Diretrizes para novos algoritmos de aprendizado de sintaxe não supervisionados

Glauber José Vaz

Unicamp – Faculdade de Tecnologia
glauber@ceset.unicamp.br

Resumo – Este artigo fornece diretrizes para se implementar um novo algoritmo não supervisionado a fim de descobrir estruturas hierárquicas em dados seqüenciais. Sua entrada é formada por uma seqüência de textos em língua natural e sua saída corresponde à gramática subjacente à entrada. Este algoritmo pode dar explicações de como a linguagem é aprendida pelo ser humano e também resolver problemas que requerem estruturações hierárquicas a partir de dados seqüenciais, como agrupamento, classificação e descoberta de regras de associação.

Abstract – *This paper provides directions to implement a new unsupervised algorithm in order to discover hierarchical structures in sequential data. Its input is a natural language text sequence and its output constitutes the underlying grammar. This algorithm can explain how language is learned by human beings and also solve problems that require hierarchical structures from sequential data, like clustering, classification and association rules discovery.*

Palavras-chave: evolução de sintaxe, indução de gramática, aprendizado não supervisionado, lingüística computacional.

1. Introdução

Este trabalho estuda o aprendizado de sintaxe com a construção de um modelo computacional que, a partir de corpora de entrada, extrai regras gramaticais que regem a linguagem considerada. Assim, procura-se investigar se é possível construir um algoritmo que simule os processos de inferência de regras gramaticais utilizados pelo ser humano e também as pré-condições necessárias para se aprender sintaxe a partir da experiência. A proposta do modelo computacional é inspirada principalmente nos trabalhos de Kirby (2000,2002) e de Solan et al (2005a).

Qualquer cientista que pretenda explicar a evolução da linguagem precisa explicar, segundo Perfors (2002, 4.2), dois saltos principais na evolução: o primeiro uso de palavras como símbolos e o primeiro uso da gramática, a que ela se refere, respectivamente, por evolução da comunicação e evolução da sintaxe. A primeira está intimamente ligada ao problema de *symbol grounding* (Harnad, 1990; Taddeo e Floridi, 2005) e a questões semânticas. Já a evolução da sintaxe, objeto de estudo deste trabalho, pode ser estudada, de acordo com algumas linhas de pesquisa, sem maiores preocupações com outros aspectos da linguagem, como a semântica e o léxico.

Kirby (2002, p. 1) afirma que a sintaxe fornece a habilidade de produzir uma quantidade infinita de expressões através da composicionalidade e da recursão. Ele define a composicionalidade como a propriedade de que o significado das expressões é função dos significados de suas partes e da maneira como são colocadas juntas, e a recursão como a propriedade de linguagens com léxico finito e conjunto de regras, em que os constituintes de uma expressão podem ter constituintes da mesma categoria.

Perfors (2002, 1.3) afirmara que a corrente dominante na lingüística era de que muito do conhecimento lingüístico humano é inato, e que os aspectos mais significativos da linguagem evoluem por meio da evolução biológica clássica e não simplesmente pela cultura. Entretanto, muitos trabalhos têm defendido a idéia de que a evolução da linguagem dá-se predominantemente através da transmissão cultural e não por evolução biológica, e que mecanismos de aprendizado da linguagem dependentes da experiência são mais relevantes do que os mecanismos inatos, não dependentes da experiência. Kirby (2002, p. 2) defende que as propriedades estruturais da linguagem emergem ao longo do tempo através dos complexos processos dinâmicos de transmissão social e que se deve procurar cuidadosamente por processos alternativos antes de se utilizar a seleção natural para explicar a evolução da linguagem. Seidenberg (1997, p. 1603) afirma que não se pode determinar o que é inato na linguagem sem a exploração do papel da experiência até seu limite. Portanto, parece haver consenso de que algumas pré-adaptações ocorreram no ser humano para que fosse possível a emergência da linguagem (Christiansen e Kirby, 2003, p. 2). No entanto, saber quais são estas pré-adaptações necessárias e

suficientes para o surgimento da linguagem continua sendo um tema de pesquisa.

Algumas das evidências consideradas nesta visão não-nativista estão ligadas ao desenvolvimento da linguagem em crianças. Por exemplo, as palavras e frases usadas com maior freqüência pelos pais serão – com probabilidade relativamente alta – as primeiras palavras, frases e estruturas gramaticais aprendidas pela criança (Perfors, 2002, 3.13). Além disso, crianças apenas generalizam uma regra ou estrutura depois de terem sido expostas a elas muitas vezes e de diferentes maneiras (Perfors, 2002, 3.13).

