Pablo Picasso Feliciano de Faria

UM MODELO COMPUTACIONAL DE AQUISIÇÃO DE PRIMEIRA LÍNGUA

CAMPINAS,
2013

UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE ESTUDOS DA LINGUAGEM

Pablo Picasso Feliciano de Faria

UM MODELO COMPUTACIONAL DE AQUISIÇÃO DE PRIMEIRA LÍNGUA

Tese de doutorado apresentada ao Instituto de Estudos da Linguagem da Universidade Estadual de Campinas para a obtenção do Título de Doutor em Lingüística.

Orientadora:
Profa. Dra. Ruth Elisabeth Vasconcellos Lopes
Coorientadora:
Profa. Dra. Charlotte Marie Chamberlland Galves

CAMPINAS,
2013

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Estudos da Linguagem
Teresinha de Jesus Jacintho - CRB 8/6879

Faria, Pablo, 1978-
F225m Um modelo computacional de aquisição de primeira língua / Pablo Picasso Feliciano de Faria. – Campinas, SP : [s.n.], 2013.

Orientador: Ruth Elisabeth Vasconcellos Lopes.
Coorientador: Charlotte Marie Chamberlland Galves.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Estudos da Linguagem.

1. Aquisição de linguagem. 2. Linguística computacional. 3. Gramática gerativa. 4. Aprendizado do computador. 5. Linguagens formais. I. Lopes, Ruth Elisabeth Vasconcellos,1960-. II. Galves, Charlotte,1950-. III. Universidade Estadual de Campinas. Instituto de Estudos da Linguagem. IV. Título.

Informações para Biblioteca Digital

Título em outro idioma: A computational model of first language acquisition
Palavras-chave em inglês:
Language acquisition
Computational linguistics
Generativa grammar
Machine learning
Formal languages
Área de concentração: Linguística
Titulação: Doutor em Linguística
Banca examinadora:
Charlotte Marie Chamberlland Galves [Orientador]
Marcelo Barra Ferreira
Sérgio de Moura Menuzzi
Edson Françozo
Leticia Maria Sicuro Correa
Data de defesa: 19-11-2013
Programa de Pós-Graduação: Linguística

BANCA EXAMINADORA:

