguísticos", nos quais a língua é tratada sob perspectivas diferenciadas, por ferramentas e recursos que se complementam.

Dias da Silva *et al.* (2007) esquematizam a arquitetura dos sistemas de interpretação de línguas naturais na figura abaixo. O material de entrada (*input*) é geralmente uma sentença, que passa por diversos analisadores automáticos até formar um material de saída (*output*) que represente o sentido de forma inteligível pelo computador. Enquanto o material flui pelos módulos linguísticos, recursos com informações sobre a estrutura da língua — como léxico e

gramática — são acessados, bem como modelos do domínio e do usuário que ajudam a compreender o "microuniverso" no qual as entidades conceituais e os interlocutores humanos estão inseridos.

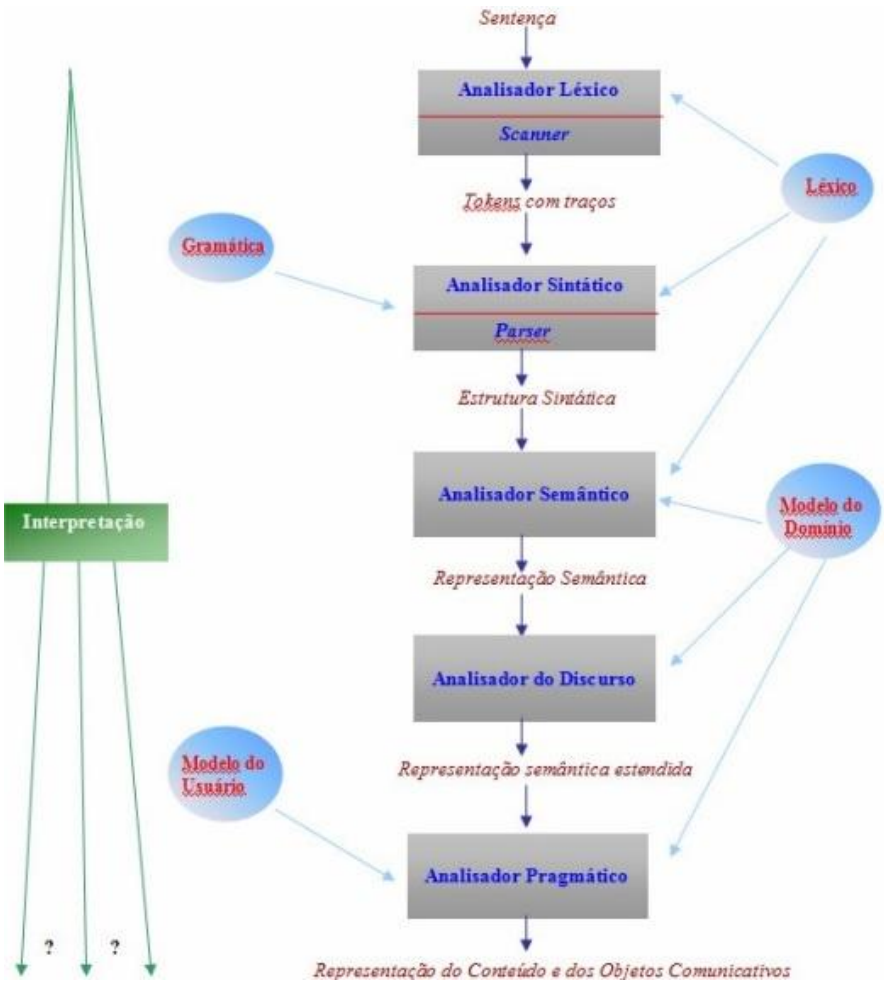


Figura 7 - Arquitetura de um sistema genérico de interpretação de línguas naturais. Fonte: DIAS-DA-SILVA *et al.* (2007).

Nos sistemas de geração de línguas naturais, os módulos são, todavia, menos delimitados. O percurso do material linguístico é aproximadamente inverso àquele realizado na interpretação, pelo que o processo de geração se inicia com representações do significado e finaliza com a estrutura superficial da língua, mas difere substancialmente nos mecanismos de tratamento automático. Os recursos e modelos associados são os mesmos, mas os analisadores dão lugar a outros tipos de ferramentas, tendo em vista que o objetivo então não é analisar, mas produzir sentenças e textos em línguas naturais. Pela flexibilidade desses sistemas, Dias da Silva *et al.* (2007) dividem a arquitetura de geração de línguas naturais em apenas três passos básicos, como se observa na figura 8: seleção do conteúdo, no qual o sistema determina as informações que quer transmitir ao usuário, ainda a nível abstrato; planejamento

do texto, quando ele formula a mensagem em notações lógicas; e realização do texto, em que, apoiado na gramática e no léxico, ele reescreve a mensagem em língua natural.

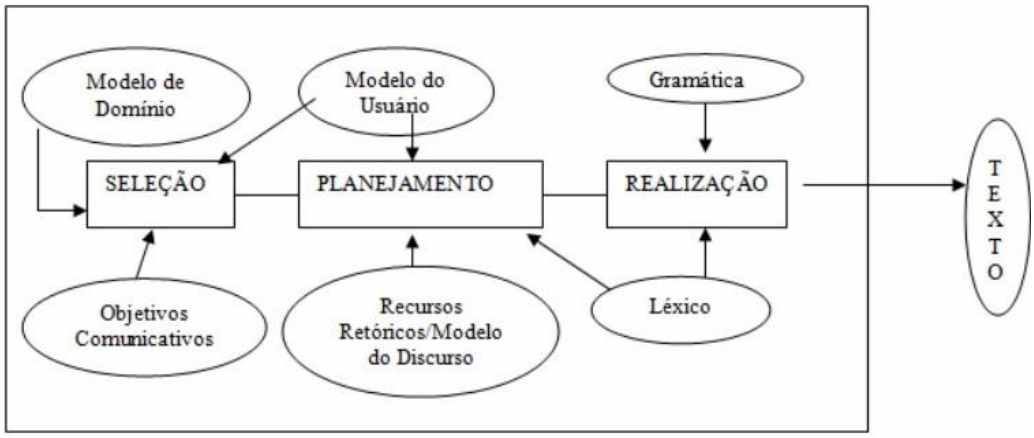


Figura 8 - Fases principais de um sistema de geração de línguas naturais. Fonte: DIAS-DA-SILVA *et al.* (2007).

À parte dos esquemas genéricos, outras variações são observáveis na arquitetura de sistemas de línguas naturais. Mesmo pela figura 7, é possível concluir que nem todos os sistemas de interpretação envolvem os módulos fonético-fonológico ou morfológico, por exemplo, por já receberem *inputs* na forma escrita ou por dispensarem a análise de morfemas, simplesmente separando as palavras por tokenização e acessando um léxico com as formas íntegras dos vocábulos. A composição do sistema dependerá, portanto, da compatibilidade dos recursos internos e dos objetivos aos quais a aplicação computacional se destina, se incluirá reconhecimento de fala, interpretação de sentenças ou compreensão de textos, entre outros aspectos. Nos sistemas de geração de línguas naturais, além da flexibilidade interna às três fases, pode haver trechos de texto pré-delineados — conhecidos como *canned texts* — dos quais os sistemas têm de apenas produzir material para preencher as lacunas, poupando o trabalho de geração em aplicações mais simples (DIAS-DA-SILVA *et al.*, 2007). Por outro lado, em aplicações mais sofisticadas, os sistemas podem envolver tanto a interpretação quanto a geração de línguas naturais, com a completa interação homem-máquina.

2.7 Aplicações

O conhecimento teórico e as técnicas do Processamento de Línguas Naturais possibilitaram o desenvolvimento de diversas aplicações computacionais hoje disponíveis no mercado, algumas das quais de uso cotidiano e popular. Seguem, abaixo, exemplos de aplicações de interpretação, de geração de línguas naturais ou de ambas, mais comumente

citados na literatura (JURAFSKY; MARTIN, 2008; MANNING *et al.*, 2008; RUSSELL; NORVIG, 2010).

a) Tradutores automáticos — Consistem em ferramentas para traduzir *inputs* linguísticos em formato escrito, oral ou visual (em imagens) sem a intervenção humana. Quando baseados em regras, podem seguir o modelo direto, em que as palavras são traduzidas uma a uma com o auxílio de um léxico bilíngue; o modelo de transferência, que formula uma árvore sintática da sentença na língua de partida, aplica regras de equivalência e desambiguação, transpõe o resultado para uma árvore sintática na língua-alvo e, finalmente, escreve a sentença; ou o modelo de interlíngua, em que uma representação abstrata do sentido é intermediária no processo de tradução. Se baseados em métodos estatísticos, os sistemas aprendem de forma supervisionada com *corpora* paralelos, nos quais um mesmo texto é reproduzido em duas ou mais línguas e as sentenças podem ser alinhadas para compreensão dos padrões de tradução. Um método comum nessa abordagem foi instituído pelos modelos de tradução da IBM (*IBM translation models*), que calculam a versão de tradução mais provável com base no teorema de Bayes. Exemplos da aplicação são o Google Tradutor e o Skype Translator.

b) Analisadores de sentimentos — São ferramentas que analisam o conteúdo emocional de sentenças e textos geralmente emitidos por usuários de mídias sociais. O objetivo desse tipo de aplicação pode ser averiguar a avaliação de consumidores sobre determinados produtos e marcas, pesquisar a opinião pública a respeito de programas governamentais ou fazer previsões como na eleição de candidatos a cargos políticos e sobre as tendências do mercado econômico para investimento em compra de ações. A estratégia dos analisadores é categorizar palavras como *like*, *disappointing* e *okay* por polaridade: positiva, negativa e neutra. Para tanto, utiliza recursos como um classificador Naïve Bayes, que, de um *corpus* de treino, extrai o vocabulário desejado, e máquinas de suporte vetorial, de aprendizagem supervisionada. Além disso, léxicos de sentimentos e tesauros como o WordNet, com sinônimos e antônimos, são usados para verificar a polaridade dos itens lexicais ou encontrar relações de similaridade (*strong* e *weak*, nesse sentido, teriam cargas opostas). Um exemplo de analisador é o Semantria, que pode ser aplicado ao Twitter, Facebook e outros *sites* de relacionamento.

c) Sumarizadores automáticos — Objetivam construir uma versão menor de um *input* escrito preservando os tópicos de maior interesse para o usuário. Ferramentas do tipo

podem ser classificadas segundo o material de entrada, a finalidade a que se propõem e o modo de produção. Segundo o material de entrada, podem ser sumarizadores de texto contínuo (*single-document summarizers*) ou de textos múltiplos (*multiple-document summarizers*), como é o caso de sumarizadores que fornecem resumos de notícias. Nestes, deve haver especial cuidado para evitar redundâncias, por tratarem textos de conteúdo semelhante. Pela finalidade, classificam-se em sumarizadores genéricos (*generic summarizers*), para simples resumo do material submetido, ou direcionados a perguntas (*query-focused summarizers*), que elaboram resumos conforme e especificamente para responder a questões do usuário, geralmente como parte de um sistema de perguntas e respostas. Já pelo modo de produção, os sumarizadores podem ser extrativos (*extractive summarizers*), quando criam os resumos reutilizando trechos do texto-fonte, ou abstrativos (*abstractive summarizers*), quando expressam as ideias com palavras diferentes. Exemplos dessa aplicação são os sistemas NeATS e LetSum.

d) Sistemas de recuperação de informações — São ferramentas de busca de materiais entre dados não estruturados, em geral, textos de um *corpus*. Por "não estruturados" entende-se dados em formato dificilmente acessível pelo computador, como é o caso dos documentos de *corpora*, cujo teor é mantido da forma como foi produzido por pessoas no processo de comunicação. Sistemas de recuperação são observáveis, por exemplo, na Web, com as ferramentas de busca on-line, ou em computadores pessoais, onde são usados para a localização de arquivos. Nas aplicações que utilizam o método booleano, a pesquisa é feita por meio de palavras-chave combinadas pelos operadores AND, OR e NOT. De forma semelhante aos números binários, os documentos são, então, classificados em dois extremos: interessantes, se incluírem a expressão exata enviada pelo usuário; ou desinteressantes, caso não a contenham. Em modelos de recuperação por *ranking*, por outro lado, como o modelo de espaço de vetores, a pesquisa dispensa o uso de operadores e o sistema gera uma lista seguindo a ordem de importância dos documentos, que podem ser mais ou menos interessantes. Em geral, quanto mais a palavra aparece no texto, mais importante ele é considerado. Nesse tipo de aplicação, ademais, algumas medidas são importantes: a precisão (*precision*) diz respeito à fração dos documentos encontrados que são realmente relevantes para o usuário; e a revocação (*recall*) é a fração dos documentos relevantes que chegam a ser encontrados. Ambas fazem parte da avaliação dos sistemas. Exemplos de ferramentas de recuperação na Web: Google Search, Yahoo! Busca e Bing.

e) Sistemas de extração de informações — Têm por objetivo construir bancos de conhecimento com as informações contidas em dados não estruturados ou semiestruturados em língua natural, de maneira a organizá-las para o trabalho humano e computacional. A extração de informações também é conhecida como análise de texto (*text analytics*) e mineração de dados textuais (*text mining*) no mercado, podendo, porém, ser estendida a áudios, imagens e vídeos em certos tipos de aplicação. O processo é baseado no reconhecimento de entidades nomeadas (*named entity recognition*), isto é, organizações, pessoas, eventos (ações ou acontecimentos), lugares, datas e quantidades específicas citados no texto, aos quais as informações estão associadas. Tal tarefa consiste, primeiro, na segmentação das palavras do *input* e, então, na busca e classificação de menções de entidades, com o uso de técnicas como os modelos ocultos de Markov, expressões regulares (de gramáticas do tipo 3) e características como dígitos e letras iniciais maiúsculas. A resolução de anáforas e outros tipos de referências também é realizada, já que nem sempre as entidades estão explícitas ao longo do texto. Além disso, é possível que o sistema compreenda a extração de relações (*relation extraction*), reconhecendo expressões como *located at*, *works for* e *was born in*, que conectam semanticamente as entidades nomeadas. A extração de informações é visível, por exemplo, no Google Agenda, que identifica datas e horários no Gmail e assinala atividades ao calendário do usuário automaticamente.

