Uma história da linguística computacional no âmbito das ciências cognitivas

Marco Rocha ·

Resumo

Este trabalho procura descrever como a área de pesquisa denominada linguística computacional evoluiu desde a sua concepção inicial, associada à tradução de máquina, até os dias de hoje. A área de pesquisa foi criada como uma forma de dar credibilidade científica à tradução de máquina, após a constatação de baixa confiabilidade por parte do relatório ALPAC (1966). O percurso subsequente demonstra o esforço para recuperar esta confiabilidade através de uma fundamentação científica considerada adequada, a qual deveria derivar da teoria linguística e da ciência cognitiva. O artigo procura descrever este processo evolutivo de natureza pendular. A conclusão expõe a necessidade do desenvolvimento de métodos híbridos, integrando métodos lógicos e probabilísticos, na busca de um novo caminho mais profícuo na evolução da área de pesquisa.

Palavras-chave

Linguística computacional; Ciência cognitiva; Computação simbólica; Métodos probabilísticos; Métodos híbridos.

A history of computational linguistics in cognitive science

Abstract

The aim of this paper is to describe how the field known as computational linguistics evolved from its inception, associated with machine translation, to this day. This field was created as a way to afford scientific credibility to machine translation after low reliability was demonstrated in the ALPAC report (1966). The following developments evidence efforts to improve reliability through adequate scientific foundations, which ought to come from linguistic theory and cognitive science. This paper describes the pendulum-like nature of this process and finally demonstrates the need for hybrid methods integrating logic and probabilistic methods in the search for a more profitable path for the future development of this field of research.

Keywords

Computational linguistics; Cognitive science; Symbolic computation; Probabilistic methods; Hybrid methods

· Centro de Comunicação e Expressão, Universidade Federal de Santa Catarina, Brasil. ✉ marco@cce.ufsc.br

> "Do not be bullied by authoritative pronouncements about what machines will never do. Such statements are based on pride, not fact."
> Marvin Minsky, 1982

Introdução

A epígrafe acima serve como uma espécie de inspiração para a descrição histórica apresentada neste artigo. A palavra *pride*, em particular, desempenha um papel especial na situação descrita, quando a referência é a maneira pela qual os seres humanos usam a língua. A capacidade da fala é, em termos simples, aquilo que distingue os seres humanos dos outros animais. Não é surpreendente, portanto, que os seres humanos tenham orgulho desta capacidade. Ao mesmo tempo, a citação ilustra a dificuldade substancial que pesquisadores e implementadores encontram para fazer com que computadores e robôs falem como qualquer ser humano. Consequentemente, seria possível argumentar, diante da modéstia de resultados que a área apresenta atualmente: de fato, as máquinas nunca serão capazes de falar e entender línguas humanas. Não obstante, aplicações como reconhecimento de fala, interfaces de diálogo e tradução de máquina continuam a ser encaradas como objetivos relevantes para o direcionamento dessa área do conhecimento.

As três aplicações mencionadas têm situação distinta no âmbito de um objetivo tecnológico da linguística computacional, isto é, construir sistemas de computador capazes de processar línguas humanas com eficiência. A palavra eficiência significa, em última análise, com desempenho semelhante ao de um ser humano especialista. A tradução de máquina é a aplicação mais antiga. Na verdade, teve o papel de servir como motivação para o surgimento do termo linguística computacional, como será exposto no decorrer do trabalho. O reconhecimento de fala é provavelmente a aplicação de maior êxito no âmbito da área de investigação, já que existem alguns sistemas que são comercializados regularmente por realizarem esta função, ou seja, dada uma entrada sonora em uma língua humana, o sistema é capaz de realizar a digitalização do texto. A terceira aplicação é o sonho que caracterizaria a conquista de uma capacidade caracteristicamente humana por parte das máquinas, as quais seriam capacitadas a conversar, responder perguntas, fazer perguntas e, em termos gerais, utilizar um banco de dados, por exemplo, uma versão digitalizada do *Código do Consumidor*, como meio de fornecer informações a um usuário que não tem especialização em computação. Em termos simples, estas máquinas não existem, a despeito de aproximações eficientes para domínios restritos.