Assim, muitos estudos têm explorado aspectos estatísticos e probabilísticos da linguagem (Saffran et al, 2001), muitas vezes combinados a representações de regras gramaticais (Seidenberg et al, 2002; Peña et al, 2002; Niyogi e Berwick, 2009; Solan et al, 2005a) e em modelos conexionistas (Seidenberg, 1997; Morris et al, 2000).

Em relação a estas abordagens que relacionam estatísticas a gramáticas no aprendizado da linguagem, Seidenberg et al (2002) fazem vários questionamentos ainda não esclarecidos:
- O aprendizado gramatical começa quando o aprendizado estatístico termina?
- Que tipo de estatísticas as pessoas são capazes de computar?
- Que tipo de informação pode ser aprendido por estatísticas?
- Qual é a diferença entre uma regra e uma generalização estatística?
- Existe um procedimento realizado pelas crianças de maneira que elas encontram as regras corretas sob certas condições de aprendizagem realistas?

Peña et al. (2002) afirmam que para aprender linguagem, são necessárias computações estatísticas para identificar palavras em discurso e computações de formas algébricas para descobrir estruturas gramaticais, sugerindo então, que computações estatísticas são poderosas em um nível mais baixo, mas não são suficientes em um nível mais alto, para obter generalizações e estruturas na forma de regras. A maior dificuldade, portanto, parece estar na tentativa de se explicar a maneira como o ser humano deduz regras gramaticais a partir de suas experiências lingüísticas.

Seidenberg et al (2002) afirmam que representaria um progresso substancial uma teoria que, dadas as complexidades das experiências, explicasse como são aprendidas as regras que fazem as generalizações corretamente.

Este trabalho procura ajudar na compreensão de como são extraídas estas regras pelos seres humanos, estudando modelos computacionais bem sucedidos e apontando caminhos para novos modelos.

1.1. Aprendizado individual

Christiansen e Kirby (2003, p. 302) esclarecem que a evolução de linguagem envolve três sistemas adaptativos: evolução biológica, transmissão cultural e aprendizado individual. Portanto, é possível lidar com este tema em três diferentes escalas de tempo: tempo de vida de um indivíduo, de uma linguagem e de uma espécie.

Neste trabalho, considera-se o aprendizado individual da linguagem. Espera-se que dentre as três abordagens, a mais fácil de se obter melhores resultados seja a de aprendizado individual, por vários motivos:
- Pode-se considerar que já existe uma linguagem estabelecida,
- E que é possível aprendê-la com indivíduos que já têm domínio sobre a linguagem;
- Não há necessidade de se entender a origem da linguagem no ser humano ou as adaptações biológicas necessárias para se compreender como um indivíduo aprende a linguagem;
- Há uma fonte de inspiração mais concreta para possíveis teorias e experimentos: as crianças;
- Têm-se parâmetros mais claros de comparação para as simulações computacionais, pois é possível comparar simulações computacionais com o aprendizado de crianças, e as escalas de tempo de processamento são mais facilmente comparáveis com o aprendizado individual da linguagem do que com as escalas de tempo consideradas na evolução de uma linguagem ou de uma espécie.

Além disso, é possível que a melhor compreensão sobre o aprendizado individual de uma linguagem por um ser humano dê pistas de como surgiram as línguas naturais utilizadas atualmente e sobre como o homem evoluiu a ponto de desenvolver habilidades lingüísticas.

Neste contexto de aprendizado individual da linguagem, Saffran et al (2001, p. 12874) destacam três linhas de pesquisa:
- Como as crianças conseguem encontrar as palavras em uma seqüência acústica que corresponde à entrada para o aprendizado da linguagem;
- Como as crianças adquirem a habilidade para rapidamente combinar elementos lingüísticos e determinar as relações entre eles;

- Como as crianças impõem estrutura gramatical na entrada percebida.

O primeiro item não é foco deste trabalho, uma vez que se pretende trabalhar com textos escritos, que já apresentam a divisão das palavras através de elementos como espaços em branco e pontuação. Peña et al (2002) afirmam que se os itens de um corpus estão delimitados, os seres humanos são capazes de extrair estruturas sintáticas. No entanto, para isso, precisam ser dotados de uma capacidade diferente da computação estatística. Esta capacidade de extrair regularidades em um nível mais alto pode ser investigada através de modelos computacionais.