f) Sistemas de perguntas e respostas — São aplicações que visam a responder perguntas do usuário sobre determinados temas, mediante o acesso a um *corpus* ou à Web. As perguntas podem ser classificadas em dois tipos, para os quais os métodos de tratamento são diferentes: simples ou complexas. Perguntas do tipo simples (*simple/factoid questions*) são objetivas e pedem respostas sintéticas, geralmente constituídas por entidades nomeadas, como *How far is the Moon?*, cuja resposta é de cerca de 240 mil milhas de distância entre os centros da Terra e da Lua. É o tipo mais explorado pelas aplicações comerciais e pode ser processado por métodos baseados em recuperação, em conhecimento ou métodos híbridos. Segundo o método de recuperação de informações, o sistema começa extraindo dados da pergunta que indiquem a classe de entidade nomeada que a resposta deve conter (como distância, em *How far*); segue recuperando documentos previamente indexados no *corpus* e extraindo-lhes passagens relevantes; e finaliza convertendo as passagens em uma resposta que respeite a classe de entidade nomeada detectada na pergunta (*240,000 miles*). Segundo o método baseado em conhecimento, o

sistema realiza uma etapa de *parsing* da pergunta e constrói uma representação semântica do sentido apreendido; então, mapeia o resultado representativo a recursos estruturados como bases de dados e ontologias, que contenham informações suficientes sobre a entidade nomeada de que deve tratar a resposta. E em métodos híbridos, o sistema combina as duas formas de processamento, comumente utilizando recuperação de informações para encontrar possíveis respostas e técnicas baseadas em conhecimento para decidir a mais convincente. Perguntas do tipo complexas, por outro lado, são discursivas e exigem respostas longas o bastante para explicar fatos baseando-se em evidências. É o caso de *How is the Earth affected by meteorites?*, cuja resposta deve reunir diversas informações sobre os impactos geográficos e biológicos que meteoritos costumam causar à Terra. São perguntas tratadas, sobretudo, por sistemas de apoio à pesquisa científica, que utilizam técnicas de sumarização automática para coletar dados e agrupá-los em uma resposta concisa. Exemplos da aplicação: a ferramenta Wolfram|Alpha e a assistente pessoal Apple Siri.

g) Sistemas de diálogo — Também chamados "agentes conversacionais", são ferramentas de atendimento ao usuário capazes de manter comunicações fluidas através da fala. Geralmente são empregados em serviços de telefonia, nos quais os dispositivos não incluem teclados completos, e podem auxiliar, por exemplo, na aquisição de pacotes de viagens adaptados ou ao encaminhar ligações de consumidores aos departamentos corretos. Na configuração dos sistemas, o diálogo é tratado como uma atividade conjunta de dois ou mais participantes, com falas alternadas, silêncios e uma base de conhecimento comum aos interlocutores, que pode restringir-se a domínios específicos. A Pragmática se insere com aplicações da Teoria dos Atos de Fala e diversos aspectos discursivos são observados, como a coerência. Nesse contexto, duas fases constituem o ato linguístico: a apresentação, em que um interlocutor emite a mensagem executando um dos tipos de ato de fala (afirmação, pedido, enunciado de comprometimento, expressão de emoções ou declaração proclamativa); e a aceitação, em que o ouvinte dá sinal de que a mensagem foi recebida e efetivamente compreendida. Além disso, a estrutura do diálogo pode ser formulada em "pares adjacentes", como saudação-saudação, cumprimento-agradecimento, pedido-concessão. Para moldar essa estrutura, os sistemas utilizam, entre outros recursos computacionais, os *frames* e os autômatos finitos, modelo baseado em grafos e parâmetros de uma gramática regular. Exemplo: CMU Communicator, uma ferramenta de apoio ao planejamento de viagens.

3 IBM Watson

3.1 O desafio Jeopardy!

Em 2007, anos após ter vencido o campeão mundial de xadrez Garry Kasparov com o Deep Blue, um computador que rendeu técnicas inovadoras para a ciência, a International Business Machines (IBM) decidiu lançar-se em um novo desafio: construir um sistema inteligente o bastante para competir com humanos no Jeopardy!. No ar desde 1984, o programa de perguntas e respostas em conhecimentos gerais é um clássico da televisão nos Estados Unidos, conhecido como uma competição para pessoas inteligentes e rápidas. A ideia era, efetivamente, levar a Computação para novos horizontes. Se dominar as estratégias do xadrez pode parecer difícil, à primeira vista, lidar com a língua humana em sua profundidade é muito mais complicado, pelas ambiguidades e irregularidades já discutidas, praticamente opostas ao raciocínio lógico-matemático tanto do jogo quanto do computador. Para conseguir cumprir o desafio, portanto, a empresa teria que conceber uma nova metodologia em sistemas de perguntas e respostas, que pudesse munir toda uma máquina, sem limites de domínio, dos artifícios que faltavam às pequenas aplicações computacionais da época, sempre dirigidas a tarefas e contextos bem específicos.

O programa Jeopardy! é baseado em uma competição de três participantes, cuja proficiência intelectual é avaliada previamente em um teste de 50 questões. No cenário, a tela principal reproduz uma tabela com 30 pistas do jogo, divididas em 6 categorias e ocultadas por valores em dólar até que os participantes optem por selecioná-las. Além disso, cada competidor possui um botão que lhe dá o direito de resposta quando pressionado antes daqueles dos concorrentes, em momento devido. Ao responder corretamente, o participante ganha o valor que ocultava a pista; ao errar, o mesmo valor é subtraído do seu total. Há três rodadas: na primeira, o quadro contém, entre as opções, uma pista de *Daily Double*, que deve ser respondida apenas por quem a conseguir revelar e pode dobrar o total acumulado pelo participante, a menos que ele decida indicar um valor estrategicamente menor (até 5 dólares); na segunda, todos os valores do quadro são duplicados em relação à primeira rodada, e há dois *Daily Doubles*; e na terceira, chamada Final Jeopardy!, uma única questão, mais difícil, é respondida por todos, após indicarem para si um valor igual ou menor do que o já acumulado. As categorias são imprevisíveis e podem ser explícitas, como *history*, *science* e *politics*, ou implícitas, como *tutu much*, com pistas sobre balé. Há também categorias do tipo *puzzle*, em

que as pistas (*clues*) implicitamente se dividem em duas subpistas (*subclues*), com duas respostas que devem se combinar de modo específico para formar a resposta final. Na categoria *Before and After*, por exemplo, ambas as respostas se intersecionam. Na *Rhyme Time*, elas devem rimar uma com a outra. É o caso dos exemplos abaixo (FERRUCCI *et al.*, 2010):

> Category: *Before and After Goes to the Movies*
> Clue: *Film of a typical day in the life of the Beatles, which includes running from bloodthirsty zombie fans in a Romero classic.*
> Subclue 1: *Film of a typical day in the life of the Beatles.*
> Answer 1: *(A Hard Day's Night)*
> Subclue 2: *Running from bloodthirsty zombie fans in a Romero classic.*
> Answer 2: *(Night of the Living Dead)*
> Answer: *A Hard Day's Night of the Living Dead*

> Category: *Rhyme Time*
> Clue: *It's where Pele stores his ball.*
> Subclue 1: *Pele ball (soccer)*
> Subclue 2: *where store (cabinet, drawer, locker, and so on)*
> Answer: *soccer locker*

As respostas, porém, devem ser dadas em forma de pergunta. Sendo assim, para uma pista *This drug has been shown to relieve the symptoms of ADD with relatively few side effects.*, a resposta seria *What is Ritalin?* (FERRUCCI *et al.*, 2010). Não obstante a complexidade, os participantes têm apenas 5 segundos para responder, mas costumam fazê-lo em menos tempo, pela segurança que devem ter ao pressionar o botão. Não fosse assim, correriam o risco de errar a questão e perder dinheiro.

Para chegar a um sistema com as características de um campeão do Jeopardy!, a IBM começou com pesquisas sobre o estado da arte da área de aplicações do PLN em perguntas e respostas — *Question Answering* ou QA. O sistema PIQUANT (Pratical Intelligent Question Answering Technology) havia sido desenvolvido pela empresa para competir na Text REtrieval Conference (TREC) e foi avaliado entre os melhores do segmento entre os anos de 1999 e 2005, mas quando submetido a condições similares às do programa Jeopardy!, apresentou resultados ruins. Outro experimento foi feito com o OpenEphyra, um *framework* de código aberto para o desenvolvimento de aplicações em QA, mas, tendo sido elaborado pela Universidade Carnegie Mellon a partir do Ephyra System, também condicionado à TREC, apresentou desempenho aquém do esperado (FERRUCCI *et al*., 2010). Em uma iniciativa conjunta com a universidade, a empresa deu início, então, à iniciativa Open

Advancement of Question Answering (OAQA), com o intuito de engajar a comunidade científica para avançar mais rapidamente a área de QA. Entre os planos da iniciativa, estava a elaboração de uma metodologia universal para a avaliação dos sistemas de perguntas e respostas, não adstrita a competições públicas como a conferência TREC, tendo as discussões culminado em 8 métricas (FERRUCCI *et al.*, 2009):

1. Complexidade Linguística da Pergunta (*Query Language Difficulty*) — Nível de dificuldade para extrair o significado da questão, separando-a da estrutura linguística. É mais alto quando há ambiguidade e formações sintáticas fora da norma culta.
2. Complexidade Linguística do Conteúdo (*Content Language Difficulty*) — Nível de dificuldade ao acessar e interpretar os documentos de referência, como em dados não estruturados.
3. Complexidade da Questão (*Question Difficulty*) — Nível de profundidade das inferências necessárias para determinar e justificar a resposta a partir do conteúdo disponível.
4. Usabilidade (*Usability*) — Nível de interação com o usuário. É mais alto quando são oferecidas, por exemplo, opções de refinamento das perguntas, respostas alternativas e justificativas.
5. Confiança (*Confidence*) — Probabilidade de o sistema estar certo ao analisar uma resposta como correta. É mensurável em sistemas que incluem indicadores de confiança (*confidence scores*), baseados em fatores com a qualidade da fonte dos materiais de referência e a quantidade de traços em comum entre a pergunta e a resposta.
6. Adequação (*Accuracy*) — Porcentagem de questões para as quais o sistema gera um *ranking* de possíveis respostas e a resposta correta é exibida em primeiro lugar.
7. Rapidez (*Speed* ou *Response Time*) — Velocidade na elaboração de respostas às perguntas.
8. Domínios de Abrangência (*Broad Domain*) — Nível de abrangência das áreas de conhecimento cobertas pelo sistema.

Aplicadas aos parâmetros dos sistemas participantes da TREC e ao contexto de perguntas, respostas e competidores do Jeopardy!, essas oito métricas foram reunidas pela

empresa no gráfico abaixo, pelo qual pôde observar, com maior clareza, o quanto deveria avançar em tecnologias de *Question Answering* e em qual direção.

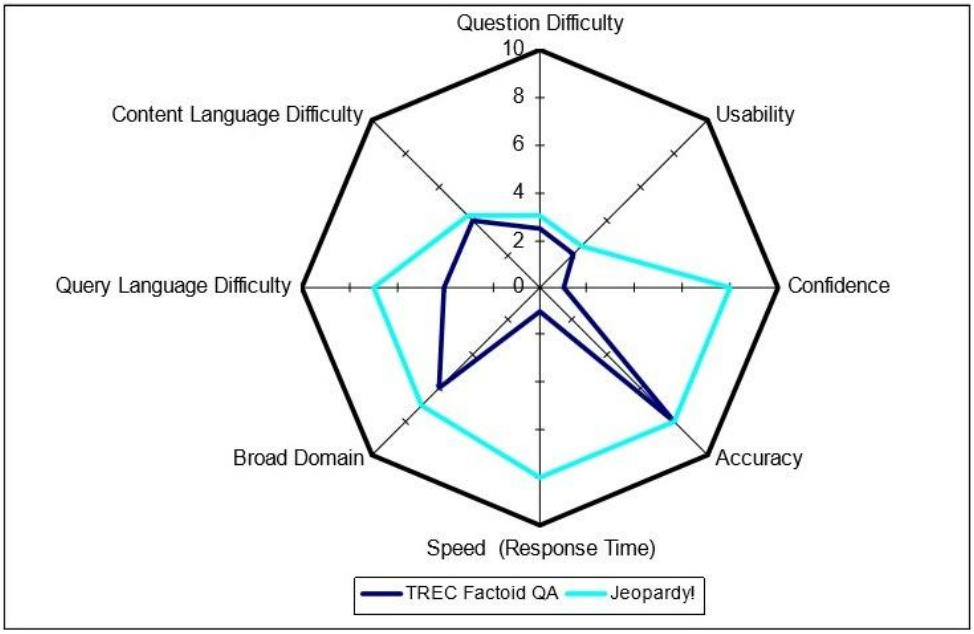


Figura 9 - Métricas da conferência TREC em sistemas de perguntas e respostas do tipo simples e do programa Jeopardy! segundo a iniciativa OAQA. Fonte: FERRUCCI *et al.* (2009).

O resultado dos avanços foi uma arquitetura de processamento profundo de línguas naturais chamada DeepQA, a qual seria a base de funcionamento do sistema competidor no Jeopardy!. Em apenas 3 segundos, tempo médio de resposta dos participantes, o sistema teria de efetuar uma análise sintática da pista, interpretar o que estivesse sendo perguntado, relacionar o significado às referências que "leu" previamente, determinar a melhor resposta e avaliar se estava suficientemente confiante para pressionar o botão. Tal processo seria realizado por componentes da arquitetura e recursos acoplados à DeepQA. Além disso, outros mecanismos seriam necessários para que o sistema, após aprovar seu nível de confiança, pressionasse o botão antes dos concorrentes, pronunciasse a resposta e, baseado no resultado anunciado pelo apresentador do programa, selecionasse uma nova pista ou passasse o controle para outro competidor (FERRUCCI, 2012).

Uma vez que aos jogadores não é permitido qualquer tipo de consulta para responder às questões, o sistema teria de internalizar as próprias referências — dicionários, enciclopédias e livros, por exemplo —, sem acesso à Internet. Para tanto e para suportar o volume de dados que é processado simultaneamente pelo sistema, um *hardware*[5] potente seria

[5] Parte física dos sistemas. Distingue-se dos *softwares*, parte lógica, da qual fazem parte os programas computacionais.

necessário, capaz de lidar com computação paralela massiva, e foi escolhido o processador POWER7, da própria IBM. Após a construção do supercomputador, ele foi submetido a treinos intensivos com o material de conteúdo e com simulações dos jogos do Jeopardy!, inclusive contra ex-participantes do programa.