As limitações dos sistemas que processam línguas humanas, como será discutido em maior detalhe no corpo do artigo, levaram a uma exigência de fundamentação científica mais aprofundada como caminho para atingir um patamar superior de desempenho operacional. Esta fundamentação científica viria da teoria linguística, estabelecendo um elemento de interdisciplinaridade intrínseco à linguística computacional. Em termos mais abrangentes, um aperfeiçoamento da compreensão existente sobre as línguas humanas, como também de seu uso em contextos reais de comunicação, deveria estender-se para uma compreensão da natureza do que se convencionou chamar de inteligência. As semelhanças identificadas entre a inteligência humana e outros sistemas de manipulação de símbolos,

como os computadores, passaram a fazer parte dos objetivos de pesquisa. O objetivo cognitivo passou a fazer parte do universo da linguística computacional. A combinação interdisciplinar primordial das ciências cognitivas (a saber, linguística, ciência da computação, psicologia e filosofia) seria, ao mesmo tempo, fonte de um aprofundamento da fundamentação científica, o que levaria a avanços no objetivo tecnológico de construir sistemas melhores. A complexidade da interação entre os dois objetivos ao longo da evolução da área do conhecimento é o assunto deste artigo.

Linguística computacional e ciência cognitiva

Na página da *Association for Computational Linguistics,* a definição da área apresentada é: "o estudo científico das línguas a partir de uma perspectiva computacional"[1]. Em termos menos formais, Kay reforça o essencial da definição: "a linguística computacional procura fazer o que os linguistas fazem de uma maneira computacional"[2]. Conforme apontado por Abney,[3] a linguística computacional, com base nestas definições, é um ramo da linguística. Porém, alguns fatos relacionados à área não parecem autorizar esta conclusão. Estes fatos são apresentados em seguida, ainda fazendo uso do material em Abney, como também de vivência direta.

Primeiramente, a enorme maioria dos pesquisadores em linguística computacional é formada em ciência da computação e trabalha em departamentos de ciência da computação. Além disso, a história da linguística computacional é claramente distinta da história recente dos estudos linguísticos. A metodologia de pesquisa é baseada em coleta intensiva de dados, seguida de análise sistemática do uso da língua em busca de padrões que possam servir para elaborar uma teoria deste uso, a qual é testada experimentalmente segundo parâmetros específicos. Isto não significa que a linguística computacional tenha necessariamente deixado de ser linguística. O uso da tecnologia na análise intensiva de dados seria uma forma alternativa de linguística, a qual poderia ser comparada ao uso de computação em biologia ou astronomia. A tecnologia é, portanto, um meio para o fim de compreender melhor as línguas humanas.

Segundo Kay, o termo linguística computacional foi criado por Hays (1962) em resposta ao que viria a aparecer no Relatório ALPAC (*Automatic Language Processing Advisory Committee*, 1966).[4] O relatório avaliou os resultados e perspectivas da pesquisa em tradução de máquina desde os primórdios até então, já que o investimento havia sido considerável, mas os resultados, modestos. Os esforços para criar um sistema capaz de traduzir do russo para o inglês automaticamente esbarravam nas limitações dos computadores dos anos 50 e 60. Porém, a comissão, indicada pela *National Academy of Sciences* e presidida por John Pierce,

[1] No original: "the scientific study of language from a computational perspective", www.aclweb.org. As traduções ao longo do texto são do próprio autor

[2] No original: "computational linguistics is trying to do what linguists do in a computational manner"; Martin Kay, "A Life of Language," *Computational Linguistics*, 16, nº1 (1994): 1-13, em 5.

[3] Steven Abney "Data-intensive Experimental Linguistics," *Linguistic Issues in Language Technology* 6, nº2 (2011): 1-27.

[4] Kay, "Life of Language", 5.

percebeu que não havia, no funcionamento do sistema, fundamento linguístico que pudesse guiar a sua construção. Portanto, não havia uma base científica adequada. Os elementos desta base científica deveriam vir da teoria linguística.

Hays conheceu antecipadamente o conteúdo do relatório (na verdade, tinha ajudado a escrever o documento) e sabia das dificuldades políticas que traria para o financiamento da tradução de máquina. O termo e o campo de investigação são produto, portanto, de uma iniciativa preventiva para criar um 'braço científico' que desse mais credibilidade à tradução de máquina. A linguística computacional teria sido criada, deste modo, sem ter maiores relações com a linguística propriamente dita. Hays cunhou o termo e logo em seguida fundou a *Association for Machine Translation and Computational Linguistics*, a qual tornou-se simplesmente *Association for Computational Linguistics* em 1968. O termo 'processamento de linguagem (língua) natural' (PLN), oriundo da inteligência artificial, foi rejeitado por Hays, que o considerou insuficientemente erudito.

