A Lingüística Computacional

Giselle Cardoso

Departamento de Engenharia de Computação e Automação Industrial - Faculdade de Engenharia Elétrica e de Computação - Universidade Estadual de Campinas
13.083-970 – Campinas – SP – Brazil

gcardoso@dca.fee.unicamp.br

Resumo. *A necessidade de entender os mecanismos da linguagem começou há muitos anos e ganhou uma abordagem mais formal através dos estudos de Saussure, Jakobson, Spir, Chomsky, entre outros. Na realidade, a linguagem é a única característica que separa os seres humanos dos animais, sendo, que alguns estudiosos dizem que a linguagem é nata do ser humano e outros que ela é inata. A evolução da linguagem vem trazendo muitos desafios nos seus estudos, pois há a necessidade da colaboração de várias disciplinas, sendo estas disciplinas de diversas áreas diferentes. Uma dessas disciplinas que contribuem para o estudo da sua evolução é a computação com a lingüística computacional, onde podemos encontrar inúmeras aplicações.*

1. Introdução

A busca por entender os mecanismos da linguagem iniciou-se com os primeiros estudos de gramática na Grécia antiga, ganhou uma abordagem mais formal através dos estudos de Saussure, Jakobson, Bloomfield, Spir, Whorf e desenvolveu-se notoriamente através dos estudos de Chomsky.

A utilização da linguagem é um fato real evidente que realmente separa os seres humanos dos animais. Os seres humanos são notáveis não só por terem habilidades lingüísticas em relação a outros animais, mas também por terem habilidades intelectuais significativamente mais poderosas. Atualmente muitos lingüistas, filósofos e cientistas cognitivos debatem se as habilidades lingüísticas do ser humano são natas (a linguagem é específica de um órgão no cérebro) ou inatas (resultado de uma aplicação de habilidades cognitivas).

Os pontos de vista dos nativistas é que a habilidade para usar gramática e sintaxe é um instinto, ou dependente de módulos específicos ("órgãos") do cérebro, ou ambos. O único elemento essencial à visão dos nativistas, porém, é a idéia que habilidade de linguagem está de certo modo em alguma sensação não trivial diretamente dependente da biologia do organismo humano, isso é, separável de suas adaptações cognitivas gerais.

Alguns estudiosos como Noam Chomsky e Steven Pinker defendem que as habilidades lingüísticas humanas são inatas. Outros sugerem que a competência da linguagem é um subproduto de habilidades intelectuais gerais ou uma instância de linguagem que se adapta às mentes dos seres humanos. Segundo [Perfors, 2002] há lingüistas que acreditam que a linguagem é inata e específica do cérebro e há a existência aparente de um período crítico para apreender essa linguagem.

A linguagem surge das interações de três sistemas adaptáveis: aprendizado individual, transmissão cultural, e evolução biológica ([Christiansen e Kirby, 2003]), como ilustrado na Figura 1.

A evolução da linguagem traz muitos desafios para a ciência e a necessidade de colaborações interdisciplinares. Atualmente não são os lingüistas que contribuem nestas pesquisas. A maioria os investigadores da evolução da linguagem não são da área de lingüística, vem de muitas outras disciplinas dentro das ciências cognitivas e de outras ciências. Essa natureza interdisciplinar de pesquisa da evolução da linguagem está ilustrada Figura 2. O estudo desta área é emergente e de interesse crescente com a utilização do computador para explorar assuntos pertinentes e entender a sua origem e evolução.

O artigo está organizado da seguinte forma. A próxima seção apresenta a definição

da lingüística computacional. A seção 3 aborda algumas de suas aplicações. Finalmente, a seção 4 conclui o artigo.