2. Modelos Computacionais

A utilização de modelos computacionais com o objetivo de explicar aspectos da evolução da linguagem já é amplamente aceita. Atualmente, é comum o uso deste recurso, que representa, segundo Perfors (2002, 6.2), um meio termo entre as teorias abstratas e as rigorosas abordagens matemáticas, e permite a avaliação de quais fatores são importantes e sob quais circunstâncias.

Cangelosi (2001, p. 94-95) enumera três tipos de modelos computacionais para a evolução de linguagem: um que se baseia simplesmente na associação entre símbolos com objetos do mundo real, um que se baseia nas relações entre símbolos e objetos e nas relações entre os símbolos, e um terceiro que relaciona apenas símbolos. Este último, utilizado normalmente para o estudo da evolução da sintaxe, é a abordagem adotada neste trabalho. Uma de suas vantagens é que não é necessário lidar com o problema de *symbol grounding*.

Os principais modelos considerados neste trabalho são aqueles propostos por Kirby (2000, 2002) e Solan et al (2005a). São destacados a seguir os principais aspectos destes modelos.

2.1. Modelo de Kirby

De acordo com Kirby (2002, p. 16, 20), a linguagem na população evolui de um vocabulário idiossincrático associado a significados complexos para uma sintaxe composicional com categorias nominais e verbais. Isto ocorre porque linguagens que são transmitidas mais facilmente de geração para geração persistem, de maneira que regras mais genéricas são priorizadas.

O estado final inevitável desse processo de evolução é uma linguagem com uma sintaxe composicional e recursiva. A gramática para esta linguagem parece ser a menor em termos de números de regras capaz de expressar todo o espaço de significados (Kirby, 2002, p. 22).

O modelo de Kirby (2000, 2002) é extremamente limitado. Cada significado na simulação é uma tripla envolvendo Agente, Paciente e Predicado, e o conjunto de possíveis valores são divididos em classes de Objetos, para Agentes e Pacientes, ou Ações, para Predicados, em um mundo que conta com apenas 5 Objetos e 5 Ações, o que resulta em apenas 100 possibilidades de significados em uma linguagem não recursiva e que não possui relações reflexivas (relações em que agente e paciente são o mesmo objeto).

No entanto, este modelo apresenta características interessantes:
- Cada indivíduo – o algoritmo é populacional – começa com uma gramática vazia.
- Cada indivíduo representa seu sistema de comunicação como uma gramática livre de contexto, o que permite a expressão tanto de sistemas completamente não comoposicionais quanto de sistemas altamente composicionais. (Kirby, 2000, p. 5)
- O processo de aprendizado ocorre por um modelo de indução de sintaxe que envolve as operações de inclusão de regras não composicionais à gramática, de maneira a acrescentar novos fatos experienciados pelo indivíduo, e de junção de regras existentes na gramática para criar generalizações (Kirby, 2000, p.7). Além disso, com a possibilidade de se ter predicados como argumentos de outros predicados, Kirby (2002, p. 17) conseguiu um modelo em que a recursão também emerge.

Além da evolução da sintaxe, Kirby trata o problema de associar símbolos a significados, o que não é considerado pelo presente trabalho.

2.2. Modelo de Solan et al

Solan et al (2005a) usam corpora de dados brutos, simbólicos e seqüenciais para inferir as regras subjacentes que governam sua produção. Estes dados podem ser tanto textos em língua natural, como seqüências de proteínas, por exemplo. O algoritmo implementado neste trabalho, o ADIOS (*automatic distillation of structure*), descobre padrões estruturados hierarquicamente com uma abordagem não supervisionada. Ele se baseia em métodos de extração de padrões e generalização estruturada, dois processos envolvidos na aquisição de linguagem. Portanto, mais do que descobrir gramáticas para

linguagens, um algoritmo como este pode ser utilizado em problemas de mineração de dados, envolvendo, por exemplo, regras de associação, agrupamentos ou classificações.

O algoritmo começa carregando o corpus em um grafo direcionado e não simples, cujos vértices são entradas léxicas. Cada sentença define um caminho que começa em um vértice com o símbolo especial de início e termina em um vértice com o símbolo especial de fim. Depois do carregamento do grafo, é feita uma busca iterativa para encontrar padrões significativos.