Em 14 de janeiro de 2011, após quatro anos de esforços de uma equipe com cerca de 25 pesquisadores e engenheiros, o computador finalmente competiu contra dois dos melhores participantes da história do Jeopardy!: Ken Jennings, que havia permanecido por mais tempo no programa sem que fosse derrotado, e Brad Rutter, que havia acumulado a maior quantia em dinheiro. A máquina foi intitulada Watson, em homenagem a Thomas John Watson, fundador da IBM, e, pelo seu grande porte, o programa foi gravado no próprio laboratório da empresa em Nova York.

Além da arquitetura DeepQA, responsável pelo processamento de perguntas e respostas do sistema, dois componentes fizeram parte do Watson, um deles responsável pelas decisões estratégicas e o outro, pela interface do supercomputador com os participantes do jogo (FERRUCCI, 2012). O componente estratégico continha modelos de simulação dos adversários e do ambiente competitivo do programa, elaborados com o uso de técnicas de aprendizagem de máquina e dos métodos Monte Carlo, comumente aplicados a previsões estatísticas de computação paralela (TESAURO *et al.*, 2012). Dado o foco dos pesquisadores no processamento linguístico, o Watson não foi capacitado para ver ou ouvir, portanto, o componente de interface teve de buscar formas alternativas para acompanhar e interagir com o programa. Quando as categorias e pistas eram exibidas na tela, eram também enviadas eletronicamente para o Watson. O computador também monitorava sinais gerados quando o sistema de resposta era ativado para os competidores e quando qualquer deles pressionava o botão. Se o Watson estivesse confiante de sua resposta, ativava um dispositivo para pressionar o botão e, então, um mecanismo de síntese de fala para responder. Além disso, para saber se a resposta foi aceita como correta, ele verificava as variações de sua pontuação, enviadas pelo Jeopardy! (LEWIS, 2012).

O jogo foi televisionado apenas em fevereiro — do dia 14 ao dia 16 —, quando o supercomputador foi divulgado como campeão, tendo acumulado US$ 77.147 contra os US$ 24.000 de Jennings e os US$ 21.600 de Rutter. Ao total, a empresa recebeu o prêmio de 1 milhão de dólares, que doou integralmente para instituições de caridade.

3.2 Arquitetura DeepQA

A DeepQA é definida pelos seus criadores como uma *massively parallel probabilistic evidence-based architecture* (FERRUCCI, 2010). É, portanto, uma arquitetura baseada em evidências, na qual possíveis respostas são elaboradas e posteriormente fundamentadas em referências adicionais. É probabilística porque produz indicadores de confiança para todas as respostas e evidências, só então decidindo qual sentença corresponde melhor à pergunta de entrada. Além disso, é massivamente paralela, ao trabalhar com inúmeras possibilidades e processos ao mesmo tempo.

Para alcançar a eficiência que possui no tratamento de perguntas e respostas, a DeepQA foi implementada a partir da Unstructured Information Management Architecture (UIMA), uma arquitetura de software desenvolvida pela IBM e licenciada pela Fundação Apache como *framework* de código aberto, com a finalidade de "integrar diversas coleções de análises de texto, fala e imagem, independentemente da abordagem algorítmica, linguagem de programação ou modelo de domínio" (FERRUCCI, 2012, p. 2). Em síntese, a UIMA foi elaborada para facilitar o processamento de dados não estruturados e o faz associando-se a um conjunto de anotadores, que identificam e assinalam traços semânticos — e sintáticos, quando cabíveis — ao material. Pela estrutura escalável que oferece para aplicações de análise multimodal, tal plataforma foi fundamental para a rápida integração dos componentes da DeepQA, além de ter conferido flexibilidade e robustez ao sistema: a quantidade e a diversidade de componentes da arquitetura é tanta, que o Watson não depende estritamente de qualquer deles.

Ainda sobre a DeepQA, antes que o sistema de perguntas e respostas fosse disponibilizado para os treinos do Jeopardy!, foi necessário identificar e reunir conteúdo de referência para as respostas e respectivas evidências. Tal processo pode ser chamado Aquisição de Conteúdo (*Content Acquisition*) e envolve uma combinação de passos manuais e automáticos. O primeiro deles é elaborar uma descrição prévia dos tipos de perguntas que deverão ser respondidas, com uma análise manual de exemplos, e uma caracterização do domínio de aplicação, com análises estatísticas automáticas. Sendo o Jeopardy! uma competição de conhecimentos gerais, os domínios de aplicação são vários e uma estratégia utilizada foi a análise de *lexical answer types* (LATs), palavras das pistas que indicam o tipo de resposta que se deve obter. Na pista a seguir, por exemplo, o LAT é a palavra *maneuver*, indicando que a resposta deve ser o nome de uma manobra de xadrez (FERRUCCI, 2010).

Category: *Oooh… Chess*
Clue: *Invented in the 1500s to speed up the game, this maneuver involves two pieces of the same color.*

A descrição das perguntas e a caracterização dos amplos domínios do Jeopardy! definiram a composição do *corpus* de base, constituído por enciclopédias, dicionários, tesauros, notícias, obras literárias e outros tipos de referência do conhecimento popular e científico, em formatos preservados. Feito isso, o *corpus* foi expandido com informações encontradas na Web que fossem relacionadas aos textos originais, extraindo-se passagens relevantes do *corpus* de base e da Web e mesclando-as em documentos concisos sobre os mesmos assuntos. Além das referências não estruturadas, dados semiestruturados e estruturados também foram utilizados, incluindo bases de dados, taxonomias e ontologias, como o WordNet, a DBpedia e a YAGO (FERRUCCI, 2010). A empresa construiu, ademais, um recurso próprio de conhecimento estruturado, chamado PRISMATIC, que, utilizando *parsing*, extração de informações e técnicas estatísticas, deduz axiomas dos textos brutos do *corpus*. Tais axiomas são, por exemplo, *books are found on shelves*, *people visit museums*, *people visit websites* e *candidates win elections*, os quais também facilitam a geração de possíveis respostas pelo sistema (FERRUCCI, 2012).

Com um *corpus* amplo, variado, atualizado e bem articulado, o sistema pode prosseguir com a interpretação das perguntas e a geração das respostas, conforme a DeepQA. O processo pode ser organizado em quatro passos: Análise da Pergunta (*Question Analysis*); Geração de Hipóteses (*Hypothesis Generation*); Pontuação de Hipóteses e Evidências (*Hypothesis and Evidence Scoring*); e Compilação e Classificação Final (*Final Merging and Ranking*). Como se observa na figura a seguir, após a Análise da Pergunta, é possível que a pergunta de entrada seja decomposta em outras duas ou mais, cada uma das quais passando pelas fases de Geração de Hipóteses e Pontuação de Hipóteses e Evidências de modo paralelo, até que sejam sintetizadas antes da Compilação e Classificação Final. Além disso, da Geração de Hipóteses à Compilação e Classificação Final, há etapas intermediárias de processamento, bem como estão associadas referências não estruturadas e estruturadas e modelos de aprendizagem de máquina. O material de saída é uma resposta bem fundamentada, acompanhada de um indicador de confiança para maior segurança do usuário quanto à informação. Nos tópicos que seguem, analisa-se melhor cada um desses passos.

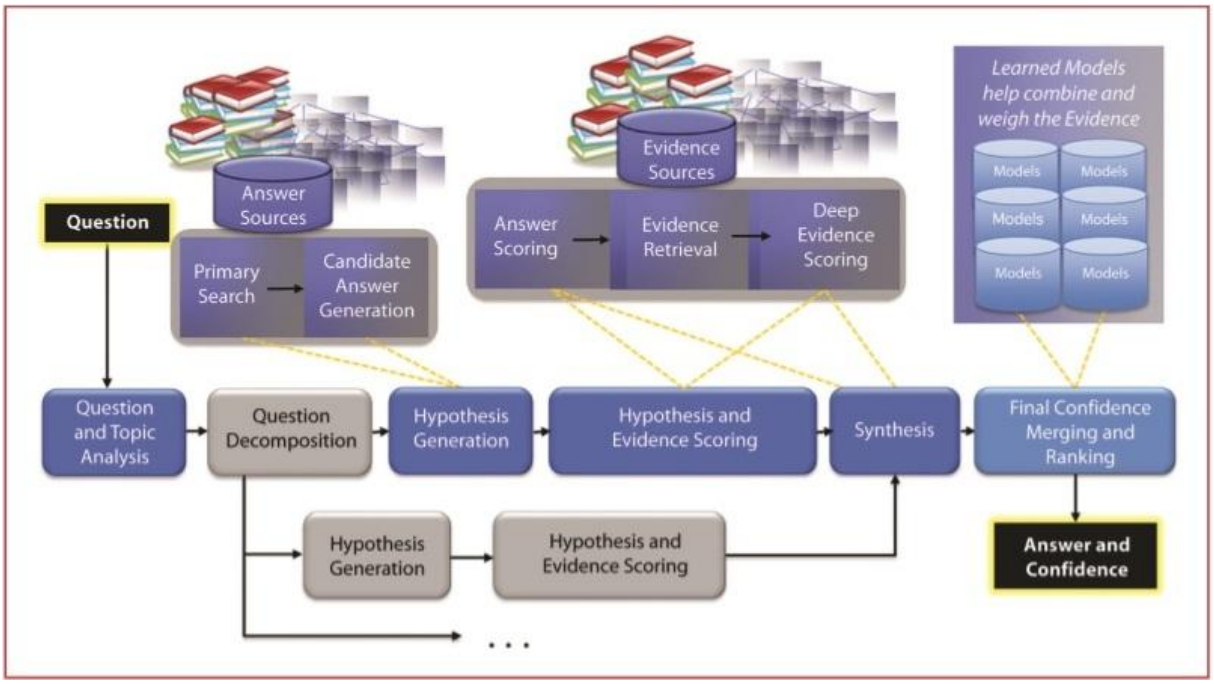


Figura 10 - Arquitetura DeepQA. Fonte: FERRUCCI (2012).

3.2.1 Análise da Pergunta

Nessa etapa, o objetivo do sistema é depreender a melhor interpretação possível da pergunta de entrada e determinar como ela deverá ser processada pelo resto do sistema. Para tanto, o sistema inclui um *parser* profundo da English Slot Grammar (ESG), associado a um construtor de estruturas predicado-argumento, um reconhecedor de entidades nomeadas, um componente de resolução de correferências e um componente de extração de relações (LALLY *et al.*, 2012). Nesse processo, inúmeras regras de detecção são aplicadas, para determinar, por exemplo, a classificação da pergunta, o foco, o LAT, as relações linguísticas presentes na sentença, e decidir se é o caso de decompor a pergunta para que receba um tratamento mais preciso.

A ESG é uma gramática de aplicações computacionais da língua inglesa cujo *parser* é capaz de analisar tanto traços sintáticos quanto morfológicos e semânticos em um mesmo processo. Tal gramática foi melhorada pela equipe especialmente para o desafio Jeopardy!, embora tenham tido o cuidado de manter a compatibilidade com o inglês geral, ao tentar adaptá-la para as questões do programa (LALLY *et al.*, 2012). Os principais passos do *parsing* são: tokenização e segmentação; análise léxico-morfológica; e análise sintática. Por compreender também o nível morfológico, o *parser* da ESG dispensa o uso de *POS taggers* e já inclui um léxico de base com aproximadamente 87 mil entradas, que se desdobram em

outras formas lexicais a partir da derivação, flexão e recombinação dos morfemas. No projeto, o léxico de base foi, ainda, expandido com o uso de recursos como o WordNet. Cada árvore sintática do *parser* é uma *dependency tree* que exibe duas dimensões de estrutura. A estrutura superficial é composta por *slots*, papeis sintáticos como sujeito e objeto. A estrutura profunda, por sua vez, diz respeito aos predicados e argumentos da sentença, representados semanticamente pela lógica. Em sentenças ambíguas, as árvores geradas são classificadas segundo um sistema de pontuação e apenas a melhor qualificada é levada adiante pelo sistema (MCCORD, 2010; MCCORD *et al*., 2012).

A ferramenta de construção de estruturas predicado-argumento (*predicate-argument structure*s, PAS) recebe o resultado do *parsing* e o simplifica, retirando termos pouco relevantes e alterando as estruturas frasais para formações mais simples. Alternâncias de voz ativa e passiva como as sentenças *John sold a fish* e *A fish was sold by John* teriam árvores sintáticas diferentes conforme a ESG, mas apresentariam a mesma estrutura após o construtor de PAS. Além das alterações lexicais e gramaticais que efetua, o PAS também reúne as dimensões superficial e profunda da árvore sintática, facilitando a aplicação das regras de detecção e classificação ligadas à análise da questão (MCCORD *et al.*, 2012).

Ainda nessa etapa da arquitetura, outras tarefas são realizadas para depreender da questão pontos e características que indiquem a aplicação de técnicas extras ou que determinarão como a resposta deverá ser construída. A classificação da pergunta (*question classification*) procura identificar tipos de questões e partes das questões que requerem um processamento especial. É o caso de perguntas do tipo *puzzle*, perguntas de cálculo ou mesmo palavras da sentença que sejam polissêmicas. A detecção de foco e LAT (*focus and LAT detection*) consiste no reconhecimento de trechos da pergunta diretamente relacionados com a resposta. LAT, já mencionado, é geralmente uma palavra ou sintagma nominal da pista que especifica o tipo e o domínio da resposta. Foco, por outro lado, é uma parte da questão que, quando substituída pela resposta, torna a questão uma sentença declarativa de sentido completo. Na pista *When hit by electrons, a phosphor gives off electromagnetic energy in this form*, por exemplo, o foco é a expressão *this form*. Substituída por *light* ou *photons*, resposta correta, a sentença ganha sentido. A detecção de relações (*relation detection*) busca compreender as relações sintáticas, como a ligação sujeito-verbo-objeto, e as relações semânticas entre as entidades da pergunta, também presente em outras etapas da DeepQA. A habilidade do Watson, no entanto, é limitada nessa tarefa, já que o sistema depende de bases de dados com a Freebase para ditar as relações que serão detectadas. Outra tarefa é a

decomposição (*decomposition*), que combina o *parsing* com métodos de classificação estatísticos para decidir como pistas do tipo *puzzle* e questões similares podem ser segmentadas (FERRUCCI, 2010). Tais tarefas são possíveis, também, graças à aplicação de regras de detecção, que os pesquisadores optaram por escrever em Prolog, pela simplicidade e expressividade dessa linguagem de programação (LALLY; FODOR, 2011).