Não obstante, o termo foi adotado pela comunidade de pesquisadores diretamente ligados à inteligência artificial para se referirem a qualquer componente de processamento de língua dentro de inteligência artificial, com destaque para o acesso a banco de dados através de uma língua humana. Os termos PLN e linguística computacional, em algumas situações, podem ser vistos como semelhantes e até mesmo intercambiáveis. Não resta dúvida, contudo, que a pesquisa em PLN está associada a uma engenharia da linguagem, ou engenharia de software, capaz de lidar com línguas humanas, o que constitui um ramo da inteligência artificial. Já a denominação linguística computacional incorpora explicitamente um elemento de interdisciplinaridade.

A linguística computacional absorve, deste modo, elementos da inteligência artificial que ultrapassam a tradução de máquina e reforçam a necessidade de tratamento interdisciplinar. A possibilidade de integrar, por exemplo, conhecimento oriundo da psicolinguística aos fundamentos de uma psicologia computacional, com reflexos em aplicações no campo da interação humano-máquina, torna as pesquisas de natureza interdisciplinar atraentes. O tratamento de computadores, robôs e seres humanos como processadores de símbolos estabelece ligações entre os objetos de pesquisa dos diferentes componentes da ciência cognitiva e enriquece significativamente a discussão sobre a maneira pela qual as línguas humanas são utilizadas pelos processadores. A pesquisa volta-se para a compreensão da lógica do conhecimento linguístico.

Abordagens lógicas em linguística computacional

As décadas de 70 e 80 são caracterizadas pelas abordagens lógicas, também chamadas de abordagens baseadas em regras. Nestas abordagens, esperava-se que uma representação adequada do que é o conhecimento linguístico levasse a uma capacidade de compreender línguas humanas por parte das máquinas. Isto significa que uma representação dada sob a forma de sons ou caracteres precisa ser transformada numa outra representação que expresse o conhecimento linguístico necessário para de fato entender uma sentença ou declaração. O primeiro passo seria determinar o significado de cada um dos

elementos que compõem uma sentença, as palavras. Esta compreensão depende de uma consulta a alguma forma de dicionário ou léxico que registre o significado das palavras de uma língua.

Em seguida, a organização destas palavras na sentença, a sintaxe, deve ser estabelecida através de vínculos entre as palavras. As palavras são portanto, agrupadas em unidades sintáticas, muitas vezes chamadas pelos linguistas de sintagmas ou, mais simplesmente, grupos. As sentenças seriam compostas por dois tipos básicos de sintagma, o sintagma ou grupo nominal ('véu de noiva') e o sintagma ou grupo verbal ('fiz o jantar'). A partir de gramáticas previamente especificadas sob a forma de regras, inspiradas inicialmente na gramática gerativa chomskyana, seria possível definir quais as sentenças gramaticais de uma determinada língua com base nas várias formas que estes dois sintagmas podem aparecer na organização sintática das sentenças de uma língua. Diante de uma nova sentença a ser processada, o significado das palavras e a estrutura sintática, se especificados corretamente, permitiriam a compreensão do significado da sentença como um todo.

O léxico deveria prever os significados possíveis de palavras polissêmicas, como 'quadro', através de formalismos de representação do conhecimento lexical. A gramática deveria prever as estruturas sintáticas possíveis através de formalismos de representação do conhecimento sintático. Em combinação com os modelos formais de representação do conhecimento, deveriam ser especificados também modelos de processamento deste conhecimento, ou seja, algoritmos para a realização da análise sintática. O passo subsequente é a especificação de novos formalismos de representação do conhecimento semântico e discursivo. A noção de representação do conhecimento tornou-se, portanto, o aspecto fundamental das abordagens lógicas. Este conjunto de conhecimento linguístico deve estar representado na teoria computacional e implementado sob forma de programa.

O programa deveria ser capaz, por exemplo, de distinguir a expressão 'quadro clínico' (situação) de 'quadro antigo' (obra de arte). Ambas as expressões deveriam ser interpretadas de modo diferente da expressão 'quadro abaixo' em uma página de livro. A especificação das estruturas sintáticas de sentenças como 'João viu a moça no parque com o telescópio' tornaram-se exemplos famosos das dificuldades que as línguas humanas apresentam para o processamento computacional. Sem alguma forma de processamento contextual, não é possível determinar se a análise sintática desejada é 'viu com um telescópio', 'moça com o telescópio' ou 'no parque com o telescópio'. Ainda mais grave: a dificuldade é relativamente trivial se comparada às complexidades contidas em estruturas sintáticas com seis ou sete orações ligadas por uma variedade de mecanismos de coordenação e subordinação.