ADIOS utiliza informações estatísticas presentes nos dados e extrai regularidades na forma de regras, o que dá suporte a uma generalização estruturada. Apesar de estas características já estarem presentes em outros trabalhos, os autores afirmam que o algoritmo foi o primeiro a apresentar simultaneamente certas características cruciais.

ADIOS obteve sucesso com gramáticas artificiais, com corpora de língua natural, e ainda na extração de estruturas sintáticas de seqüências de proteínas. Além disso, apresentou resultados superiores aos modelos estatísticos padrões de linguagem, como os baseados em probabilidades *n-gram*, uma vez que estes modelos apresentam estimativas não confiáveis para eventos raros ou não vistos, e porque modelos com *n* pequeno não conseguem capturar dependências entre palavras mais distantes entre si (Solan et al, 2005b).

Outras vantagens apontadas pelos autores do algoritmo são a fácil escalabilidade para corpora maiores e a complexidade computacional estimada, que, empiricamente, no caso médio, cresce linearmente com o tamanho do corpus (Solan et al, 2005b).

Apesar de seu sucesso, o algoritmo não pode ser considerado definitivo, uma vez que apresenta características que podem ser melhoradas. Por exemplo:

- Cada padrão é estruturado na forma de árvore, o que elimina a possibilidade de recursão de padrões. Apesar de Solan et al (2005, p. 11630) afirmarem que é possível introduzir recursão em um estágio pós-processamento, a recursão infinita não está implementada na versão corrente do algoritmo. Considerando-se a importância da recursão nas gramáticas, deve-se considerar a implementação de um novo algoritmo para reconhecimento de padrões e/ou a utilização de representações mais adequadas.
- Cada descoberta de padrão gera um agrupamento de unidades léxicas no grafo, de maneira que unidades originalmente distantes podem se aproximar entre si e tornar possível que o algoritmo encontre dependências de longo alcance entre estas unidades. No entanto, é possível que duas unidades dependentes entre si nunca sejam verificadas se não ocorrer agrupamentos suficientes para reduzir a distância entre elas. Isso acontece porque o algoritmo utiliza um parâmetro *L* que estabelece a largura da janela de contexto onde classes de equivalências são buscadas. Assim, apesar de conseguir extrair dependências entre palavras mais distantes do que os modelos *n-gram*, na prática há um limite para estas distâncias. Devem ser considerados, portanto, novos algoritmos para que dependências entre unidades da mesma sentença, independentemente das distâncias envolvidas, possam ser sempre capturadas.
- O grafo final gerado pelo ADIOS inclui tantos caminhos quanto o número de sentenças originais, mas ainda podem ser gerados muitos novos caminhos devido à descoberta de padrões (Solan et al, 2005 p. 11630). Da mesma maneira que as sentenças são conservadas no grafo, os padrões construídos são irreversíveis. Assim, todos os padrões extraídos e as sentenças da entrada ficam em memória. Em uma aplicação que requer um fluxo contínuo de sentenças, é necessário haver uma política de descarte daqueles elementos que parecem não ser importantes ou não são muito utilizados. Esta abordagem parece ser mais compatível com os métodos de aprendizado de um ser humano, que recebe continuamente seqüências de palavras e normalmente esquece sentenças que são irrelevantes, não lhe fazem sentido ou não possuem uma estrutura sintática aparente.
- O algoritmo não manipula explicitamente gramáticas livres de contexto, mas grafos que podem ser vistos como um conjunto de regras gramaticais. Talvez, manipulando diretamente regras gramaticais, seja possível descobrir padrões de maneira mais rápida do que na estrutura de grafo ou padrões que seriam impossíveis de se descobrir nesta estrutura.
- Apesar de apresentar resultados excelentes com gramáticas livres de contexto simples, o ADIOS ainda pode ser melhorado quando trata o aprendizado de gramáticas livres de contexto complexas. Além disso, como o algoritmo é guloso, a ordem das sentenças no conjunto de treinamento influi o resultado. Isto não é desejável se a gramática geradora das sentenças é a mesma.

Estes são alguns dos aspectos que podem ser melhorados em novas abordagens para o problema.