3.2.2 Geração de Hipóteses

Após ter interpretado a pergunta e determinado a melhor maneira de respondê-la, o sistema consulta o *corpus* em busca de fontes de informações sobre o assunto em questão e, dos documentos encontrados, extrai fragmentos para construir sentenças candidatas à resposta. O objetivo do sistema, portanto, é gerar diversas possibilidades de resposta para só então, baseado em evidências, decidir qual delas é mais adequada à questão. Isso reduz as chances de erro do processo que, com a escolha de uma alternativa de resposta já nos passos iniciais, poderia equivocar-se pela falta de dados para embasamento. "Cada uma das candidatas à resposta, associada à pergunta de entrada, constitui uma hipótese, que o sistema prova estar correta gerando indicadores de confiança" (FERRUCCI, 2010, p. 70). Para construí-las, a etapa se desmembra em duas: a Busca Inicial, de acesso ao *corpus*, e a Geração de Candidatas à Resposta, propriamente dita.

A Busca Inicial (*Primary Search*) consiste em recuperar referências que contenham traços correspondentes aos identificados na Análise da Pergunta. Uma série de técnicas e recursos podem ser utilizados com esse fim: busca de documentos, com ferramentas como Indri e Lucene; busca de passagens de textos; buscas em bases de conhecimento, com a linguagem SPARQL; e buscas em texto livre com codificações que satisfaçam os critérios da pergunta. No caso das bases de conhecimento, as buscas dessa etapa consistem principalmente de entidades nomeadas, cujas informações podem ser facilmente encontradas em dados estruturados (FERRUCCI, 2010).

As técnicas da Geração de Candidatas à Resposta (*Candidate Answer Generation*), a seu turno, dependem dos resultados da Busca Inicial. Em resultados da busca de documentos que sejam orientados por títulos, os próprios títulos são extraídos como candidatos à resposta. Em resultados da busca de passagens, o reconhecimento de entidades nomeadas também é aplicado para distinguir, das palavras do texto, aquelas que são compatíveis com o *lexical answer type* da questão. Já nas buscas em bases de conhecimento e em definições de

dicionários, ambos os casos com conteúdo assinalado a expressões enxutas, os resultados constituem, por si só, possibilidades de resposta. Nessa etapa, é importante produzir o maior número possível de hipóteses — em geral, centenas —, pois, se a resposta correta não estiver entre as candidatas, o sistema será incapaz de acertar a questão (FERRUCCI, 2010).

3.2.3 Pontuação de Hipóteses e Evidências

Para decidir qual das hipóteses é a ideal, o passo seguinte da arquitetura é reunir evidências que atestem ou refutem cada uma delas, além de pontuá-las segundo o grau de correção. Com centenas de candidatas à resposta, no entanto, o processo de busca de evidências pode se tornar dispendioso e prejudicar a agilidade e a eficiência do sistema. À vista disso, antes que evidências sejam colhidas e as hipóteses, pontuadas, uma filtragem leve (*soft filtering*) é aplicada ao conjunto de hipóteses iniciais, com o objetivo de eliminar as candidatas mais facilmente identificáveis como incorretas. Nesse caso, algoritmos[6] de ação superficial podem computar, por exemplo, a chance de uma candidata à resposta ser um exemplo do LAT. Uma margem pré-determinada, baseada em aprendizagem de máquina com dados de treino, decide o destino das hipóteses: candidatas cujas chances superam a média, seguem na etapa de Pontuação de Hipóteses e Evidências; por outro lado, candidatas com chances abaixo da média são encaminhadas diretamente para a Compilação e Classificação Final, onde provavelmente receberão colocações mais baixas (FERRUCCI, 2010).

Feito isso, cerca de 100 hipóteses permanecem na etapa e carecem de comprovação. A recuperação de evidências (*evidence retrieval*) surge como uma tarefa necessária e utiliza uma série de técnicas para dar respaldo às candidatas restantes. Entre essas técnicas, uma particularmente efetiva é a busca de passagens da Geração de Hipóteses, à qual é enviada, porém, como termo de pesquisa, a palavra ou sentença candidata já produzida. “Isso irá recuperar passagens que contenham a candidata à resposta usada no contexto dos termos da pergunta original” (FERRUCCI, 2010, p. 72). A evidência encontrada é, então, direcionada aos marcadores de pontuação, que avaliam a candidata inserida nos contextos.

A tarefa de pontuação (*scoring*) concentra esforços massivos para uma análise profunda do conteúdo. O objetivo é associar às evidências indicadores de confiança sobre o grau em que elas atestam ou refutam a correção das candidatas e, para tanto, múltiplos marcadores trabalham em conjunto (FERRUCCI, 2012). O Watson emprega mais de 50

[6] Códigos de comando escritos em linguagem de programação.

componentes de pontuação, que envolvem desde probabilidades formais até contagens simples de traços em comum, baseando-se em evidências oriundas de fontes de dados estruturados, semiestruturados e não estruturados e considerando fatores como a compatibilidade das estruturas predicado-argumento da passagem e da pergunta, a confiabilidade da fonte das evidências, a localização geoespacial, as relações temporais e a classificação taxonômica dos itens lexicais. Sobre a localização, por exemplo, o sistema averigua as relações entre locais citados na questão e nas passagens das evidências para determinar se eles se tratam do mesmo território — como a relação de uma capital com o país. Nas relações temporais, igualmente, o sistema compara datas da pista com as do contexto que envolve a candidata à resposta, em busca de compatibilidade ou inconsistência (FERRUCCI, 2010). Todos esses fatores contribuem para a Pontuação de Hipóteses e Evidências e para a subsequente Compilação e Classificação Final.

3.2.4 Compilação e Classificação Final

O objetivo dessa etapa é agrupar as hipóteses geradas e pontuadas nas fases anteriores segundo o indicador de confiança de suas evidências para determinar uma resposta final. Antes de elaborar uma classificação, ou um *ranking*, das hipóteses, o sistema realiza uma tarefa de compilação das respostas possíveis (*answer merging*) para identificar as candidatas que tenham estruturas de frase ou formas lexicais diferentes, mas signifiquem a mesma coisa. Nesse caso, graças à aplicação de um conjunto de algoritmos de detecção de correspondências, resolução de correferências e normalização, as hipóteses são mescladas e os indicadores de confiança são combinados para apenas uma acepção (FERRUCCI, 2010).

Então, o *ranking* de possíveis respostas é produzido e a melhor colocada, junto ao indicador de confiança, é entregue ao usuário como material de saída para a pergunta submetida. No caso do Jeopardy!, durante o programa, o Watson exibia não só a primeira, mas as três mais bem classificadas aos telespectadores, que podiam acompanhar a importância dos indicadores para a seleção da melhor resposta. Tal classificação se deve a modelos de aprendizagem de máquina, treinados com exemplos de pistas com respostas conhecidas e capacitados para lidar com classes diferentes de indicadores, já que indicadores usados no levantamento da resposta correta de uma pergunta do tipo simples, por exemplo, podem não ser tão úteis a perguntas do tipo *puzzle* (FERRUCCI, 2010). O resultado é um sistema

flexível, robusto e de processamento profundo para a interpretação e geração de línguas naturais, claramente extensível a aplicações além do Jeopardy!.

3.3 Watson no mercado

O desafio Jeopardy! foi significante: ter construído uma máquina que venceu humanos em aspectos linguísticos, de conhecimento e estratégia representa um marco para a história da Computação. Todavia, os pesquisadores acreditavam que o sistema seria mais útil aplicado à resolução de problemas reais do cotidiano humano, se levado ao mercado para operar em diálogos com o usuário e auxiliá-lo na tomada de decisões a partir de grandes volumes de dados em formato natural (FERRUCCI, 2012). Eles chegaram à conclusão de que o Watson poderia ser chamado de cognitivo pelas características que tinha em comum com a mente humana, como a capacidade de criar e testar hipóteses, decompor e recompor construções linguísticas e extrair e avaliar informações relevantes, as quais, aliadas à potência de um sistema computacional, permitiriam resolver problemas de grande escala e com maior precisão e adequação.

A aplicação do Watson consistiria em três serviços cognitivos: com *Ask*, o usuário poderia enviar perguntas — ou pistas ou outros materiais textuais de base — e o sistema reuniria dados para providenciar-lhe uma resposta, ganhando utilidade, por exemplo, em diagnósticos de doenças, análises de crédito e pesquisas acadêmicas; com *Discover*, o sistema ajudaria o usuário a chegar a *insights* sobre um conteúdo, isto é, perceber aspectos e correlações nunca antes imaginados, ao realizar inferências adicionais automaticamente sobre o tema em foco; e com *Decide*, o usuário poderia tomar decisões mais fundamentadas sobre um problema, graças às diversas alternativas de solução que o sistema apresenta acompanhadas de indicadores de confiança (INTERNATIONAL BUSINESS MACHINES [IBM], 2012).

Para que a tecnologia do Watson pudesse chegar às empresas, o primeiro passo foi capacitá-lo para operar na nuvem, de modo que o simples acesso à Internet suprisse a necessidade de reproduzir *hardwares* pesados e onerosos para a comercialização do sistema. Além disso, sem a proibição de consulta a fontes externas que havia nas regras do Jeopardy!, o sistema passou a utilizar dados da Web em tempo real. Os produtos fornecidos hoje são: Watson for Clinical Trial Matching, Watson for Oncology, Watson Discovery Advisor

(também na versão Watson Discovery Advisor for Life Sciences), Watson Explorer, Watson Engagement Advisor, Watson Analytics e Watson Curator[7]. Em síntese, a maioria auxilia na otimização do trabalho administrativo, no gerenciamento e visualização de dados, no processo de inovação científica, na interação automática inteligente com consumidores, em previsões analíticas ou em atividades da assistência médica, especialmente em testes clínicos e em opções de tratamento de câncer. Há também o Watson Knowledge Studio, ainda em construção, que permitirá aos usuários customizar o Watson para a sua própria área ou empresa com aprendizagem supervisionada.

Além dos produtos prontos para uso, recentemente, a IBM disponibilizou uma série de interfaces de programação de aplicações (*application programming interfaces*, APIs) cognitivas, que possibilitam a desenvolvedores de *softwares* e aplicativos incorporar as habilidades do Watson a seus produtos. O serviço, chamado Watson Developer Cloud, utiliza a plataforma de desenvolvimento Bluemix para abrigar as APIs, que se ligam às criações dos programadores através da nuvem. As opções se classificam atualmente em quatro tipos: APIs de processamento de texto (Concept Expansion, Concept Insights, Dialog, Document Conversion, Language Translation, Natural Language Classifier, Personality Insights, Question and Answer, Relationship Extraction, Retrieve and Rank e Tone Analyzer); APIs de processamento de fala (Speech to Text e Text to Speech); a API de análise de dados Tradeoff Analytics; e APIs de análise de imagens (Visual Insights e Visual Recognition), com a mais nova capacidade do Watson de visão[8]. Entre as opções, há também APIs da empresa AlchemyAPI (AlchemyLanguage, AlchemyVision e AlchemyData News), que a IBM adquiriu em 2015. Essas têm o diferencial de associarem as técnicas de PLN a conhecimentos de aprendizagem profunda (*deep learning*), tecnologia de ponta que aplica aprendizagem de máquina e redes neurais artificiais para simular as competências da mente.

A revolução que o Watson provocou no modo de funcionamento computacional — e, consequentemente, nos serviços de Tecnologia da Informação — foi tamanha, que a empresa afirma o início de uma Era da Computação Cognitiva, a terceira da Computação, após a Era dos Sistemas de Tabulação e Cálculo e a Era dos Sistemas Programáveis (GANDOUR, 2014). Para que um público cada vez maior usufrua dessas inovações, os próximos desafios da IBM são aplicar o Watson a outras áreas de conhecimento, levá-lo a outros povos e fazê-lo processar novas línguas. Entre essas línguas, está o português do Brasil.

[7] http://www.ibm.com/smarterplanet/us/en/ibmwatson/offerings.html
[8] http://www.ibm.com/smarterplanet/us/en/ibmwatson/developercloud/services-catalog.html

4 Watson em português brasileiro

4.1 Internacionalização e localização do sistema

Em se tratando da exportação e importação de produtos, simplesmente traduzir o conteúdo para a língua do destino não basta. Aspectos culturais, financeiros e legais do local podem ser completamente diferentes daqueles aos quais o produto se submeteu durante a produção, e tais divergências, se não observadas, podem determinar o fracasso da recepção por parte dos consumidores. Em *softwares*, *websites* e sistemas computacionais, por exemplo, o tamanho dos ícones pode variar segundo o costume, a direção do texto e a localização do menu dependem da língua, alguns símbolos e cores devem ser evitados por estarem ligados a religiões e crenças, os formatos de data e hora devem seguir os padrões sociais e mesmo o calendário adotado pode ser diferente. Nesses casos, são aplicadas as técnicas de internacionalização e localização, nas quais a tradução é apenas mais um dos passos.

Conforme Prudêncio *et al.* (2004), a internacionalização é uma etapa do processo de desenvolvimento de *softwares* — e outros produtos — que visa a torná-lo culturalmente neutro. É conhecida como "i18n" no jargão da área, simbolizando as 18 letras que há entre o "i" e o "n" de *internationalization*. Já a localização,"l10n", consiste na adaptação dos produtos a determinados países e culturas e, em regra, sucede a etapa de internacionalização. Os alvos da localização são chamados *locales*, representados por pares língua-região no setor de *software*, como português do Brasil (pt-BR) e inglês dos Estados Unidos (en-US).