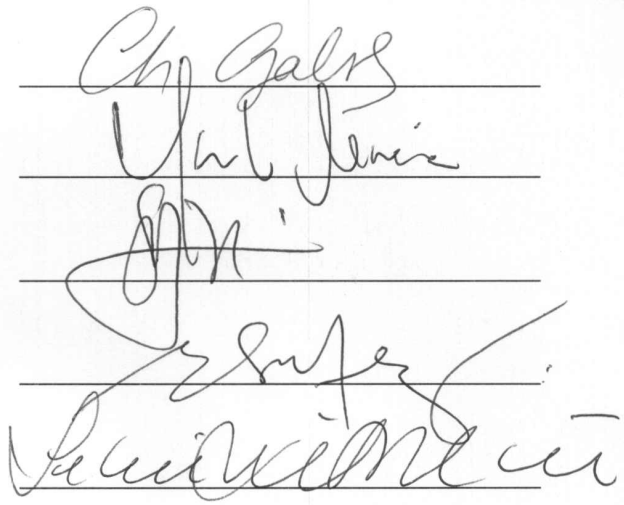
Charlotte Marie Chambelland Galves

Marcelo Barra Ferreira

Sérgio de Moura Menuzzi

Edson Françozo

Leticia Maria Sicuro Correa



Plínio Almeida Barbosa

Andrew Nevins

Cilene Aparecida Nunes Rodrigues

IEL/UNICAMP
2013

Resumo

Neste trabalho, o fenômeno de aquisição de uma língua natural é investigado através de uma modelagem computacional. O aprendiz modelado – apelidado de IASMIM – se caracteriza como um modelo computacional integrado de aquisição de primeira língua, visto que integra os processos de aquisição lexical e sintática. Além disso, o modelo foi concebido de modo a atender certos critérios de plausibilidade empírica e psicológica. A perspectiva teórica que norteia a investigação é a da Gramática Gerativa (cf. Chomsky, 1986) e este é um modelo voltado para a *competência linguística*, não um modelo de processamento ou de performance (i.e., de uso do conhecimento linguístico). O aprendiz modelado é capaz de adquirir um conhecimento gramatical relativamente abrangente e demonstra algum potencial translinguístico, particularmente no que diz respeito a variações de ordem. As simulações para avaliação do modelo permitem observar a emergência de padrões de adjunção e de recursividade na gramática, considerados aqui como as principais evidências de um conhecimento sintático mais elaborado. Finalmente, o modelo incorpora algumas noções caras à teoria sintática no âmbito do Programa Minimalista (cf. Chomsky, 1995b), tais como *set-Merge*, *pair-Merge*, "traço seletor" (cf. Chomsky, 1998), em conjunto com assunções sobre a binariedade das representações sintáticas e a hipótese de que a ordem linear não tem papel na sintaxe (cf. Uriagereka, 1999). O modelo incorpora, ainda, uma versão da representação semântico-conceitual proposta em Jackendoff (1990). Nesta modelagem, estas noções e assunções ganham uma interpretação concreta e integrada, interagindo na determinação das propriedades do conhecimento adquirido.

Palavras-chave: Aquisição da linguagem, Linguística computacional, Gramática Gerativa, Aprendizado do computador, Linguagens formais.

Abstract

In the present work, the acquisition of natural languages is investigated through a computer simulation. The modelled learner – dubbed IASMIM – is characterized as an integrated computational model of first language acquisition, in the sense that it integrates the processes of lexical and syntactic acquisition. Furthermore, the model was conceived in order to be empirically and psychologically plausible. The theoretical perspective of this enterprise is that of Generative Grammar (cf. Chomsky, 1986) and this is a model concerned with *linguistic competence*, rather than language processing or performance (i.e., how the acquired knowledge is put to use). The modelled learner is capable of acquiring a relatively broad grammatical knowledge and shows some crosslinguistic abilities, in particular, the ability to handle languages with distinct word orders. In the simulations for evaluation of the model we can observe the emergence of adjunction and recursive patterns in the grammar, taken here as the main pieces of evidence of a more elaborated syntactic knowledge. Finally, the model embodies some central notions for syntactic theory under the Minimalist Program (cf. Chomsky, 1995b), such as *set-Merge*, *pair-Merge* and "selector feature" (cf. Chomsky, 1998), together with the assumptions that syntactic representations are strictly binary branching and that linear word order has no significant role in syntactic phenomena (cf. Uriagereka, 1999). The model also embodies a version of the semantic-conceptual representation proposed in Jackendoff (1990). They take a concrete and integrated existence in this model, interacting with one another to determine the properties of the acquired grammatical knowledge.

Keywords: Language acquisition, Computational Linguistics, Generative Grammar, Machine learning, Formal languages.

Sumário

Dedicada à Maria Fernanda e Iasmim,
minhas flores, meus amores.

Agradecimentos

Enfim, concluo esta jornada longa, árdua, mas plena de crescimento e amadurecimento. Não posso deixar – como sempre – de iniciar meu agradecimento expressando o quanto sou grato a Deus pelas incontáveis dádivas em minha vida. Nunca me faltaram saúde ou condições materiais para que eu desse o melhor de mim, para que "multiplicasse os talentos". Nem tampouco faltaram pessoas maravilhosas, tanto na família, como entre amigos e colegas de trabalho, que nas horas difíceis me amparassem, seja com palavras carinhosas de incentivo, seja ouvindo com paciência meus desabafos, seja me apoiando materialmente. É a Tua divina presença, ò Pai, que vejo em cada um destes anjos. Viverei para retribuir com trabalho e amor tudo que tenho recebido de Ti.