3. Diretrizes para novos algoritmos

Como não se pode estabelecer que há uma solução definitiva para o problema considerado, é necessário tentar encontrar novos algoritmos que busquem resultados melhores do que os alcançados pelos modelos já existentes. No entanto, as idéias essenciais dos principais modelos devem continuar sendo utilizadas. Assim, algumas das características que podem levar a melhores resultados são as seguintes:

- O sistema deve usar uma representação de gramáticas livres de contexto, que é iniciada com uma gramática vazia. Se os padrões são representados por gramáticas em vez de árvores, torna-se mais fácil descobrir propriedades recursivas destes padrões.
- O processo de aprendizado deve ocorrer por um modelo de indução de sintaxe que envolve operações de inclusão de regras e de junção de regras existentes na gramática para criar generalizações.
- Assim como o algoritmo de Solan et al (2005a, p. 1), o corpus de entrada deve ser carregado em um grafo cujos vértices são unidades léxicas e cujas arestas ligam as unidades que aparecem em seqüência no texto. No entanto, deve haver em cada nó deste grafo, uma lista de triplas indicando um código da sentença em que aparece a palavra, um ponteiro para a palavra anterior e um ponteiro para a palavra posterior nesta sentença. Desta maneira, não se perde dependências entre unidades léxicas, mesmo que elas estejam distantes entre si.
- Devem ser considerados testes em que haja fluxo contínuo e ininterrupto de seqüências de dados, o que corresponde a problemas envolvendo mineração de *data streams*, que são definidos por Guha et al (2003) como seqüências ordenadas de pontos que devem ser acessadas em ordem e podem ser lidas apenas uma vez ou poucas vezes.
- Os dois últimos itens levam a um problema que normalmente é ignorado nos modelos anteriores: a capacidade de memória. Uma vez que ela pode ser esgotada, vários aspectos importantes devem passar a ser considerados, como por exemplo, políticas sobre o que deve ser descartado nos momentos em que é necessário liberar espaço.
- Os algoritmos devem ser testados com textos em língua natural e outros tipos de dados para se atacar tanto o problema de evolução da sintaxe quanto problemas de reconhecimento de padrões.
- Conforme já exposto, Saffran et al (2001) destacou a importância das pesquisas na separação de dados em unidades léxicas, a determinação de relações entre estas unidades e, por fim, a extração da gramática que rege os dados. Um possível foco de pesquisa para um novo algoritmo é determinar se a partir da seqüência de unidades léxicas, já se parte para descobrir regras gramaticais ou se passa por uma fase intermediária para relacionar essas unidades. No contexto da evolução da linguagem no ser humano, estas três etapas estão relacionadas à compreensão de palavras, ao estabelecimento de regras sintáticas simples, e ao aprendizado de regras gramaticais de mais alto nível. Fases intermediárias aparecem, por exemplo, nos trabalhos de Morris et al (2000), com os conceitos de mini-gramáticas e com os estágios de evolução, e de Kirby (2000) com os três estágios que marcam diferentes padrões de comportamentos dos indivíduos. Estas fases podem simplesmente emergir durante o processo de evolução de sintaxe, mas também podem ser determinadas por diferentes algoritmos que são executados em diferentes períodos do aprendizado da linguagem. Isto ainda não foi bem explorado pelos modelos considerados e pode esclarecer se os processos envolvidos para adquirir regras sintáticas mais simples são os mesmos envolvidos para aprender as regras gramaticais mais elaboradas.

4. Conclusão

Apesar do sucesso já alcançado por alguns algoritmos, como o ADIOS, não se pode considerar que o problema de se descobrir estruturas hierárquicas em dados seqüenciais com abordagem não supervisionada tenha sido resolvido de maneira definitiva. Novos algoritmos devem ser criados com o objetivo de se conseguir resultados ainda melhores do que os obtidos pelos modelos existentes. Este problema é importante não só para se entender melhor a evolução da sintaxe, no domínio da lingüística, mas também para a resolução de problemas computacionais envolvendo reconhecimento de padrões.

Uma vez que este novo algoritmo consiga gerar gramáticas de língua natural adequadamente, criam-se fortes evidências de que a sintaxe pode ser de fato estudada separadamente de outros aspectos da linguagem e de que o conjunto de pré-adaptações necessárias para o aprendizado de sintaxe seja formado pela capacidade de identificação de unidades léxicas através de sinais que separam estas unidades e que identificam início e fim de sentença, pela capacidade de representar gramáticas e pelos processos de computação estatística e de regras gramaticais.