Mais dificuldades desafiadoras teriam que ser enfrentadas no nível do processamento semântico. É necessário lidar com sentenças semanticamente ambíguas, como 'O Flamengo não é campeão por acaso', e escolher a interpretação apropriada para o contexto de uso. Inevitavelmente, num tratamento lógico, é necessário usar conhecimento de mundo. É preciso saber se 'o Flamengo é ou não é campeão'. Aspectos do processamento discursivo ou pragmático também precisam fazer parte da representação do conhecimento linguístico. Deste modo, uma pergunta como 'Você sabe as horas?' deve resultar numa

resposta que explicite 'que horas são', e não apenas uma resposta como 'Sei', sem nenhum complemento. Porém, a análise sintática pura não detecta esta exigência de natureza pragmática. Diante destes problemas, como também de muitos outros que não há necessidade de detalhar aqui, Elaine Rich conclui que "entender uma língua natural é difícil"[5]. É difícil, sem dúvida, para os computadores, já que crianças ainda muito jovens são perfeitamente capazes de compreender sua língua materna e comunicar-se fazendo uso desta língua materna.

As dificuldades e, parece razoável dizer, a frustração de expectativas em relação ao desempenho das máquinas no processamento de línguas humanas, não raro muito otimistas, foram gradualmente diminuindo a confiança antes depositada nas abordagens baseadas em representação do conhecimento. Não obstante, as abordagens lógicas atingiram resultados reais no âmbito dos dois objetivos da linguística computacional. Formalismos baseados em unificação, como a gramática léxico-funcional (LFG) e a gramática de estrutura de constituintes comandada pelo núcleo (HPSG), obtiveram resultados consistentes em implementações com o objetivo de realizar a unificação de representações de conhecimento linguístico. A noção de unificação tornou-se um dos avanços marcantes da história da linguística computacional. As gramáticas de unificação, sob várias formas, passaram a predominar no contexto dos esforços de base lógica para processar línguas humanas em máquinas.

A unificação consiste essencialmente em verificar se duas representações de conhecimento linguístico distintas são compatíveis. A noção de compatibilidade, neste contexto, significa que existe algum objeto para o qual ambas as descrições se aplicam. Um módulo de processamento lexical em um sistema qualquer conteria uma representação de conhecimento lexical. Esta representação poderia, por exemplo, prever os três significados de 'quadro' mencionados acima. Ao tentar unificar o sintagma nominal 'quadro grave', seria possível descartar tanto o significado 'obra de arte' quanto o significado gráfico, uma vez que somente o significado situação poderia ser aplicado a algum objeto de um mundo descrito pelo texto 'quadro grave'. A operação pode ser realizada com uso de algoritmos relativamente simples, se comparados aos necessários para lidar, por exemplo, com gramáticas gerativas de inspiração chomskyana.

No âmbito da sintaxe, é possível especificar, em uma representação, que as formas verbais do verbo 'saber' podem ser unificadas a objetos oracionais, enquanto formas verbais do verbo 'conhecer' não podem, só podem ser unificadas a objetos nominais. A especificação é inegavelmente útil em tradução de máquina. A língua inglesa possui um único verbo, *to know*, para ambos os contextos sintáticos. Em consequência, a tradução sei deveria ser escolhida sempre que estruturas como '*I know he's not coming*' tivessem que ser traduzidas. A tradução 'conheço' deve ser preferida quando a estrutura a ser traduzida for '*I know him*'. A forma 'sei ele' não funciona em português. A unificação não é aceita, portanto, quando as duas instâncias de representação especificam uma forma verbal de conhecer e um complemento oracional. Também não é aceita a unificação entre formas do verbo 'saber' e representações de objetos como pessoas, lugares, empresas, livros e mais uma variedade

[5] Elaine Rich, *Artificial Intelligence* (Singapore: McGraw-Hill, 1983).

considerável de outros tipos de objeto. A integração do aspecto semântico gera dificuldades maiores, mas é possível, pelo menos em certa medida.

Há casos em que ambas as traduções são possíveis, como em '*I know the facts*', mas muito frequentemente é preciso fazer ajustes. A maioria dos falantes de português brasileiro estaria à vontade com 'Eu conheço os fatos', mas 'Sei os fatos' pode ser inadequado em determinados contextos. O ajuste 'Sei quais são os fatos' é frequentemente necessário. É possível argumentar que 'Conheço quais são os fatos' não constitui erro. Porém, o contraste entre as formas do verbo 'saber' e as formas do verbo 'conhecer' continua indubitavelmente presente, ainda que possa ser difícil de formalizar. Por exemplo, 'Sei o caminho' e 'Conheço o caminho' parecem perfeitamente intercambiáveis. Já a combinação 'Conheço a casa' parece definitivamente mais apropriada, se comparada a 'Sei a casa'. Em última análise, ainda que possa ser uma simplificação, a separação entre os verbos com base nos complementos usados permanece útil.