Em segundo lugar, meus mais sinceros e profundos agradecimentos à minha amada esposa, Maria Fernanda, por estar ao meu lado durante tantos anos e especialmente durante este último ano do doutorado que tanto exigiu de ambos. Obrigado por acreditar em mim e na importância desta etapa para nossas vidas. Por cuidar tão bem de nossa querida Iasmim, esta linda flor e certamente o grande presente que a vida nos deu. Nunca se esqueça de como você foi fundamental para que eu concluísse esta pesquisa, Fê. Eu garanto que nunca esquecerei.

À minha família, amadas irmãs e especialmente minha querida mãezinha, Conceição, a quem sempre buscarei honrar e retribuir por todo o apoio, todas as orações, toda a luta para criar a mim e minhas irmãs e, principalmente, o exemplo de trabalho, de honestidade e abnegação. À minha segunda família, nas figuras do Zé Luiz, dona Eva e cunhad@s, também meu enorme agradecimento, pois também sempre estiveram presentes, apoiando de todas as formas. Sem todos vocês esta conquista talvez fosse impossível e certamente não teria o mesmo valor. Obrigado!

Aos colegas de Unicamp nesta longa caminhada, incluindo professores, não tenho como externar o quanto todos vocês foram e tem sido importantes. Seria impossível nomear a todos aqui, colegas do início da caminhada em Campinas e outros mais recentes. Mas há aqueles com os quais convivi mais de perto e pelos quais nutro um carinhoso sentimento de

amizade, tais como Aroldo de Andrade, André Antonelli, Aline Gravina, Gilcélia, Cynthia, Carlos Felipe, Elisângela, Lilian, Marcos, Juliana Kepitxo e Sabrina. A tod@s sou grato pelas inúmeras trocas de ideias, cafés, pela paciência em ouvir minhas ideias excêntricas e também com meu pessimismo generalizado nos momentos mais difíceis da pesquisa.

Há ainda alguns amigos muito queridos e fundamentais nessa caminhada, como meu querido Joaquim (Mr. Kim!), que me hospedou durante os últimos oito meses de pesquisa e a quem estarei sempre em dívida. Nice, essa querida amiga e (fada-)madrinha da Iaiá, companheira de inúmeros momentos agradáveis, seja nas reuniões em casa, nas idas ao cinema, nas meditações da Deeksha e nas pizzas na Di Capri. Amigos valiosos e que me propiciaram deliciosos interlúdios de descontração, fundamentais em momentos mais extressantes da pesquisa.

Agradeço também a todos os funcionários do IEL por sempre me atenderem com carinho, com presteza, tornando mais fácil a realização de minhas atividades acadêmicas. Em especial, deixo um forte abraço aos dois Carlos, o Carlão e o Carlinhos, ao pessoal da secretaria de pós, Miguel, Rose e Cláudio, à Malu, ao Emerson e às atenciosas funcionárias da secretaria de projetos, Sueli e Francis. Não posso deixar de mencionar e agradecer também a minha querida Tiana, a "Tianinha", como gosto de chamar, por me receber quase como uma segunda mãe, sempre carinhosa e receptiva, nas minhas fugas para um cafézinho lá na copa do instituto. Cheia de bons causos pra contar, dei muitas risadas com ela e fiquei muito feliz de privar de tantos momentos agradáveis. Obrigado, Tiana!

Dentre os professores, fico triste por não citar todos aqui, mas quero evitar injustiças. Sou um verdadeiro fã dos docentes do IEL, todos imensamente competentes e dos quais levo o exemplo de excelência acadêmica e, principalmente, o exemplo de humildade, pois sempre estiveram disponíveis e me trataram como um colega, apesar do abismo de conhecimento e experiência que nos separam. Em especial, entretanto, destaco Charlotte Galves, uma pessoa com quem acabei estabelecendo uma proximidade maior, por ter trabalhado alguns anos em seu projeto de pesquisa e que tão importante foi para mim neste tempo que passei em Campinas. Tenho-a na mais alta conta de carinho, respeito e admiração, e estou ansioso pelo pós-doutorado, em que voltaremos a trabalhar juntos novamente.