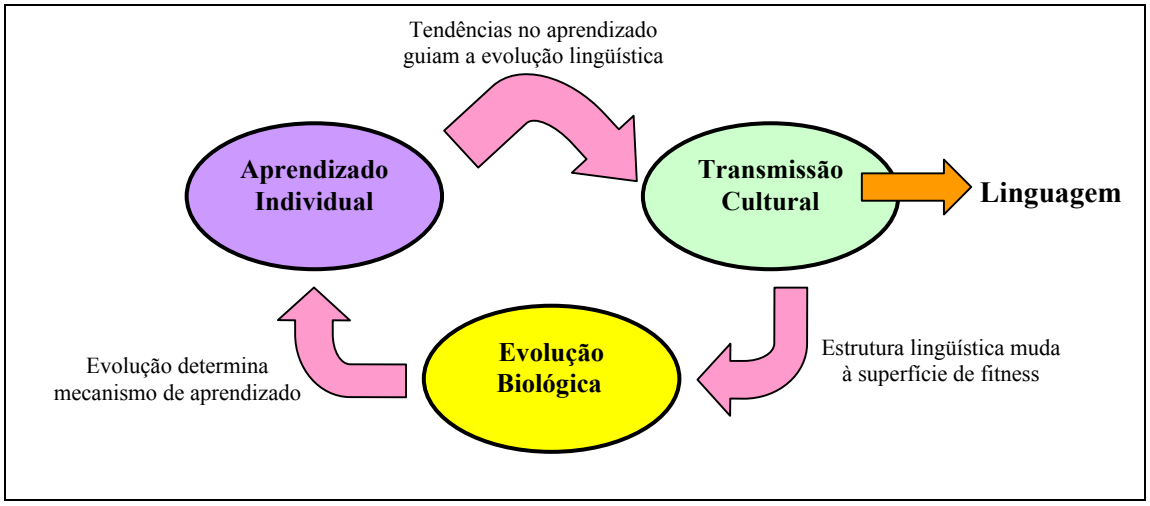

Figura 1: Modelo de Aprendizagem Iterativo (adaptado de [Christiansen e Kirby, 2003])

2. Lingüística Computacional

Com a necessidade de entender os mecanismos da língua e o interesse em desenvolver sistemas computacionais com capacidade de compreender o ser humano em sua própria linguagem, a lingüística computacional é um campo interdisciplinar que se concentra em torno do uso dos computadores para processar ou produzir a língua humana.

Segundo [Vieira, 2001] a lingüística computacional pode ser definida como uma área de conhecimento que explora as relações entre lingüística e informática, tornando possível à construção de sistemas com capacidade de reconhecer e produzir informação apresentada em linguagem natural. Ela pertence à ciência cognitiva e se sobrepõe com o campo da inteligência artificial que visa desenvolver modelos computacionais da cognição humana.

Esta lingüística trata as teorias formais sobre o conhecimento lingüístico que um ser humano necessita para gerar e compreender a língua. Quando ela é aplicada, focaliza-se no resultado prático de modelar o uso humano da língua. Os seus objetivos principais são criar produtos de software que tem algum conhecimento da língua humana e fazer com que a máquina seja capaz de processar a língua natural que falamos no dia-a-dia.

O interesse em desenvolver um sistema computacional que possuísse a capacidade de compreender os objetivos de um usuário em sua própria linguagem surgiu juntamente com os primeiros sistemas computacionais.

As pesquisas computacionais que modelam essa evolução da linguagem podem ser divididas em basicamente três categorias segundo [Christiansen e Kirby, 2003]:

(1) Avaliação: modelos computacionais, como modelos matemáticos, que tem o objetivo de forçar a formulação de explicações claras.

(2) Exploração: Podem ser usadas em simulações computacionais (com precaução) para explorar de modos gerais em quais mecanismos explicativos ou teóricos se constrói a interação. Neste modo, podem existir simulações que se dirijam para novas teorias.

(3) Exemplificação: simulações computacionais podem ser uma valiosa ferramenta por demonstrar como são os trabalhos de explicação. Eles podem multiplicar-se teorizando verbalmente e matematicamente para prover modelos para propósitos pedagógicos.

Modelos computacionais provêm uma ferramenta poderosa para o estudo de evolução da linguagem. Porém, eles devem ter ao seu lado considerações teóricas, modelos matemáticos, experimentos e coleção de dados (dados lingüístico, dados arqueológico, etc.). Por exemplo, alguns modelos computacionais podem conduzir a modelos matemáticos

[Brighton, 2002], ou vice-versa [Briscoe, 2000]. Experiências psicológicas modernas podem ser sugestionadas por modelos de computacionais [Christiansen e Ellefson, 2002] e assim por diante. Provavelmente o interesse que esses modelos aumentem mais adiante, especialmente ficando mais sofisticado em termos de mecanismos psicológicos e complexidade lingüística.

Figura 2: Evolução de Linguagem(adaptado de [Christiansen e Kirby, 2003])

Embora a matemática e os modelos computacionais possam ajudar nas discussões em relação a adaptação biológica e transmissão cultural na evolução da linguagem, tal modelo é menos provável ser capaz de enviar assuntos relacionados às origens da linguagem. Porém, evidência de outras disciplinas como arqueologia, psicolingüística neurociência cognitiva, entre outras podem prover pistas para uma resposta.