Com o sucesso desta proposta, novas fronteiras de pesquisa podem ser consideradas:

- Tratamento do nível léxico no algoritmo, de maneira que seja possível considerar não somente

palavras inteiras como unidades, mas também prefixos, sufixos e radicais de palavras.
- Desenvolver e aplicar algoritmos baseados na evolução da sintaxe para resolver problemas reais específicos, fora do domínio da lingüística.
- Implementar mecanismos de aprendizado formal através dos quais seja possível ensinar ao sistema regras gramaticais diretamente, assim como as crianças aprendem gramática na escola. Este aprendizado deve adaptar as regras gramaticais já conhecidas às regras que são aprendidas. Desta maneira, tem-se um mecanismo de aprendizado em um nível mais alto.

Referências Bibliográficas

Cangelosi, A. (2001) Evolution of communication and language using signals, symbols, and words. *IEEE Transactions on Evolutionary Computation*, vol. 5, no. 2. p. 93-101. Disponível em <http://www3.isrl.illinois.edu/~junwang4/langev/localcopy/pdf/cangelosi01evolutionOf.pdf >.

Christiansen M.H. e Kirby, S. (2003) Language evolution: consensus and controversies. *Trends in Cognitive Sciences*, vol. 7, no. 7, p. 300-307. Disponível em <http://www3.isrl.illinois.edu/~junwang4/langev/localcopy/pdf/christiansen03trends.pdf >.

Guha, S. et al. (2003) Clustering data streams: Theory and Practice. *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 3.

Harnad, S. (1990) The Symbol Grounding Problem. *Physica D*, vol. 42, p. 335-346. Disponível em <http://cogprints.org/3106/>.

Kirby, S. (2000) Syntax without natural selection: how compositionality emerges from vocabulary in a population of learners. C. Knight, editor, *The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form*, p. 303-323. Cambridge University Press. Disponível em <http://www3.isrl.illinois.edu/~junwang4/langev/localcopy/pdf/kirby00syntaxWithout.pdf>.

Kirby, S. (2002) Learning, bottlenecks and the evolution of recursive syntax. T. Briscoe, editor, *Linguistic Evolution through Language Acquisition: Formal and Computational Models*. Cambridge University Press. Disponível em <http://www3.isrl.illinois.edu/~junwang4/langev/localcopy/pdf/kirby02learningBottlenecks.pdf>.

Morris, W.C., Cottrell, G.W. e Elman, J.L. (2000) A connectionist simulation of the empirical acquisition of grammatical relations. *Hybrid Neural Systems*, Springer-Verlag. Disponível em <http://crl.ucsd.edu/~elman/Papers/morris.pdf>

Niyogi, P. e Berwick, R.C. (2009) The proper treatment of language acquisition and change in a population setting. *Proceedings of the National Academy of Sciences (PNAS)*, vol. 106, no. 25, p. 10124-10129. Disponível em <http://www.pnas.org/content/106/25/10124>.

Perfors, A. (2002). Simulated evolution of language: a review of the field. *Journal of Artificial Societies and Social Simulation*, vol. 5, no. 2. Disponível em <http://jasss.soc.surrey.ac.uk/5/2/4.html>.

Peña, M et. al. (2002) Signal-driven computations in speech processing. *Science*, vol. 298, no. 5593, p. 604-607.

Saffran, J.R., Aslin, R.N. e Newport, E.L. (1996) Statistical learning by 8-month-old infants. *Science*, vol. 274, no. 5294, p. 1926-1928.

Saffran, J.R., Senghas, A. e Trueswell, J.C. (2001) The acquisition of language by children. *Proceedings of the National Academy of Sciences (PNAS)*, vol. 98, no. 23, p. 12874-12875. Disponível em <http://www.pnas.org/content/98/23/12874>.

Seidenberg, M.S. (1997) Language acquisition and use: learning and applying probabilistic constraints. *Science*, vol. 275, no. 5306, p. 1599-1603.

Seidenberg, M.S., MacDonald, M.C., Saffran, J.R. (2002) Does grammar start where statistics stop? *Science*, vol. 298, no. 5593, p. 553-554.

Solan, Z et. al. (2005) Unsupervised learning of natural languages. *Proceedings of the National Academy of Sciences (PNAS)*, vol. 102, no. 33. p. 11629–11634. Disponível em <http://www.pnas.org/content/102/33/11629>.

Solan, Z et. al. (2005) Unsupervised learning of natural languages. *Supporting Text*. Disponível em<http://www.pnas.org/content/suppl/2005/08/02/0409746102.DC1/09746SuppText.pdf>

Taddeo, M. e Floridi, L. (2005) Solving the symbol grounding problem: a critical review of fifteen years of research. *Journal of Experimental and Theoretical Artificial Intelligence*, vol. 17, p. 419-445.