A unificação e as abordagens de base lógica, de um modo geral, desenvolveram-se tendo como referência de implementação a linguagem de programação *Prolog*, ainda que a linguagem LISP também tenha sido importante no período. Há autores, como Kay, que consideram este período como um momento em que a interação entre linguistas e cientistas de computação atingiu resultados relevantes.[6] Porém, Abney aponta o processo gradual de afastamento entre os dois grupos, causando uma erosão crescente nas iniciativas conjuntas.[7] A separação entre linguistas e cientistas da computação torna-se mais radical, levando a uma divisão de trabalho estrita, que reduz a interação entre as atividades. Os linguistas tornam-se projetistas de gramática, os cientistas da computação são implementadores que transformam as gramáticas em programas. Além disso, as críticas relativas à dificuldade de superar os problemas de ambiguidade e portabilidade aumentam, com ocasionais requintes de crueldade, como a expressão 'sistemas de brinquedo' para referir-se à fragilidade dos sistemas de base lógica diante das exigências de implementações do mundo real.

No final da década de 80, dois novos elementos integraram-se à linguística computacional, com grande impacto intelectual: reconhecimento de fala e aprendizado de máquina, levando, com o uso de métodos probabilísticos, a uma mudança de paradigma nos métodos de pesquisa. A área testemunha o retorno do uso sistemático de estatística e teoria da informação,[8] (como métodos de resolver problemas de robustez dos sistemas, tanto no que diz respeito à resolução de ambiguidades quanto à portabilidade. A sofisticação, em termos de matemática, aumenta bastante. Nos anos 90, a expressão 'revolução estatística' passou a ser usada para caracterizar a mudança nos paradigmas de pesquisa da área. A expressão refere-se ao renascimento do uso de métodos probabilísticos na resolução de problemas de processamento computacional das línguas humanas, discutido na próxima seção.

[6] Martin Kay, "Zipf's Law and l'Arbitraire du Signe," *Linguistic Issues in Language Technology* 6, nº 8 (2011): 1-25.
[7] Abner, "Data-intensive".
[8] Claude Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal* 27, nº 3 (1948): 379-423.

Métodos probabilísticos em linguística computacional

Os métodos probabilísticos em questão têm como elemento fundamental coletâneas de textos digitalizados conhecidos como **corpus**. O termo aportuguesa a forma latina *corpus*. Trata-se, deste modo, de um corpo de textos. Os pesquisadores de Portugal na área frequentemente preferem a forma **corpo**. Dentro de uma abordagem empirista em linguística computacional, o corpus passou a ser o elemento básico. Isto resulta de uma filosofia analítica para a investigação das línguas humanas, resumida por Firth na frase "Diga-me com quem a palavra anda, e eu te direi quem ela é"[9]. O significado das palavras é especificado pelas demais palavras que ocorrem à sua volta, sobretudo aquelas que ocorrem com um grau de regularidade, ou seja, uma co-ocorrência frequente, possivelmente associada a um significado.

Falantes do português brasileiro, ao processar a expressão 'folha em branco', associam a co-ocorrência sistemática de 'folha', 'em' e 'branco' ao significado 'folha de papel em que não há nada escrito'. Embora seja possível imaginar um contexto em que a interpretação da co-ocorrência fosse 'folha de árvore de cor branca', a solução é altamente improvável. Expressões como 'lua de mel' são casos mais óbvios, já que praticamente nunca há variação de significado, ainda que seja possível imaginar uma situação de uso na qual houvesse. O corpus é a fonte de dados que permite que estas co-ocorrências sejam mapeadas, contabilizadas e associadas a probabilidades de interpretação. Estas tarefas, diferentemente do uso de regras lógicas para representar conhecimento linguístico, são muito mais facilmente transformadas em programas de computador. As informações assim empiricamente levantadas sustentam métodos de processamento que resolvem ambiguidades.