O IBM Watson™ não foi inicialmente idealizado para a aplicação em outros países, exteriores aos Estados Unidos, por isso, não passou por uma etapa de internacionalização quando da elaboração da DeepQA. O objetivo, à época, era cumprir o desafio Jeopardy!, o que exigia conhecimento profundo e específico da língua inglesa na variedade americana. Com o advento da Era da Computação Cognitiva, todavia, surgiu com ela o interesse de levar o Watson a pessoas de todo o mundo, para tanto, realizando-se projetos de internacionalização e localização do sistema. A seguir, são abordados o Watson Multilíngue e projetos paralelos de adaptação do sistema a outras línguas, os quais excedem as práticas tradicionais de internacionalização e localização de *software*, por interfaces e códigos-fonte, e fundamentam-se essencialmente nos recursos e ferramentas internas de tratamento linguístico, por se tratar de um sistema altamente especializado em Processamento Automático de Línguas Naturais.

4.1.1 Watson Multilíngue

A internacionalização do Watson é tarefa do projeto Watson Multilíngue, cuja proposta é chegar a um sistema capaz de processar quaisquer ou a maioria dos idiomas utilizados internacionalmente. O desafio reside em adaptar, sobretudo, os componentes de análise morfossintática e as ferramentas de detecção de entidades e relações semânticas da DeepQA, profundamente dependentes da gramática e do léxico da língua inglesa. De forma a neutralizar linguisticamente o sistema, Cortis *et al.* (2014) elaboraram adaptações e construíram um protótipo do Watson Multilíngue, incluindo um método robusto para detecção de características em questões (como o LAT) em alternativa às regras de detecção manuais feitas para o inglês e um mecanismo para detectar entidades nomeadas iguais em textos de línguas diferentes. O resultado foi apresentado na 25ª Conferência Internacional de Linguística Computacional (COLING), na Irlanda, e demonstrou eficiência em línguas como o inglês, espanhol, francês e português brasileiro, abrindo espaço para avanços no mesmo sentido.

A modificação mais elementar diz respeito ao *parser* da English Slot Grammar, responsável por analisar as questões que o Watson recebe e encaminhá-las a outros componentes. O formalismo XSG por trás do *parser* da ESG comporta a geração de gramáticas e regras de atribuições sintáticas para línguas além do inglês, mas a atividade demandaria um grande esforço de linguistas habilidosos, que deveria ser contínuo à medida que o Watson fosse aplicado em novos domínios, quando as regras teriam de ser revisadas e estendidas. Uma alternativa estudada pela equipe foi elaborar um *parser* estatístico com as mesmas qualidades de adequação do *parser* baseado em regras da ESG, mas com maior facilidade de extensão a outras línguas. Para tanto, seria utilizado o *dependency parsing* estatístico proposto por McDonald *et al.* (2013) e formas de representação sintática não dependentes da língua, no caso, um conjunto harmonizado de rótulos de dependência para *parsing* multilíngue (MCDONALD *et al*., 2013 *apud* CORTIS *et al*., 2014). *Treebanks* com etiquetas morfológicas e rótulos de dependência compatíveis seriam necessários para treinar o *parser* e, apesar de haver vários desses recursos em diversas línguas, seria preciso adaptá-los às tarefas de QA. Outra alternativa considerada foi elaborar um subsistema independente do próprio *parsing*, que se baseasse, por exemplo, nos traços morfológicos de *lexical answer types*.

Conforme FERRUCCI *et al.* (2010 *apud* CORTIS *et al*., 2014), detectar corretamente o LAT, aliás, é importante para que o sistema produza uma resposta adequada, além de oferecer-lhe insumos para a pontuação de candidatas e para a reparação de erros. O foco da pergunta, o qual a resposta terá de substituir para constituir sentenças convincentes, é igualmente relevante. No sistema original, um componente em Prolog era responsável por identificar ambos, LAT e foco, mas suas regras de detecção eram específicas para construções na língua inglesa e o Watson Multilíngue requer um método robusto. Desse modo, os pesquisadores optaram por utilizar, no protótipo, a ferramenta de construção de regras IBM LanguageWare, cuja expressividade é similar à do Prolog e que demonstrou ser 4 vezes mais rápido que o componente. Além disso, um método estatístico que foi usado no Jeopardy! também estaria sendo adaptado para o contexto multilíngue.

Para a detecção de entidades nomeadas e a identificação de conceitos equivalentes em textos de múltiplas línguas, seria preciso ligá-los a identificadores linguisticamente neutros. Segundo Cortis *et al.* (2014), a Wikipédia e o Wikcionário disponibilizam traduções de palavras de uma língua para outra, mas os identificadores de seus conceitos são dependentes dos idiomas. A solução, então, foi utilizar a versão estendida do Open Multilingual Wordnet (OMW) que vincula o WordNet a dados do Wikicionário por meio de identificadores neutros. Os pesquisadores também se inspiraram nos rótulos alfanuméricos do Unified Medical Language System, incorporando, ademais, as recomendações de Web Semântica da World Wide Web Consortium (W3C).

Nos testes do protótipo, eles transformaram textos da Wikipédia, originalmente em formato XML, para os padrões da TREC e, então, para os índices de pesquisa do Lucene. A DeepQA admite apenas caracteres em ASCII, retirando acentos e outros símbolos durante a normalização dos textos. O protótipo, porém, foi capacitado para utilizar o Unicode na versão Normalization Form Compatibility Composition (NFKC), possibilitando um tratamento eficiente de línguas além do inglês. Processados os textos-fonte da Wikipédia, os testes valeram-se de perguntas e respostas conhecidas, em inglês, automaticamente traduzidas pelo serviço n.Fluent Translation da IBM. Algumas respostas previsivelmente não seriam respondidas — por falta de artigos correspondentes em determinada língua, por exemplo — e, para identificá-las, foi utilizada a MediaWiki API, uma API da Wikipédia, que, além disso, auxiliou no direcionamento correto das perguntas aos artigos. De resto, o protótipo contou com revisões manuais das traduções e para checar se as perguntas eram realmente passíveis de resposta.

4.1.2 Watson em outros idiomas

Atualmente, há projetos de localização do Watson para vários idiomas, sendo os principais o japonês, o espanhol e o português brasileiro, em razão de parcerias estratégicas. A localização do sistema para o japonês é fruto de uma parceria da IBM com o SoftBank, uma empresa de telecomunicações com um amplo histórico de desenvolvimento em tecnologias no Japão. O objetivo da parceria é levar a nova era da Computação ao país, implantando um capítulo local do IBM Watson Ecosystem para que empreendedores, desenvolvedores de *software* e acadêmicos japoneses possam inovar com a tecnologia cognitiva, a começar pela educação, saúde, serviços financeiros e comércios de varejo, nos quais o uso da informação é mais importante para a entrega de produtos e serviços de qualidade. As aplicações serão destinadas, sobretudo, a *smartphones*, *tablets* e robôs, dentre os quais, o Pepper, da empresa, primeiro robô pessoal a reconhecer emoções (IBM, 2015a).

A língua japonesa representa um desafio à parte para o Watson, pelo conjunto de ideogramas Kanji. A leitura dos caracteres, na verdade, é complicada para qualquer sistema computacional, por não seguir um padrão morfológico e por não haver espaços entre uma palavra e outra, exigindo técnicas especiais de demarcação dos itens lexicais. Além disso, as regras do discurso são mais obscuras que as do inglês, com inúmeras expressões idiomáticas e um certo grau de cortesia que requer interpretações e pronúncias diferentes a depender do contexto do diálogo, do gênero e idade dos interlocutores e das relações que eles mantêm entre si (IBM, 2015a; IBM, [2015]b). Apesar das dificuldades, a localização do sistema permanece ativa e o processo tem se baseado nos seguintes passos (IBM, [2015]b):

1. O Watson assimila uma sequência de 250 mil palavras, as quais transforma em 10 mil sentenças diagramadas. Ele identifica sujeitos, verbos e objetos, assim como o contexto que é necessário para entender japonês.
2. Falantes nativos de japonês corrigem as falhas do Watson, possibilitando-lhe aprender com próprios seus erros.
3. Então, com uma compreensão mais avançada do material, o Watson assimila uma segunda sequência de 250 mil itens lexicais e produz mais 10 mil sentenças diagramadas.
4. Quanto mais o Watson pratica, analisando palavras sintaticamente e diagramando sentenças, melhor ele fica em categorizá-las.
5. Após o quarto estágio de diagramação de sentenças e de correções, o Watson terá compilado integralmente 1 milhão de palavras. Essa base de conhecimento permite ao Watson dar início a um processo mais intensivo de compreensão da língua como um todo para, finalmente, responder perguntas em japonês.

Os diagramas indicam a estrutura sintática e semântica do conteúdo, a partir da desconstrução das sentenças. Nos termos do PLN, tratam-se de árvores sintáticas e

representações do significado. Seguindo a mesma lógica, ao final do processo, o Watson terá construído um *treebank* ou *corpus* de 1 milhão de palavras. A parte mais difícil e longa da localização, porém, será o processo de compreensão dos conceitos e a subsequente aplicação dos conhecimentos gramaticais e extralinguísticos, os quais ainda constituem enigmas (IBM, [2015]b).

As localizações para o espanhol e o português brasileiro, por sua vez, começaram com parcerias entre a IBM e instituições interessadas em aplicar o Watson a serviços de atendimento ao consumidor. Para o espanhol, o trabalho está sendo feito junto ao Centro de Inovação Digital do CaixaBank, um banco da Espanha já reconhecido por inovar em seus processos com o uso de tecnologias (IBM, 2014). Para o português do Brasil, o trabalho começou com o Bradesco, em outubro de 2014, para o qual a IBM espera entregar serviços cognitivos de atendimento telefônico a partir do segundo semestre de 2016. Graças à parceria, a IBM decidiu, inclusive, priorizar o português ante o espanhol. Todavia, para que os serviços do banco atendam a todo o país, será necessário mais tempo até que ele se adapte aos regionalismos (MATSU, 2015). Além do Bradesco, a empresa Ixia de soluções de atendimento ao público anunciou uma parceria com a IBM para aplicar as habilidades do Watson a *call centers* e outros canais de relacionamento, por meio dos quais será possível analisar dados de comportamentos dos consumidores de forma avançada (COMPUTERWORLD, 2015).

Paralelamente aos projetos de localização do sistema, as APIs do Watson são constantemente atualizadas com extensões para novos idiomas. A AlchemyLanguage é um conjunto de 12 APIs, metade das quais admitindo *inputs* em inglês, francês, alemão, italiano, português, russo e espanhol, além da Language Detection API, capaz de reconhecer 97 línguas. No caso das APIs desenvolvidas pela IBM, quando incluído o português — Speech to Text API e Tradeoff Analytics API —, a variedade adotada é a brasileira. Outras APIs podem ser aplicadas a outros idiomas mesmo quando não houver localização, graças à possibilidade de conectá-las à Language Detection API, já mencionada, e à Language Translation API, capaz de traduzir 62 línguas.

4.2 Particularidades do português brasileiro

Para que o Watson possa interpretar e gerar conteúdo em português brasileiro, algumas particularidades do idioma e da variedade devem ser observadas. Tais particularidades, visíveis quando se compara a língua ao inglês americano e ao português europeu, vão desde aspectos gerais da fala e da escrita próprias do Brasil até os regionalismos já mencionados pela IBM. O sistema deve ser capaz, enfim, de lidar com a língua em sua ampla diversidade de formas, da norma culta ao português coloquial, de quaisquer ou da maioria das regiões do país, para que o atendimento direto ao público, a que o Watson é frequentemente destinado, seja eficiente e inclusivo. As seções a seguir citam algumas dessas particularidades do português do Brasil (PB) sob a ótica da gramática descritiva, as quais constituem potenciais desafios para a localização do sistema.

4.2.1 Sujeito

Conforme Branco *et al.* (2012, p. 10), "o português é uma língua que permite sujeitos nulos, [...] [o que] representa um desafio acrescido para a análise sintática automática dos textos e da fala". Tanto o sujeito oculto ou elíptico quanto o indeterminado não aparecem de forma explícita na sentença, dificultando o trabalho de *parsing*. Além disso, o sujeito pode ser deslocado da posição inicial da oração ou pode não concordar com o verbo que o acompanha, seja pela flexibilidade da norma gramatical culta, seja pela autonomia da fala cotidiana. Nos exemplos abaixo, de Dias da Silva *et al.* (2007), observa-se casos de sujeito elíptico (11, 19 e 21), indeterminado (7), anteposto ao verbo (2, 4 e 9) ou que não concordem com ele (1, 6, 8, 9 e 17), em nuances de difícil percepção para uma ferramenta computacional.

(1) A alegria e o contentamento era enorme.
(2) Aconteceu um acidente terrível na estrada.
(3) Alguém sempre sai ganhando.
(4) Vendem-se casas.
(5) Ele desapareceu.
(6) O pessoal foram no cinema.
(7) Ø dizem que a inflação vai voltar.
(8) Flores não tem acento.
(9) Falta dois dias pra acabar o ano.
(10) Fumar provoca câncer.

(11) Ø comprei um carro novo.

(12) Mais de um deputado votou contra a proposta.

(13) Mateus, Marcos, João e Lucas foram apóstolos de Jesus Cristo.

(14) O príncipe dos sociólogos virou presidente.

(15) O quiabo desapareceu dos supermercados.

(16) O menino que vimos ontem passeando na rua quando estávamos a caminho do teatro desapareceu.

(17) Os Lusíadas é um livro de Luís de Camões.

(18) Walter Benjamin se matou.

(19) Paulo saiu de casa e Ø desapareceu.

(20) Sair de casa, em São Paulo, à tarde, durante o mês de março, quando o céu está cinzento, é pedir para ficar preso na chuva.

(21) Não faça Ø isso, Maria!

Sobre o sujeito nulo, especificamente, Mário A. Perini (2010) esclarece que a maioria dos verbos do português brasileiro permite a omissão do sujeito, mas que há verbos que raramente ocorrem com ele. São os casos de verbos com significado de existência (como em "Teve dois acidentes na minha rua"); verbos que indicam fenômenos da natureza (como em "Trovejou muito ontem de noite"); os verbos "ser" e "estar" com complementos de tempo ou estado meteorológico (como em "Já é tarde"); o verbo "ir" acompanhado de "para" (como em "Vai para cinco anos que eu moro aqui"); e o verbo "fazer" na construção fazer + expressão de tempo + expressão de tempo + que + oração (p. ex., "Faz sete anos que não vejo a minha irmã"). "Ter" com sentido de existência, aliás, é específico do português brasileiro, segundo Ataliba T. de Castilho (2010), pelo que os europeus fazem uso apenas de "haver".