A próxima seção abordará algumas aplicações da lingüística computacional.

3. Aplicações da Lingüística Computacional

Dentro das várias modelagens computacionais existentes, muitos trabalhos se concentram na explicação simbólica do significado lingüístico, como os trabalhos de [Simon, 1980, Kintsch, 1998, Miller, 1995, Saint-Dizier e Viegas, 1995 e Lenat 1995]. Esses modelos definem o significado da palavra em termos de outros símbolos definições como as achadas nos dicionários. Segundo [Roy, 2005], tais trabalhos podem ser abordados em vários aspectos como a associação entre palavras e categorias perceptivas [Herzog e Wazinski, 1994, Plunkett *et al*, 1992, Cangelosi *et al*, 2000, Steels, 2001, Reiter e Sripada 2002 e Mojsilovic, 2005], palavras com dependência de contexto [Gardenfors, 2000, Regier, 1996, Regier e Zheng, 2003, Matsakis e Wendling, 1999], aprendizado infantil de palavras em perspectiva de primeira pessoa a partir de dados sensoriais [Pinker, 1989, Quine, 1960 e Bloom, 2000], verbos e ações físicas [Siskind, 2001, Bailey *et al*, 1997 e Narayanan, 1999], integração de ação e percepção [Arbib, 2003].

Pesquisadores descontentes com os modelos puramente simbólicos de significados de palavras buscaram sistemas que modelam o significado das palavras em termos dos referentes do mundo real. Esses sistemas provêem explicações computacionais de como palavras adquirem significados pelas conexões delas com percepção e ação, analisando principalmente, aspectos gramaticais e sociais da linguagem, como pode ser visto em [Regier, 2003].

Com o surgimento de robôs autônomos no mercado, foi necessário projetar sistemas de fossem capazes de executar diálogos flexíveis e diálogo em aberto com seus donos. Para solucionar este problema, [Steels, 2001] propôs a idéia de jogos de linguagem, que ele define como a sucessão de interações verbais entre dois agentes situados em um ambiente específico. Eles integram várias atividades requeridas para diálogo e formula frases em um contexto específico que o ajuda com possíveis significados. O diálogo requer uma combinação do processo sintático e semântico, junto com componentes da fala e do processo de imagem. Jogos de idioma integram todas as diferentes atividades necessárias para o diálogo efetivo: visão, reconhecimento de padrão, análise de fala e síntese, verbalização, interpretação, entre outras. O agente não pode somente jogar um jogo, invocar os componentes necessários para alcançar sucesso no jogo, mas ele também deve possuir algoritmos que possam ajudar o robô a preencher conceitos perdidos ou aprender o significado de novas frases da linguagem natural.

Estes jogos de linguagem podem ser classificados em jogos de imitação (estudo de fonética e fonologia), jogos de nomeação (estudos de formação de léxico), jogos de adivinhação (agentes não sabem que parte do mundo é referenciada), jogo de observação (falante e o ouvinte conhecem o tópico) e o jogo egoísta (não tem resposta se está certo ou não).

Em um experimento [Hinton e Becker, 1989], desenvolveu um procedimento computacional para mostrar como módulos no cérebro poderiam descobrir um protocolo de comunicação compartilhado sem um supervisor especificar como comunicar.

[MacLennan, 1992] propôs uma ferramenta de cooperação na comunicação, onde ao invés de estudar a comunicação entre animais no seu mundo natural, criou mundos artificiais e organismos simulados (simorgs), por ser mais fácil para a investigação. Estes simorgs são capazes de evoluir e aprendem comportamentos complexos. Eles são compostos de uma máquina de estado finito (FSM). Estas FSM possuem somente um estado de memória interna, ou seja, os simorgs não têm a habilidades de lembrar as respostas dadas por eles, estas respostas são completamente determinadas pelo incentivo do ambiente atual. A adaptação de cada simorg depende da adaptação entre eles mesmos, a cada geração de simorgs um simorg é escolhido para morrer e um novo é gerado apartir de dois simorgs já existentes. Seu aprendizado está ligado a mudança na tabela de transição da FSM.