A ideia de um modelo probabilístico da língua não era nova. Havia sido abandonada após as críticas de Chomsky na área da linguística,[10] e de Minsky e Papert na área da inteligência artificial.[11] Conforme mencionado anteriormente, o impacto do relatório ALPAC também é um aspecto essencial da mudança de paradigma na década de 60, uma vez que os métodos empíricos não haviam apresentado resultados satisfatórios. A avaliação da ausência de fundamentos científicos como motivo da ineficiência dos sistemas de tradução de máquina da época tornou muito mais difícil obter financiamentos para pesquisas que utilizassem métodos probabilísticos. No momento, portanto, o pêndulo dos paradigmas de pesquisa afasta-se do empirismo, em busca destes fundamentos científicos.

O uso de corpus e métodos empíricos só retornaria à linguística computacional em 1989, na conferência da *Association for Computational Linguistics*, em Vancouver.[12] Os corpora da época ainda estavam longe das coletâneas gigantescas dos dias de hoje, não apenas pelas dimensões, mas também pela natureza confusa da seleção de textos incluídos. Poucos anos

[9] No original: "You shall know a word by company it keeps," John R. Firth, *Papers in Linguistics, 1934-1951* (Oxford: Oxford University Press, 1957), em 11.

[10] Noam Chomsky, (The Hague: Mouton & Co., 1957)

[11] Marvin Minsky, & Seymour Papert, *Perceptrons: An Introduction to Computational Geometry* (Cambridge [MA]: MIT Press, 1969).

[12] Cf. Adam Kilgarriff, & Gregory Grefenstette, "Introduction to the Special Issue on the Web as Corpus," *Computational Linguistics* 29, nº 3 (2003): 333-43.

depois, em 1993, o periódico *Computational Linguistics* publica um número especial sobre linguística computacional com uso de corpora de grande porte. A capacidade de armazenamento dos computadores havia aumentado consideravelmente desde os tempos da tradução russo-inglês. A redução do custo para obter máquinas com esta capacidade de armazenamento e processamento efetivo de grandes corpora também desempenhou papel determinante.

Na introdução para esse número especial, Church e Mercer (descrevem como a comunidade de pesquisa em reconhecimento de fala, neste caso fundamentalmente ligada a empresas como a IBM, concluiu que o retorno aos métodos empíricos traria ganhos substanciais para os resultados do processamento de fonemas.[13] É fácil perceber que 'atenção' e 'a tensão' soam de maneira muito parecida. Os seres humanos utilizam o contexto de fala para distinguir estas palavras. Há também muitos casos de realização alterada de palavras, como, por exemplo, em uma afirmação como 'Eu não vou', a palavra 'não', em algumas realizações, soa como 'num'. Informações sintáticas, semânticas e pragmáticas permitem que o som alterado da palavra seja reconhecido como idêntico a 'não', garantindo a interpretação correta do significado.

As técnicas de processamento baseadas em uso de conhecimento linguístico, sistematizado como representações formais, predominavam na década de 70. O grupo da IBM baseado em Raleigh, Carolina do Norte, procurava desenvolver um sistema de reconhecimento de fala segundo esta perspectiva. A entrada deste sistema convertia o sinal da fala em etiquetas que se aproximavam de fonemas, utilizando regras complexas ajustadas uma a uma. No estágio subsequente, uma gramática artificial convertia as etiquetas fonêmicas em palavras. Esta gramática, ainda segundo Church e Mercer[14], era bem pequena, cabia em uma única folha de papel. Não obstante, era usada para dar forma final à saída, apresentando o resultado sob forma de palavras. Um sistema consideravelmente intrincado de pontos a favor e contra um determinado fonema deveria corrigir imperfeições nas etiquetas, causadas por complexidades do sinal de fala semelhantes às mencionadas acima.

Conforme o relato em Church e Mercer,[15] a IBM decidiu criar um novo grupo de pesquisa em reconhecimento de fala. Este grupo foi instalado em uma localização também nova, Yorktown Heights. As mesmas concepções de reconhecimento de fala baseadas em conhecimento linguístico formalizado foram levadas de Raleigh para Yorktown Heights. A esta altura, o sistema de Raleigh era capaz de reconhecer 77% das palavras e 35% das sentenças em textos usados para testar o desempenho do sistema. Isto significa dizer que, dado o parâmetro de avaliação de uma transcrição do sinal de fala semelhante à de um foneticista, o sistema tinha um desempenho bastante inferior ao esperado. Vale dizer também que as avaliações com base em percentuais ocultam problemas às vezes muito mais graves, uma vez que o texto é uma unidade de significado. A dificuldade de reconhecer uma única palavra pode ter um impacto, do ponto de vista da compreensão deste significado, muito maior do que os percentuais revelam.