4.2.2 Pronomes

O português brasileiro apresenta particularidades também em relação aos pronomes. Os pronomes pessoais do caso reto — "eu", "tu", "ele", "nós", "vós" e "eles" — são substituídos pelas formas "eu", "você", "ele", "nós", "vocês" e "eles", na fala, havendo uma tendência progressiva pela permutação de "nós" por "a gente" (CASTILHO, 2010). "Vós" nem mesmo ocorre na escrita, exceto em contextos religiosos específicos, segundo Perini (2010). Os pronomes demonstrativos são reduzidos às formas "esse" e "aquele" e respectivas flexões, "perdendo-se a distinção lexicalmente marcada entre a primeira e a segunda pessoa"

de "este" e "esse" (CASTILHO, 2010, p. 207). Além disso, há uma generalização do pronome relativo "que", em detrimento de "cujo" e "onde", e um desaparecimento do pronome possessivo "seu" em referência à terceira pessoa.

Os pronomes clíticos são um desafio à parte para o processamento automático da língua portuguesa, conforme Branco *et al.* (2012), pela complexidade das posições que podem ocupar na frase. Segundo ele (p. 11):

> Os pronomes clíticos podem ocorrer antes ou depois do verbo, exceto nos tempos futuro e condicional, em que podem ocorrer antes ou no meio da forma verbal (*dar-lho-ei*). A presença de um clítico de terceira pessoa no meio ou após o verbo pode afetar a forma do próprio verbo. Por exemplo, na sequência final *-ar*, o *-r* cai e a vogal é acentuada (*dá-lo-ei*).

No português brasileiro, em particular, são preferíveis as posições de próclise e ênclise na escrita, em que o pronome se insere antes ou depois do verbo principal da oração. Na fala, observa-se apenas a próclise, mesmo quando a norma culta pediria uma ordem diferente. É o caso de "Me passa o bife", em contradição à regra de que pronomes clíticos ou átonos não podem iniciar sentenças (CASTILHO, 2010). A mesóclise, posição na qual o sujeito se insere dentro do verbo — a exemplo de "dar-lho-ei" e "dá-lo-ei" — é pouco praticada no PB e, em geral, até evitada, por caracterizar-se como preciosismo.

Quanto à forma lexical desses pronomes, verifica-se o uso de "me", "te", "nos" e "se" na fala brasileira como um todo. Devido à ausência de "vós", o clítico "vos" também é raramente utilizado e há uma tendência de desaparecimento do pronome acusativo "o", correspondente a "ele", como em "Ainda não vi ∅ hoje" (CASTILHO, 2010). Em vez disso, os falantes suprimem a referência pronominal ou utilizam o pronome "ele" (e flexões) na forma reta. São exemplos as sentenças: "Eu chamei ele para ajudar na cozinha" e "De repente eu vi eles chegando de táxi" (PERINI, 2010). Fato semelhante ocorre com o pronome "você", que pode tanto ser remetido à forma "te", quanto permanecer reto no objeto. A esse respeito, Perini esclarece que as sentenças "Eu queria te levar no concerto" e "Eu queria levar você no concerto" são sinônimas. O plural "vocês", por outro lado", não tem clítico correspondente: "Eu queria levar vocês no concerto".

4.2.3 Paradigma flexional

De acordo com Branco *et al.* (2012, p. 10), "o paradigma flexional do português é muito mais rico que o de línguas como o inglês, em particular no que diz respeito aos verbos". Um mesmo verbo pode variar conforme o "aspecto, tempo, modo, pessoa, número, gênero ou polaridade, atingindo mais de 160 formas flexionadas diferentes, incluindo as simples e compostas". Dois paradigmas de flexão verbal nem mesmo existem em outras línguas românicas, a saber, o infinitivo flexionado (como em "para eles irem") e o futuro do conjuntivo, cujas flexões podem ser idênticas ao infinitivo não flexionado (como em "quando eu acordar"), provocando ambiguidades. Além dos verbos, os substantivos e os numerais do português podem variar em gênero e número (como "cachorro" e "cachorras"; "primeiro" e "primeiras") e os adjetivos, em gênero, número e grau (como "divertido", "divertidas" e "divertidíssimas").

Por outro lado, a simplificação da concordância verbal e nominal é um traço praticamente universal da fala do PB (PERINI, 2010) e deve estar incluída na descrição dos fenômenos linguísticos com vistas à aplicação computacional. Porque os pronomes pessoais são modificados nessa variedade, a morfologia verbal é reduzida a quatro formas: "falo", "fala", "falamos" e "falam", por exemplo (CASTILHO, 2010). Além disso, os verbos do futuro são geralmente conjugados na forma composta (como "vou falar") e a concordância com o sujeito, como se viu, pode não existir, quando os verbos são reduzidos à forma singular (como em "se eles pudesse ajudar") (CASTILHO, 2010; PERINI, 2010).

A concordância nominal é simplificada com a perda de marcas do plural repetitivas. Em "as aluna" (PERINI, 2010), por exemplo, observa-se o plural marcado apenas no artigo, que determina a pluralização de todo o sintagma. Conforme os estudos empíricos de Campos e Rodrigues (2002), tal variação pode ser observada tanto no interior do sintagma nominal quanto em elementos externos, o que corrobora a afirmação de que as ambas as concordâncias nominais e verbais podem ser simplificadas. São exemplos: "tem mil e um curso" e "as refeições são muito pesada". Ademais, segundo as autoras, a primeira palavra do sintagma favorece a presença de marcas do plural, sendo ela núcleo ou determinante.

4.2.4 Oração

A ordem básica dos termos da oração na língua portuguesa é semelhante à da língua inglesa: sujeito-verbo-objeto. Difere do japonês, por exemplo, cuja ordem sintática é: sujeito-objeto-verbo. Sem embargo, quando incluídos modificadores frasais, é possível notar inversões internas à ordem de ambas. Adjetivos vêm antes de substantivos e sintagmas nominais em inglês e geralmente depois deles em português. *Blue flowers*, por exemplo, seria traduzido por "flores azuis".

Perguntas, de igual modo, são formuladas com verbos modais ao início, em inglês (p. ex., *Can I help you?*), enquanto em português elas diferem de afirmações apenas pelo sinal de interrogação, na escrita, e pela entonação, na fala. Isso em orações interrogativas fechadas — as chamadas *yes-no questions* —, cuja resposta pode ser apenas "sim" ou "não" (PERINI, 2010). Nas interrogativas abertas ou *wh-questions*, por outro lado, a formação assemelha-se pela presença de pronomes interrogativos na sentença: "que", "o que", "qual", "onde", "aonde", "quem", "quantos", "como", "por que", "cadê", em português; *which*, *who*, *whom*, *when*, *where*, *why*, *what* (*for*), *how* (*old*/*much*/*many*/*long*/*far*), em inglês.

No português brasileiro, especificamente, Perini (2010, p. 108) esclarece que "a ordem dos termos na oração pode ser decorrente de pelo menos dois fatores: a distribuição diferente de papéis temáticos e a topicalização de certos constituintes". Os papéis temáticos são os já conhecidos sujeito (agente) + verbo + objeto (paciente). A topicalização, por sua vez, ocorre quando um elemento é deslocado para o início da oração, por ser considerado o tópico da mensagem. É o caso da sentença "Em Bagé, ninguém come pequi" (p. 108), em que o adjunto adverbial de lugar é enfatizado pelo emissor. Bechara (2009) chama o fenômeno de antecipação ou prolepse, que pode gerar, ainda, o fenômeno do anacoluto. No trecho "Quem quer que disser mal de D. Henrique, eu me matarei com ele" (BARROS, 1777 *apud* BECHARA, 2009), por exemplo, o anacoluto ocorre com a quebra da estruturação lógica da oração, segundo Bechara, uma anomalia que resulta do descompasso entre linguagem e pensamento.

4.2.5 Fonologia

A fonologia do português brasileiro difere do falar peculiar ao português europeu, sobretudo, nas vogais. Vogais tônicas são aquelas que levam o acento de intensidade da

palavra, a exemplo do "i" em "vestido", enquanto vogais pré-tônicas são mais fracas e antecedem as acentuadas, a exemplo do "e" no mesmo item lexical. Segundo Castilho (2010), no português de Portugal há mais vogais tônicas e pré-tônicas do que a variedade brasileira, mas todas as vogais do PB são pronunciadas. Isso permite que "de frente" e "diferente" não se confundam pela pronúncia brasileira, ao contrário do que se observa em Portugal.

Em caso de ditongos, formados pelo encontro de vogais com semivogais, alguns monotongam-se em apenas um som no PB, como "terreiro", frequentemente pronunciado como "terrêru" (Castilho, 2010). Bagno (2006) explica que essa monotongação do ditongo "ei" ocorre sempre diante das consoantes "j", "x" e "r", como em "bêjo" (beijo), "dêxa" (deixa) e "primêro" (primeiro), e que se constitui como resultado do fenômeno da assimilação, assim como os ditongos "ou" de "frouxo" ("frôxo") e "ai" de "baixinha" ("baxinha") e certos encontros consonantais, como o "nd" de "falando", que se torna "falano". Algumas vogais, ao contrário, ditongam-se em dois fonemas quando aparecem ao final da palavra seguidas de sibilante ("s" ou "z"), a exemplo do que ocorre com "luz" ['luys] e "atrás" [a'trays] (Castilho, 2010).

Outras modificações são possíveis no exterior das palavras, como o "conjunto de fenômenos fonológicos que ocorrem na junção" delas, ao que Perini (2010, p. 354) denomina "sândi". Collischonn (2005a) cita que quando a última letra de uma palavra e a primeira letra da palavra seguinte são, ambas, vogais, há ressilabação da estrutura fonológica. Tal processo pode seguir três processos distintos (COLLISCHONN, 2005a, p. 126): elisão (4.1), ditongação (4.2) ou degeminação (4.3).

(4.1) camisa usada > cami[zu]sada

(4.2) camisa usada > cami[zaw]sada

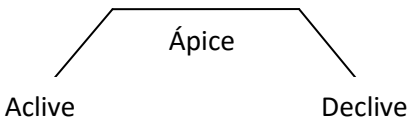
(4.3) camisa amarela > cami[za]marela

No primeiro exemplo, a vogal "a" é elidida e o qualificador "usada" é ligado ao substantivo "camisa" pela vogal "u". No segundo, o falante opta por pronunciar as duas palavras por inteiro, havendo ditongação com o encontro vocálico. Já a degeminação do exemplo 4.3 ocorre porque as vogais são idênticas, sendo a segunda átona. Bisol (2002) padroniza as situações possíveis em quatro categorias: V[9] átona + V átona geram ditongo, se as vogais forem distintas, ou elisão, se a primeira vogal for um "a" átono; V átona + V

[9] V: vogal; C: consoante.

acentuada sempre geram ditongo; V acentuada + V átona geram ditongo ou degeminação; e V acentuada + V acentuada geram ditongo ou uma variação do encontro vocálico chamada hiato, em que há pausa entre as palavras. No mesmo estudo, a autora conclui que geralmente a vogal "a" é elidida no sândi vocálico externo e, quando há ditongação, o ditongo resultante tende a ser crescente, com maior intensidade na segunda vogal.

Em busca de um padrão na formação silábica das palavras, Collischonn (2005a) explica que moldes podem ser construídos a partir da análise, principalmente, de monossílabos da língua. No inglês, por exemplo, é possível observar que a estrutura mínima de uma sílaba é VC ou VV (como "id" e "[aj]", transcrição fonológica de "I", sendo "j" uma semivogal) e a estrutura máxima é CCVVCC (como "[grajnd]", de "grind") (HOGG; MCCULLY, 1987 *apud* COLLISCHONN, 2005a). Em português, segundo a autora, não há consenso sobre o número máximo de elementos de uma sílaba, mas é possível observar formações silábicas que variam de V (como "é") a CCVVC (como "claustro"). Além disso, sabe-se que a sílaba em português constitui-se de um aclive, com uma ou duas consoantes; um ápice, com uma vogal; e um declive, com uma consoante /S/, /r/ ou /l/, a semivogal /j, w/, ou uma consoante nasal (CÂMARA JR., 1969 *apud* COLLISCHONN, 2005a):



Quanto ao acento de intensidade da língua portuguesa, Collischonn (2005b) conclui:

> a) o acento somente pode cair sobre uma das três últimas sílabas da palavra;
> b) a posição do acento na penúltima sílaba é a preferida, quando a palavra for terminada em vogal;
> c) a posição do acento sobre a última sílaba é a preferida, quando a palavra for terminada em consoante.

Proparoxítonas, isto é, palavras cuja sílaba acentuada é a antepenúltima, são menos incidentes em português, segundo a autora. Geralmente se tratam de empréstimos do grego ou do latim. Mesmo nesses casos, existe uma tendência de certos falantes brasileiros em regularizar o acento para a penúltima sílaba da palavra, uma posição mais natural. São exemplos os itens lexicais (COLLISCHONN, 2005b):

abóbra > abobra

árvore > arvri

fóforo > fosfru

xícara > xicra

cócegas > cosca

4.2.6 Aspectos diversos

Além das particularidades já esmiuçadas, diversos outros aspectos são perceptíveis. Pode-se citar mais alguns:

a) Grafia — Graças ao Acordo Ortográfico da Língua Portuguesa, assinado em 2008, foram incorporadas ao português as letras "k", "w" e "y", e o alfabeto tem, hoje, as mesmas letras do inglês. Por outro lado, a língua portuguesa admite sinais diacríticos como a cedilha e os acentos agudo, grave, til e circunflexo, enquanto na escrita inglesa se observa apenas o hífen e o apóstrofo, que também fazem parte do português.

b) Braquilogia — Segundo Bechara (2009, p. 206), "é o emprego de uma expressão mais curta equivalente a outra mais ampla ou de construção mais complexa", com o qual o falante pode simplificar a expressão de seus pensamentos. Um exemplo é a sentença "Estudou como se fosse passar" em lugar de "Estudou como estudaria se fosse passar". A braquilogia, como outros fenômenos de sintaxe, dificulta a análise de sentenças, segundo o autor, ocasião na qual pode ser desfeita.

c) Haplologia sintática — "É o desaparecimento de uma palavra em virtude de estar em contato com outra palavra (ou final de palavra) foneticamente igual ou semelhante" (BECHARA, 2009, p. 207). A haplogia sintática ocorre, por exemplo, para evitar o encontro "que que": "Antes Deus quer/ Que se perdoe um mau, que um bom padeça" (FERREIRA, 1865 *apud* BECHARA, 2009), em vez de "Antes Deus quer que se perdoe um mau que [= do que] [quer] que um bom padeça".

d) Contaminação sintática — "É a fusão irregular de duas construções que, em separado, são regulares" (BECHARA, 2009, p. 207). Por exemplo, em "Caminhar por entre as matas" a combinação de preposições resulta de "Caminhar por matas e caminhar entre as matas". Também se deve à contaminação sintática concordâncias como "As estrelas pareciam brilharem", segundo Bechara (2009).

e) Expressão expletiva — "É a que, sem função sintática, enfatiza um termo da oração ou o pensamento integral" (BECHARA, 2009, p. 208). A expressão "é que" em "Nós é que somos brasileiros" é expletiva, por exemplo.

f) Objeto elíptico — Certos verbos, comuns ao inglês e ao português, são intransitivos, ou seja, não possuem qualquer complemento. É o caso do verbo "sorrir": "A menina sorriu".