Um experimento sobre a origem e aquisição do vocabulário referencial é apresentado por [Hutchins e Hazlehurst, 1995]. Este modelo permite explorar assuntos que envolvem interação entre grupos de agentes e suas propriedades em nível de cognição, organização social e comunicação.

Uma outra abordagem é descrita por [Oliphant e Batali, 1997], que descreve que para os membros de uma população de animais apreciar o benefício de uma troca de informação, seu comportamento comunicativo deve ser coordenado - na maioria das vezes que um animal emite a um sinal em algum tipo de situação, outro responde ao sinal de uma maneira apropriada à situação que a inspirou. Eles investigaram como uma comunicação coordenada poderia emergir entre os animais, capazes de produzir e de responder aos sinais simples, e como tal coordenação poderiam ser mantidas, quando os membros novos de uma população aprendem a se comunicar observando os outros membros. Desenvolveram um procedimento de aprendizagem que permite que um indivíduo consiga a máxima exatidão possível em se comunicar com uma população. Se todos os membros novos da população usarem este procedimento, ou uma das aproximações dele, a coordenação da comunicação da população aumentará, finalmente obtendo um sistema altamente coordenado.

O experimento de [Cangelosi, 2001] descreve tipos diferentes de modelos para a evolução da comunicação e linguagem. Ele utiliza a distinção entre sinais, símbolos, e palavras para a análise de modelos evolutivos da linguagem. Em particular, mostra como podem ser usadas técnicas de computação evolutivas como vida artificial para estudar o

aparecimento de sintaxe e símbolos de sinais de comunicação simples.

O trabalho de [Vogt e Coumans, 2003] investiga como as diferentes modalidades de interações sociais influenciam a evolução da linguagem, comparando três modelos de jogos de linguagem que diferem no tipo de interações que sociais se usam.

[Loula *et* al, 2003] propôs, baseado na semiótica de Peirce e inspirado por um caso etnológico, uma metodologia para simular uma comunicação simbólica de um predador com criaturas artificiais em um mundo virtual de eventos predatórios. A fim construir esse eco-sistema digital, foi examinado o exemplo de semiose em macacos vervet do leste africano.

Acima foram apresentados alguns experimentos que analisam a linguagem, mas atualmente pode-se encontrar inúmeros trabalhos nesta área.

4. Conclusão

Este trabalho apresentou a lingüística, como ela surgiu e como está a evolução nos seus estudos, pois ela se tornou uma área na qual inúmeras disciplinas têm influenciado no seu estudo. Uma das áreas que tem influenciado muito na pesquisas da evolução da linguagem é área de computação, com a lingüística computacional. Atualmente, existem várias aplicações computacionais na área de lingüística, que contribuem muito para a evolução dos estudos da linguagem. Foram, também, abordadas algumas das inúmeras aplicações da lingüística computacional existentes atualmente.

6. Referências

Arbib, M.A. (2003), "Schema theory", In The Handbook of Brain Theory and Neural Networks (2nd edn) (Arbib, M.A., ed.), pp. 993–998, MIT Press.

Bailey, D., Chang, N., Feldman, J. e Narayanan, S. (1997), "Embodied lexical development", In Proceedings of the 19th Annual Meeting of the Cognitive Science Society.

Bloom, P. (2000), "How Children Learn the Meanings of Words", MIT Press.

Brighton, H. (2002), "Compositional syntax from cultural transmission", Artificial Life, Vol. 8, pp. 25–54.

Briscoe, E.J. (2000), "Evolutionary perspectives on diachronic syntax", In: Diachronic Syntax: Models and Mechanisms (Pintzuk, S., Tsoulas, G. and Warner, A., eds), pp. 75-108, Oxford University Press.

Cangelosi, A. (2001), "Evolution of Communication and Language using Signals, Symbols and Words", IEEE Transactions on Evolutionary Computation, v. 5, n.2, pp. 93-101.

Cangelosi, A., Greco, A. e Harnad, S. (2000), "From robotic toil to symbolic theft: Grounding transfer from entry-level to higher-level categories", Connection Science, Vol. 12, pp. 143–162.

Christiansen, M.H. e Ellefson, M.R. (2002), "Linguistic adaptation without linguistic constraints: the role of sequential learning in language evolution", In Transitions to Language (Wray, A., ed.), pp. 335-358, Oxford University Press.

Christiansen, M.H. e Kirby, S. (2003), "Language Evolution: Consensus and Controversies", Trends in Cognitive Sciences, Vol. 7. nº 7, pp. 300-307.