---

[13] Kenneth Church, & Robert L. Mercer "Introduction to the Special Issue on Computational Linguistics Using Large Corpora, *Computational Linguistics* 19, nº 1 (1993): 1-24.
[14] Ibid.
[15] Ibid.

O sistema de pontos a favor e contra uma determinada solução poderia ser tratado como um conjunto de probabilidades. Porém, seria necessário especificar um modelo probabilístico completo para transformar este tratamento possível em um modelo de processamento. A fundamentação científica viria, desta vez, do modelo do canal ruidoso de Shannon,[16] definido em sua teoria matemática da comunicação. A abordagem já havia sido utilizada anteriormente, como se sabe, nos sistemas de tradução de máquina das décadas de 50 e 60. O pêndulo havia atingido o seu ponto máximo na direção dos sistemas baseados em formalização do conhecimento linguístico. O processo de volta começa a tornar-se inevitável. O grupo de Yorktown Heights seria o impulso de mudança inicial.

O modelo do canal ruidoso parte do princípio da imperfeição de um sinal percebido, isto é, aquilo que o sistema 'ouve' não é o texto real produzido pelo falante. Isto ocorre, como discutido anteriormente, devido às várias interferências, tais como as exemplificadas acima. Trata-se, portanto, de estabelecer qual é a probabilidade de um determinado sinal detectado ser um outro sinal inicialmente produzido. O caminho seria criar parâmetros para um modelo da língua com base na computação de várias estatísticas a partir de uma amostra grande de textos. Estes parâmetros são usados para calcular a probabilidade de um determinado sinal original ter gerado a saída detectada pelo sistema. O treinamento de um sistema com uso de dados reais, em detrimento do uso de uma representação do conhecimento linguístico imaginada por um projetista de gramática, evolui com relativa rapidez para tornar-se um paradigma de pesquisa na área da linguística computacional.

As probabilidades condicionais que caracterizam o canal ruidoso foram computadas através de um *modelo oculto de Markov*. Um modelo de Markov é um autômato de estados finitos associado a probabilidades que determinam as transições entre os estados e controlam a emissão de símbolos de saída. Como não se conhece a sequência de transições, desde o sinal inicial até o sinal detectado, diz-se que o modelo de Markov é um modelo oculto. Em termos linguísticos, isto significa que, dada uma entrada sonora como 'Eu num vô', o sistema usa as estatísticas extraídas da amostra de fala usada no treinamento para avaliar as probabilidades das diferentes sequências possíveis e produzir aquela que melhor se aproxima da sequência de sinais detectada.

No verão de 1977,[17] os resultados do grupo de reconhecimento de fala de Yorktown atingiram 95% de correção em sentenças e 99,4% de correção em palavras. O impacto da mudança de abordagem espalhou-se por toda a comunidade de pesquisa em reconhecimento de fala. À medida em que quantidades maiores de dados gravados tornaram-se disponíveis, as técnicas de treinamento automático foram sendo aperfeiçoadas, expandido a base empírica do processamento implementado. A generalização do uso de métodos probabilísticos atinge, em seguida, o processamento linguístico voltado para a especificação das classes das palavras de um texto e a análise da estrutura sintática, estabelecendo vínculos entre palavras em agrupamentos sintaticamente motivados. O processamento semântico, voltado para a eliminação de ambiguidades lexicais e textuais, foi logo incorporado ao conjunto de questões da linguística computacional, cujo tratamento automático passou a ser implementado por meio de modelos probabilísticos. Estes

[16] Shannon, "Mathematical Theory".
[17] Church & Mercer, "Introduction".

resultados conduziram a avanços em aplicações tais como tradução de máquina e interfaces em língua humana. É possível argumentar que o modelo ideal de processamento de línguas humanas deveria combinar abordagens, de modo a aproveitar o que há de melhor em cada uma delas. Os modelos probabilísticos, por outro lado, sem dúvida vieram para ficar. A discussão seria, portanto, como combinar estes métodos a processos linguisticamente motivados, de modo a de fato incorporar elementos do processamento humano, isto é, de modo a fazer com que uma máquina de fato consiga 'entender' um texto.