No entanto, há verbos na língua portuguesa que, mesmo transitivos, podem ser usados sem complemento, ficando o objeto subentendido. Perini (2010) denomina as orações resultantes de "construções transitivas de objeto elíptico", como as sentenças: "Catapora também mata" e "O marido dela bebe". Tal situação é semelhante àquela do sujeito nulo e, portanto, exige tratamento especial.

g) Negação — Conforme Perini (2010), a partícula negativa costuma vir antes do verbo em português. Dos exemplos do autor: "O Eugênio não trabalha mais aqui"; "O Eugênio nunca trabalhou aqui". Todavia, é possível a adição de partículas enfáticas após o verbo, em orações com duas ou mais negações: "Eu não vou lá não"; "A professora não vai ajudar ninguém"; "Eu nunca pedi a ninguém que me ajudasse em nada". Nesses casos, o número de negações na sentença não afeta a interpretação (uma não cancela a outra), segundo o autor, sendo sempre negativa.

4.2.7 Regionalismos e variações socioletais

Tais particularidades, mesmo quando diferem da norma culta, são gerais ao português utilizado no território brasileiro. Há particularidades, porém, que dizem respeito a comunidades e grupos específicos do país. É o caso das variações regionais — ou regionalismos — e das variações peculiares a grupos socioeconômicos, de profissão ou cultura diferentes — também chamadas socioletais.

O pronome "tu", como se viu, tem sido substituído pela forma "você", para simplificação da concordância verbal. Todavia, em diversos estados brasileiros (sobretudo, do Sul e do Nordeste), esse pronome é de uso corrente no tratamento informal, enquanto "você" pode demonstrar certo distanciamento (CASTILHO, 2010; PERINI, 2010). Nessas regiões, a utilização de "tu" também é variável, podendo acompanhar conjugações verbais de segunda ou terceira pessoa, inclusive em forma lexical simplificada, como "tá" ou "tás". Outra variação se percebe na própria forma "você", com "cê" e "ocê". Além disso, o clítico "lhe" pode ser utilizado em lugar de "te", a depender do local (PERINI, 2010).

No estado de Minas Gerais, existem formas reduzidas para os pronomes pessoais do caso reto: "e el [ele] falou que não podia mais dar aula"; "o albergue tava lotado, eis [eles] não registravam mais ninguém"; "se não fosse eu, éa [ela] tinha matado a colega" (CORRÊA, 2002). No Vale do Ribeira, em São Paulo, e na baixada de Cuiabá, verifica-se falta de

concordância de gênero: "o meu sobrinha"; "cabelo grossa" (CASTILHO, 2010). Nessas e em outras regiões do Brasil, verifica-se também variantes lexicais: abóbora, jerimum e moranga; tangerina, mexerica e bergamota; mandioca, aipim e macaxeira.

Quanto às variações socioletais, pode-se citar aspectos comuns à fala de pessoas de baixa escolaridade. A regularização induzida ao acento de proparoxítonas, já mencionada, é um exemplo dos usos desses falantes. Segundo Castilho (2010), percebe-se o fenômeno da iodização da palatal "lh", como em "orelha" e "velho", pronunciados como [o'reya] e ['vɛyu]. "Há perda do valor do sufixo -ior nos comparativos de superioridade, utilizando-se o advérbio 'mais': 'mais mió', 'mais pió'" (p. 207). Distingue-se as formas verbais do pretérito perfeito daquelas do presente por meio da elevação da vogal temática. "Ficamos", por exemplo, que é invariável segundo a norma culta, é dito "fiquemu", no pretérito, e "ficamu", no presente. Além disso, Bagno (2006) adiciona o fenômeno da rotacização do "l" para o "r" em encontros consonantais, como em "pranta", "ingrês" e "frecha". Tais variações, embora mais observáveis nesse grupo social, exercem forte influência no português brasileiro como um todo.

4.3 Possíveis adaptações ao sistema

A localização do Watson para o português brasileiro, preservando-se a forma como o sistema foi constituído, requer a adaptação de cada um de seus componentes que processem, originalmente, o inglês. A English Slot Grammar e o respectivo *parser* podem receber uma versão no PB a partir do formalismo XSG, como se viu, extensível para outras línguas. A despeito do Watson Multilíngue e dos esforços intensos que a atividade de construção de uma Slot Grammar para cada língua demandaria, o Watson em português do Brasil, especificamente, pode incluir a Brazilian Portuguese Slot Grammar (BPSG), já em desenvolvimento conforme McCord (2010).

A BPSG deve observar, entre outras, as particularidades do português brasileiro aqui discutidas. Para tanto, as obras de Perini (2010), Castilho (2010), Bechara (2009), Bagno (2006), Bisol (2002; 2005) e Ilari (2002) podem ser consultadas, por fornecerem insumos para a elaboração de regras de descrição linguística do PB, geralmente não encontrados em gramáticas da norma culta da língua portuguesa. Dias da Silva *et al.* (2007, p. 44) explica que a maioria dos *parsers* atuais requerem uma etapa de pré-processamento dos *inputs*, de

maneira a suprimir os traços de flexibilidade e os fenômenos de desvio da norma padrão da língua:

> Trata-se do processo de **regularização sintática**, através do qual as informações omitidas (os sujeitos elípticos, por exemplo) são restauradas, as anáforas são indicializadas, as formas passivas são substituídas pelas formas ativas, a sentença é reorganizada a partir da ordem direta (sujeito verbo objeto), e as clivagens e topicalizações são suprimidas. (grifo do autor)

Entretanto, o autor argumenta que essa tarefa pode acarretar prejuízos semânticos na compreensão das sentenças. Além disso, segundo ele, gramáticas limitadas, que apenas aceitam *inputs* regularizados, fracassam "na análise das frases produzidas no registro oral, marcadas por falsos inícios, hesitações, repetições, retomadas, anacolutos, topicalizações e movimentos de natureza pouco previsível" (DIAS-DA-SILVA *et al.*, 2007, p. 45). Na etapa de Análise da Pergunta da DeepQA, a ferramenta de construção de PAS executa atividade semelhante de regularização sintática para a língua inglesa. Todavia, as sentenças são, primeiro, analisadas sintática e semanticamente pelo *parser*, e a estrutura profunda gerada, que diz respeito ao significado, se conserva ao longo do processo. A regularização, nesse caso, tem o objetivo apenas de simplificar a Geração de Hipóteses, na qual a resposta já terá sido interpretada pelo sistema.

Othero (2006; 2009) possui estudos descritivos sobre o PB visando à implementação computacional que também podem ser úteis à localização, todavia, as regras se baseiam na Teoria X-Barra do PLN, mais voltada para o princípio da constituência da análise sintática, enquanto o *parser* das Slot Grammars seguem o princípio da dependência. Isso traz divergências na classificação funcional dos termos da oração, que, o primeiro deve constituir-se de sintagmas e, para o segundo, deve ser formada por sujeito, verbo e objeto. De qualquer maneira, as descrições do autor podem ser utilizadas na elaboração de uma gramática equivalente para *dependency parsing*, no caso, a BPSG. Outra desvantagem nos estudos de Othero é a descrição exclusiva de orações simples declarativas, isto é, sentenças com apenas um verbo e que não sejam negativas ou interrogativas — a não ser interrogativas abertas, similares às declarativas, como se viu. Entretanto, o autor implementa suas regras em Prolog, a mesma linguagem utilizada nas regras de detecção de LAT, foco, entidades nomeadas e relações pela DeepQA.

Ainda sobre a descrição do PB, necessária ao aperfeiçoamento da BPSG, a grafia particular da língua exige que o sistema processe caracteres que excedem o padrão ASCII, sendo indispensável uma adaptação para o padrão Unicode. A fonologia, além das regras

descritivas que lhe dizem respeito, pode ser "aprendida" pelo sistema com a compreensão automática de padrões em *corpora* de fala e do PB falado escrito, como o CSLU Spoltech Brazilian Portuguese, o West Point Brazilian Portuguese Speech, o C-ORAL-BRASIL e o Corpus Brasileiro. Regionalismos e variações socioletais podem ser incorporados aos poucos nas descrições da gramática e nos componentes de fonologia do sistema.

Quanto aos recursos com dados semiestruturados e estruturados que a DeepQA utiliza, sobretudo para a detecção, classificação e extração de entidades nomeadas e relações, pode-se incorporar versões em português dos recursos já utilizados pelo sistema, quando disponíveis, e repositórios construídos especificamente para o português brasileiro. O WordNet, por exemplo, tem uma versão WordNet.Br, inicialmente apenas com verbos, e a versão WordNet.PT Global, com várias variedades do português, inclusive a brasileira. A DBpédia tem uma versão em português, entretanto, o projeto é aparentemente compartilhado entre os países de língua portuguesa, sem distinção de variedade linguística. A versão em português da Freebase inclui entidades e conhecimento relacionados a Portugal e teria de ser substituída no sistema. A Wikipédia, com dados semiestruturados, é ativa no Brasil, porém seus artigos devem ser ampliados em número e conteúdo. E a YAGO, embora alcance diversas línguas, não tem versão em português. Os axiomas do PRISMATIC, criado para o Watson, teriam de ser traduzidos, para comportar o vocabulário de relações conceituais de que o português brasileiro dispõe. Outros recursos são listados pela Linguateca[10], um projeto de promoção e apoio ao processamento computacional do português, e podem auxiliar na localização no sistema. Inclusive, há opções de léxico geral e especializado, que podem servir à composição do dicionário de base da BPSG.

Na falta de recursos ontológicos suficientes ou adequados, Gottschalg Duque (2005) propõe um sistema de recuperação de informações que produz ontologias de modo semiautomático. O Sistema de Recuperação de Informação baseado em Teorias da Linguística Computacional e Ontologia (SiRILiCO) foi construído com o *parser* PALAVRAS da língua portuguesa, capaz de produzir representações sintáticas de diversos formatos, inclusive árvores de dependência. A análise semântica do sistema é tarefa do GeraOnto, uma ferramenta própria, que, criada a partir de editor de ontologias Protégé, gera "ontologias leves" com a representação do significado. Uma vez que essas ferramentas se encontram separadas em módulos, no sistema, é possível conectar ferramentas externas, e um teste de

[10] http://www.linguateca.pt/

compatibilidade poderia ser feito com o *parser* da BPSG. Entretanto, as funcionalidades desse sistema teriam de ser usadas, apenas, em aplicações do Watson para domínios específicos, uma vez que as ontologias leves podem produzir ruídos na compreensão de certas informações pela máquina (DIN & ENGELS, 2001 *apud* GOTTSCHALG-DUQUE, 2005).

Uma maneira mais simplória de adaptar o Watson ao português brasileiro, enquanto o Watson Multilíngue é desenvolvido, seria conectá-lo às próprias APIs cognitivas disponibilizadas pela IBM: a Speech to Texto API faria a transcrição da fala do usuário em PB; a Language Detection API (da AlchemyLanguage API) reconheceria que o texto transcrito está em português; e a Language Translation API faria a tradução do texto do português brasileiro para o inglês americano. Traduzido o texto, o processo subsequente da DeepQA seria exatamente o mesmo, sem qualquer alteração das ferramentas ou recursos. Apenas o *output* do sistema, isto é, a resposta à pergunta do usuário, teria de ser traduzido novamente para o PB e pronunciado pelo Watson. Para tanto, a Text to Speech API teria de incluir o português brasileiro.

Tal alternativa, entretanto, não tem sido seguida no projeto de localização do Watson para o japonês, que se baseia, segundo as evidências (IBM, [2015]b), em aprendizagem de máquina. O sistema recebe textos em japonês, executa o *parsing*, falantes nativos corrigem quaisquer falhas e o sistema aprende com os erros. O mesmo poderia ser feito na localização para o PB, resguardadas as devidas particularidades. A aprendizagem de máquina, aliás, foi utilizada no sistema original em inglês e vem sendo intensificada nas funcionalidades do Watson, a exemplo do Watson Knowledge Studio, em construção, que permitirá aos usuários treinar o Watson para fins específicos.

Mesmo no Watson Multilíngue, a aprendizagem de máquina é bastante explorada e, segundo Cortis *et al.* (2014), *treebanks* serão necessários para treinar o *parser* estatístico, caso o sistema ainda inclua a etapa de *parsing*. Tais *treebanks* dependerão das línguas que o Watson processar após a internacionalização, incluindo material em português brasileiro, entretanto, deverão ser compatíveis com as etiquetas morfológicas e os rótulos de dependência do *parsing* universal de McDonald *et al.* (2013). A construção do *treebank* para o PB já começou segundo informações[11] da Universal Dependencies, responsável pela implementação do *parsing* universal, o que permite inferir avanços na construção do Watson Multilíngue.