Gardenfors, P. (2000), "Conceptual Spaces: The Geometry of Thought", MIT Press.

Herzog, G. e Wazinski, P. (1994), "VIsual TRAnslator: Linking Perceptions and Natural Language Descriptions", Artificial Intelligence Review, Vol. 8 nº 2/3, pp. 175-187.

Hinton, G., e S. Becker (1989), "An unsupervised learning procedure that discovers surfaces in random-dot stereograms", Unpublished manuscript, University of Toronto, Canada.

Hutchins, E. e Hazlehurst, B. (1995), "How to Invent a Lexicon: The Development of Shared Symbols in Interaction", Artificial Societies: The Computer Simulation of Social Life.

Kintsch, W. (1998), "Comprehension: A Paradigm for Cognition", Cambridge University Press.

Lenat, D. (1995), "CYC: A large-scale investment in knowledge infrastructure", Communications of the ACM, Vol. 38, nº 11, pp. 32-38.

Loula, A., Gudwin, R. e Queiroz J. (2003), "Synthetic Approach of Symbolic Creatures", S.E.E.D. Journal (Semiotics,

Evolution, Energy, and Development), Vol. 3, nº 3, pp. 125-133.

MacLennan, B. (1992), "Synthetic Ethology: An Approach to the Study of Communication", Artificial Life II: The Second Workshop on the Synthesis and Simulation of Living Systems.

Matsakis, P. e Wendling, L. (1999), "A new way to represent the relative position between areal objects", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 21, nº 7, pp. 634-643.

Miller, G. (1995), "Wordnet: A lexical database for English", Communications of the ACM, pp 39-41.

Mojsilovic, A. (2005), "A computational model for color naming and describing color composition of images", IEEE Transactions on Image Processing, Vol. 14, pp.690–699.

Narayanan, S. (1999), "Moving right along: A computational model of metaphoric reasoning about events", In Proceedings of the National Conference on Artificial Intelligence and the eleventh Innovative applications of artificial intelligence conference innovative applications of artificial intelligence, pp. 121 – 127.

Oliphant, M. e Batali, J. (1997), "Learning and the emergence of coordinated communication", The newsletter of the Center for Research in Language, Vol. 11, nº 1.

Perfors, A. (2002) "Simulated Evolution of Language: a Review of the Field", Journal of Artificial Societies and Social Simulation, Vol. 5, nº. 2.

Pinker, S. (1989), "Learnability and Cognition", MIT Press.

Plunkett, K., Sinha, C., Moller, M. F., & Strandsby, O. (1992), "Symbol grounding or the emergence of symbols? Vocabulary growth in children and a connectionist net", Connection Science, Vol. 4, nº 3 e 4, pp. 293–312.

Quine, W.V.O. (1960), "Word and Object", MIT Press.

Regier, T. (1996), "The Human Semantic Potential", MIT Press.

Regier, T. (2003), "Emergent constraints on word-learning: A computational perspective", Trends in Cognitive Sciences, Vol. 7, 263-268.

Regier, T. e Zheng, M.M. (2003), "An attentional constraint on spatial meaning", In Proceedings of the 25th Annual Meeting of the Cognitive Science Society.

Reiter, E. e Sripada, S. (2002), "Human variation and lexical choice", Computational Linguistics, Vol. 28, pp. 545-553.

Roy, D. (2005), "Grounding words in perception and action: computational insights",Trends in Cognitive Science, Vol. 9, nº 8, pp.389-396.

Saint-Dizier, P. e Viegas, E., eds (1995), "Computational Lexical Semantics", Cambridge University Press.

Simon, H. (1980), "Physical symbol systems", Cognitive Sciences, Vol. 4, pp. 135-183

Siskind, J. (2001), "Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic", Journal of Artificial Intelligence Research, Vol.15, pp. 31-90.

Steels, L. (2001), "Language games for autonomous robots", IEEE Intelligent systems, pp. 16-22.

Vieira, R. e Lima, V.L.S. (2001), "Lingüística computacional: princípios e aplicações", In: As Tecnologias da Informação e a questão social: anais. Carlos Eduardo Ferreira (Ed.), SBC, Vol. 2, pp. 47-88, Fortaleza, Ceará, Brasil.

Vogt, P. e Coumans, H. (2003), "Investigating social interaction strategies for bootstrapping lexicon development", Journal of Artificial Societies and Social Simulation, Vol. 6, nº 1.