Métodos híbridos

O uso de métodos híbridos constitui em combinar várias abordagens utilizadas no processamento de línguas humanas em computadores. Dado que nenhum método em isolamento é capaz de produzir um desempenho satisfatório, há motivação para produzir sistemas híbridos que possam obter resultados melhores. Além disso, é possível argumentar que os seres humanos, muito provavelmente utilizam métodos híbridos para processar língua. Por exemplo, alguma forma de dicionário mental deve existir. Este dicionário ou léxico mental precisa incorporar, sob alguma forma eficaz, a noção de co-ocorrência, uma vez que expressões tais como 'à medida em que' não podem ser processadas através da soma do significado de cada um dos elementos que a compõem. Este conhecimento seria, em última análise, probabilístico, no sentido de interpretar uma determinada combinação de palavras de uma maneira previamente aprendida por meio da experiência com a língua em uso. Porém, alguma capacidade de interpretar construções inteiramente novas também parece estar presente. Não seria possível reduzir toda a língua a padrões de uso, ainda que estes padrões certamente existam e sejam usados, possivelmente em intensidade maior do que previsto em teorias linguísticas existentes.

O processamento paralelo de subsistemas baseados em abordagens distintas seria uma maneira de aperfeiçoar e talvez aproximar-se do processamento humano. Não resta dúvida que, no estágio atual do conhecimento psicolinguístico, as informações disponíveis sobre o verdadeiro dicionário mental, ou a verdadeira gramática mental, são ainda bastante especulativas. Deste modo, a aproximação em questão não pode ser baseada em especificações claras de como é o processamento linguístico real no cérebro humano. Porém, a ação de estratégias baseadas em regras para, por exemplo, lidar com situações inesperadas ou incomuns, nas quais os padrões de uso não funcionam conforme o esperado, parece também fazer parte do processamento humano, aperfeiçoando, por exemplo, a produção de uma forma linguística final para uma resposta linguística em uma determinada situação. Uma solução de processamento computacional com base nestas possibilidades de processamento paralelo é viável.

É possível propor, também, uma hipótese de processamento em que a experiência com a língua sirva de fundamento para a produção de regras. Uma vez estabelecidas, estas regras podem ser usadas como se fizessem parte de um sistema baseado em regras independente, que poderia ter sido criado por um projetista de gramática. Porém, a robustez que o tratamento estatístico permite estaria na base das regras utilizadas. É muito provável, porém, que sistemas assim construídos tenham dificuldades de portabilidade, isto é, a

eficiência das regras geradas inicialmente pelo tratamento estatístico depende bastante do material de treinamento utilizado. Tarefas novas afastadas dos domínios incluídos no corpus de treinamento apresentariam maior dificuldade.

A despeito de soluções propostas em termos de teoria computacional, parece fundamental, neste momento, moderar as discussões às vezes acaloradas que tratam a escolha de um dos métodos básicos, isto é, com base em conhecimento linguístico ou com base em probabilidades. Conforme apontado por Hovy,[18] talvez seja mais adequado não perder de vista aspectos práticos da construção de sistemas reais. Considerando uma aplicação como tradução de máquina, a construção de um sistema que não seja apenas um brinquedo, isto é, um sistema que utilize mais de 5.000 itens lexicais para lidar com situações novas de modo robusto, tende a exigir um elemento de conhecimento estatístico. Ao mesmo tempo, uma saída linguisticamente adequada tende a exigir um módulo linguisticamente motivado que aperfeiçoe a qualidade do resultado final com ajustes relacionados ao domínio e à situação, em termos pragmáticos.

Conclusão

Desde os primórdios da evolução da área agora denominada linguística computacional, uma forma de arrogância parece ter estado presente no espírito das iniciativas de pesquisa. A área demorou bastante a se dar conta da complexidade do sistema linguístico humano e das consequentes dificuldades para capacitar máquinas 'lógicas' a realizar tarefas que envolvam línguas humanas. É possível que esta arrogância estivesse presente também nas pesquisas em linguística. A importância de princípios teóricos é indiscutível. Porém, uma das marcas registradas das investigações na área parece ser, mais recentemente, a sinalização da relativa ignorância dos pesquisadores em relação ao verdadeiro funcionamento das línguas humanas. A construção de sistemas realmente capazes de lidar com línguas humanas depende, na verdade, de avanços científicos e tecnológicos que não parecem estar ainda incorporados aos métodos de processamento conhecidos, como também não podem ser obtidos nas formulações teóricas da linguística e da psicolinguística de modo simples. Não há soluções fáceis nem ideias brilhantes que possam substituir a pesquisa sistemática e apropriadamente dimensionada na abrangência de suas conclusões.

[18] Eduard Hovy, "Deepening Wisdom or Compromised Principles? The Hybridization of Statistical and Symbolic MT Systems," *IEEE Expert* 11, nº 2 (1996): 16-8.