[11] https://github.com/ryanmcd/uni-dep-tb

5 Discussões finais

A escassez de recursos de processamento automático do português, em especial, do português brasileiro, que sejam aplicáveis a projetos internacionais sugere a necessidade de uma conscientização das academias, empresas de pesquisa e dos governos locais sobre a urgência da inclusão dessa língua no mundo digital com maior expressividade. A Sociedade da Informação e do Conhecimento exige cada vez mais que os seus cidadãos estejam a par dos acontecimentos à sua volta e, nesse contexto, a tecnologia é o veículo e apenas a língua suficientemente adaptada para o tratamento automático pode tornar-se um código efetivo de informações. Sendo assim, é fundamental que as diferentes culturas existentes no mundo físico façam-se presentes também no ciberespaço e nos sistemas computacionais, por meio de suas línguas, de modo a evitar que a preponderância de uma única língua, mais desenvolvida tecnologicamente, ou de um grupo pequeno determine uma hegemonia social nada condizente com a diversidade humana.

Nesse sentido, propõe-se que mais universidades abordem em suas pesquisas o Processamento Automático de Línguas Naturais e as peculiaridades do português brasileiro com insumos para o desenvolvimento de aplicações tecnológicas em língua portuguesa e para a localização de produtos já existentes exterior. Mercado há. Uma evidência disso são as parcerias já firmadas com a IBM, no país, para a aquisição do Watson e os esforços intensos da empresa em localizá-lo para o português do Brasil. Além disso, é imperioso que haja maior cooperação entre os países de língua portuguesa para a construção de projetos maiores, com resultados mais eficazes, observando-se as particularidades das variedades linguísticas de cada um. Junto a isso, mais fontes de financiamento à pesquisa científica e mais incentivos à produção empresarial no setor. Só assim será possível a essa comunidade e, particularmente, ao Brasil, tornar-se participante ativo do mundo globalizado e da Era da Informação.

6. Referências Bibliográficas

BAGNO, M. **A língua de Eulália**: novela sociolingüística. 15. ed. São Paulo: Contexto, 2006.

BATES, M. **Models of natural language understanding**. Proceedings of National Academy Science, USA, 1995, pp. 9977-9982.

BECHARA, E. **Lições de português pela análise sintática**. 18. ed. rev. e ampl., com exercícios resolvidos. Rio de Janeiro: Nova Fronteira, 2009.

BISOL, L. (Org.) **Introdução a estudos de fonologia do português brasileiro**. 4. ed. rev. e ampl. Porto Alegre: EDIPUCRS, 2005.

BISOL, L. Sândi vocálico externo. In: ILARI, R. (Org.) **Gramática do português falado**. 4 ed. rev. Campinas: Editora da Unicamp, 2002. v. II: Níveis de análise linguística.

BORENSTEIN, S; ROBERTSON, J. **IBM 'Watson' Wins: 'Jeopardy' Computer Beats Ken Kennings, Brad Rutter.** Huffpost Tech. Disponível em: <http://www.huffingtonpost.com/2011/02/17/ibm-watson-jeopardy-wins_n_824382.html>. Acesso em: 7 dez. 2015.

BRANCO, A; MENDES, A; PEREIRA, S; HENRIQUES, P; PELLEGRINI, T; MENEIDO, H; TRANCOSO, I; QUARESMA, P; LIMA, V.L.S; BACELAR, F. **A língua portuguesa na era digital** – The Portuguese Language in the Digital Age. META-NET White Paper Series: Europe's Languages in the Digital Age. Springer, Heidelberg, 2012.

CAJUEIRO, R. L. P. **Manual para elaboração de trabalhos acadêmicos:** guia prático do estudante. Editora Vozes Ltda. 1ª ed. Rio de Janeiro, 2013.

CAMBRIA, E.; WHITE, B. **Jumping NLP Curves**: a Review of Natural Language Processing Research. IEEE Computational Intelligence Magazine 9, p. 48-57, 2014.

CAMPOS, O. G. L. A. S.; RODRIGUES, A. C. S. Flexão nominal: indicação de pluralidade no sintagma nominal. In: ILARI, R. (Org.) **Gramática do português falado**. 4 ed. rev. Campinas: Editora da Unicamp, 2002. v. II: Níveis de análise linguística.

CASTELLS, Manuel. **A sociedade em rede**. Volume I; 8 edição revista e ampliada; tradução de Roneide Vanancio Majer com colaboração de Klauss Brandini Gerhardt. São Paulo: Paz e Terra, 2005.

CASTILHO, A. T. **Nova Gramática do Português Brasileiro**. 1. ed., 1ª Reimpressão. São Paulo: Editora Contexto, 2010.

COLLISCHONN, G. (2005a) A sílaba em português. In: BISOL, L. (Org.) **Introdução a estudos de fonologia do português brasileiro**. 4. ed. rev. e ampl. Porto Alegre: EDIPUCRS, 2005.

COLLISCHONN, G. (2005b). O acento em português. In: **Introdução a estudos de fonologia do português brasileiro**. 4. ed. rev. e ampl. Porto Alegre: EDIPUCRS, 2005.

COMPUTERWORLD. **IBM conquista segundo projeto de Watson no Brasil**. 7 dez. 2015. Disponível em: <http://computerworld.com.br/ibm-conquista-segundo-projeto-de-watson-no-brasil>. Acesso em: 7 dez. 2015.

COPPIN, B. **Artificial Intelligence Illuminated**. Jones and Bartlett Publishers Inc., 2004.

CORRÊA, L. T. A variação lingüística eles/es e a indeterminação de sujeito. In: COHEN, M. A. A. M.; RAMOS, J. M. (Org.) **Dialeto mineiro e outras falas**: estudos de variação e mudança lingüística. Belo Horizonte: Faculdade de Letras/ UFMG, 2002.

CORTIS, K.; BHOWAN, U.; MAC AN TSAOIR, R.; MACCLOSKEY, D. J.; SOGRIN, M.; CADOGAN, R. **What or Who is Multilingual Watson?** In: Proceedings of COLING 2014 - 25th International Conference on Computational Linguistics: System Demonstrations, p. 95-99. Dublin, ago. 2014.

DIAS-DA-SILVA, Bento Carlos. **O estudo lingüístico-computacional da linguagem**. Letras de Hoje, EDIPUCRS Porto Alegre Brasil, v. 41, n. 144, p. 103-138, 2006.

_____; MONTILHA, G.; RINO, L.H.M.; SPECIA, L.; NUNES, M.G.V.; OLIVEIRA JR., O.N.; MARTINS, R.T.; PARDO, T.A.S. (2007). **Introdução ao Processamento das Línguas Naturais e Algumas Aplicações**. Série de Relatórios do NILC. NILC-TR-07-10. São Carlos-SP, Agosto, 121p.

FELIPPO, A. Di; DIAS-DA-SILVA, B. C. **Uma introdução à engenharia do conhecimento linguístico**. Revista de Letras, v. 1, p. 57-72, 2008.

FERRRUCI, D. A. **Introduction to "This is Watson"**. IBM Journal of Research and Development, vol. 56 3.4: IBM, pp. 1–1, 2012.

_____; NYBERG, E.; ALLAN, J.; BARKER, K.; BROWN, E.; CHU-CARROLL, J.; CICCOLO, A.; DUBOUE, P.; FAN, J.; GONDEK, D.; HOVY, E.; KATZ, B.; LALLY, A.; MCCORD, M.; MORARESCU, P.; MURDOK, W.; PORTER, B.; PRAGER, J.; STRZALKOWSKI, T.; WELTY, W.; and ZADROZNY, W. 2009. **Towards the Open Advancement of Question Answer Systems**. IBM Technical Report RC24789, Yorktown Heights, NY.

_____; BROWN, E.; CHU-CARROLL, J.; FAN, J.; GONDEK, D.; KALYANPUR, A.; LALLY, A. MURDOCK, J. W.; NYBERG, E.; PRAGER, J.; SCHLAEFER, N.; WELTY, C. **Building Watson**: An overview of the DeepQA project. AI Mag., vol. 31, no. 3, pp. 59–79, 2010.

GANDOUR, Fábio. O que muda com a computação cognitiva? **Revista da ESPM**, São Paulo, ano 20, ed. 95, n. 5, p. 12-21, set./out. 2014.

GOTTSCHALG-DUQUE, Cláudio. **SiRILiCO**: uma proposta para um sistema de recuperação de informação baseado em teorias da linguística computacional e ontologia. 2005. Tese (Doutorado)- Escola de Ciência da Informação da Universidade Federal de Minas Gerais, Belo Horizonte, 2005.

ILARI, R. (Org.) **Gramática do português falado**. 4 ed. rev. Campinas: Editora da Unicamp, 2002. v. II: Níveis de análise linguística.

INTERNATIONAL BUSINESS MACHINES. IBM Software Group. **The Era of Cognitive Systems**: An inside look at IBM Watson and how it works. Whitepaper. Somers, NY, 2012.

INTERNATIONAL BUSINESS MACHINES. (2015a). **IBM, SoftBank Alliance to Bring Watson to All of Japan**. Press release, Tóquio e Nova York, 10 fev. 2015. Disponível em: <http://www-03.ibm.com/press/us/en/pressrelease/46045.wss>. Acesso em: 7 dez. 2015.

INTERNATIONAL BUSINESS MACHINES. (2014). **IBM Watson acelera su expansión global**. Nota de Prensa, Madrid, 9 out. 2014. Disponível em: <http://www-03.ibm.com/press/es/es/pressrelease/45080.wss>. Acesso em: 7 dez. 2015.

INTERNATIONAL BUSINESS MACHINES. ([2015]b). **Watson's Learning Japanese**. Disponível em: <http://www.ibm.com/smarterplanet/us/en/ibmwatson/ibm_watson_learns_japanese_with_softbank.html>. Acesso em: 7 dez. 2015.

JURAFSKY, Daniel; MARTIN, James H. **Speech and Language Processing**: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. 2ª edição. Prentice Hall, 2008.

LALLY, Adam; FODOR, Paul. **Natural Language Processing With Prolog in the IBM Watson System**. The Association for Logic Programming (ALP) Newsletter, March 2011.

_____; J. M. Prager; MCCORD, M. C.; BOGURAEV, B. K.; PATARDHAN, S.; FAN, J.; FODOR, P.; CHU-CARROLL, J. **Question analysis**: How Watson reads a clue. IBM J. Res. & Dev., vol. 56, no. 3/4, Paper 2, pp. 2:1–2:14, May/Jul. 2012.

LEWIS, B. L. **In the game**: the interface between Watson and Jeopardy!, IBM J. Res. & Dev., vol. 56, no. 3/4, Paper 17, pp. 17:1–17:6, May/Jul. 2012.

LIDDY, E. D. Natural Language Processing. In: **Encyclopedia of Library and Information Science**, 2nd Ed. Marcel Decker, Inc. 2001.

LUGER, G. F. **Inteligência Artificial**. 6ª edição. São Paulo: Editora Pearson Education do Brasil, 2013.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. **Introduction to Information Retrieval**. Cambridge University Press, 2008.

_____; C. D.; SCHÜTZE, H. **Foundations of Statistical Natural Language Processing**. Cambridge: MIT press, 1999.

MATSU, C. Exclusivo: veja como e onde o Bradesco está adotando o Watson. **Computerworld**, 9 jun. 2015. Disponível em: < http://computerworld.com.br/exclusivo-veja-como-e-onde-o-bradesco-esta-adotando-o-watson>. Acesso em: 7 dez. 2015.

MCCORD, M. C. **Using slot grammar**. IBM T. J. Watson Res. Center, Yorktown Heights, NY, IBM Res. Rep. RC23978.

_____; MURDOCK, J. W.; BOGURAEV, B. K. **Deep parsing in Watson**. IBM J. Res. & Dev., vol. 56, no. 3/4, Paper 3, pp. 3:1–3:15, May/Jul. 2012.

MCDONALD, R.; NIVRE, J.; QUIRMIBACH-BRUNDAGE, Y.; GOLDBERG, Y.; DAS, D.; GANCHEV, K.; HALL, K.; PETROV, S.; ZHANG, H.; TÄCKSTRÖM, O.; BEDINI, C.; CASTELLÓ, N. B.; LEE, J. **Universal Dependency Annotation for Multilingual Parsing**. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, p. 92-97. Sofia, Bulgária, ago. 2013.

OTHERO, G. A. **A gramática da frase em português**: algumas reflexões para a formalização da estrutura frasal em português. Porto Alegre: EDIPUCRS, 2009. 160 p.

OTHERO, G. A.; MENUZZI, S. M. **Lingüística Computacional**: Teoria & Prática. 1. ed. São Paulo SP: Parábola Editorial, 2005. v. 1. 128 p.

OTHERO, G. A. **Teoria X-Barra**: descrição do português e aplicação computacional. São Paulo: Contexto, 2006.

PERINI, M. A. **Gramática do português brasileiro**. São Paulo: Parábola, 2010.

PRUDÊNCIO, A. C.; VALOIS, D. A.; DE LUCCA, J. E. Introdução à Internacionalização e à Localização de Software. **Cadernos de Tradução**, Florianópolis, v. 2, n. 14, p. 211-242, jul. 2004.

RUSSELL, S.; NORVIG, P. **Artificial Intelligence**: A Modern Approach. Prentice-Hall, 3ª edição, 2010.

STRICKLAND, Eliza. **IBM's Watson Learns to Cook from Bon Appetit Magazine**. IEEE Spectrum. Disponível em: < http://spectrum.ieee.org/tech-talk/robotics/artificial-intelligence/ibm-watson-learns-to-cook-from-bon-appetit-magazine>. Acesso em 7 dez. 2015.

TESAURO, G.; GONDEK, D. C.; LENCHNER, J.; FAN, J.; PRAGER, J. M. **Simulation, learning, and optimization techniques in Watson's game strategies**, IBM J. Res. & Dev., vol. 56, no. 3/4, Paper 16, pp. 16:1–16:11, May/Jul. 2012.

VAHID, Frank. **Sistemas Digitais**: Projeto, Otimização e HDLs. Porto Alegre: Artmed, 2008. 560 p.

VAN DIJK, Teun Adrianus. **Cognição, discurso e interação**. Organização e apresentação de Ingedore Grunfeld Villaça Koch. 6 ed. São Paulo: Contexto, 2004.

VOLTOLINI, Ramon. **Watson, o supercomputador que promete erradicar 8 tipos de câncer**. TecMundo. Disponível em: <http://www.tecmundo.com.br/ciencia/45898-watson-o-supercomputador-que-promete-erradicar-8-tipos-de-cancer.htm>. Acesso em: 7 dez. 2015.