Agradeço muito também à minha orientadora, Ruth Lopes, por topar, assim como no mestrado, mais este desafio de orientar uma pesquisa que extrapola os limites da área de aquisição de linguagem. Somando ao mestrado, são seis anos de trabalho juntos para chegar neste momento. Por todas as dicas, as broncas, os debates, os cafés, a leitura crítica e, principalmente, pela confiança que depositou em mim, sou também imensamente grato a você, Ruth. Muito

obrigado por tudo e espero ter correspondido minimamente às expectativas.

Finalmente, meus agradecimentos sinceros à FAPESP – Fundação de Amparo à Pesquisa do Estado de São Paulo – pelo financiamento desta pesquisa de doutorado, através da bolsa 2009/17172-3, sem a qual este trabalho não teria sido possível.

Lista de Figuras

Lista de Tabelas

1

Visão geral

O presente trabalho trata do fenômeno da aquisição da linguagem na perspectiva da gramática gerativa (Chomsky, 1986). Isso implica ver o fenômeno como um processo em que o aprendiz parte do *estado inicial* da Faculdade da Linguagem (FL) passando por estágios intermediários até atingir o *estado final*, bastando para isso estar exposto aos dados linguísticos da língua-alvo, isto é, a língua de sua comunidade linguística. O processo investigado aqui compreende casos típicos de aquisição, isto é, por crianças típicas em situações típicas de exposição aos dados linguísticos, o que exclui o ensino explícito dentro ou fora do ambiente escolar. Interessa, portanto, o processo de aquisição expontânea da linguagem.

O estado inicial diz respeito aos princípios e elementos comuns aos falantes de todas as línguas naturais. A teoria linguística desse estado é chamada de Gramática Universal (GU). Os estados transitórios e o estado final, nessa perspectiva, dizem respeito à gramática interna do falante, isto é, um sistema simbólico de caráter gerativo capaz de produzir não apenas as expressões linguísticas às quais o aprendiz foi exposto, mas também expressões inéditas cujo conjunto assume-se ser infinito. Ao conhecimento incorporado neste sistema dá-se o nome de *competência* ou *língua-I*, esta em oposição ao que seria o conjunto (infinito) das expressões produzidas, designado *língua-E*.

Formalmente, o processo de aquisição é entendido como um processo de indução ou

inferência de gramática a partir de um conjunto finito de expressões bem-formadas da língua-alvo (Nowak et al., 2002). Por "gramática", faz-se referência aqui particularmente aos aspectos *lexical* (aquisição de palavras), *sintático* (aquisição de regras) e *semântico* (mapeamento entre representações semânticas e sintáticas) do conhecimento linguístico. O processo de aprendizagem é visto como majoritariamente indutivo porque o aprendiz tira suas conclusões sobre a gramática a partir da análise e da generalização sobre um conjunto de "fatos" (os dados linguísticos) que, por assunção, seriam derivados dela.

A aquisição da linguagem é investigada aqui através de uma modelagem computacional, seguindo uma tradição já razoavelmente longa de estudos desse tipo (Berwick, 1985, Gibson & Wexler, 1994, Gaylard, 1995, Villavicencio, 2002, Pearl & Sprouse, 2011, entre outros). O aprendiz modelado neste estudo foi apelidado de IASMIM e se caracteriza como um modelo computacional integrado de aquisição de primeira língua, visto que integra a aquisição lexical e a sintática. Além disso, o IASMIM foi concebido de modo a atender certos critérios de plausibilidade empírica e psicológica.

O período de aquisição compreendido pelo modelo abarca – de um modo geral – desde os estágios iniciais da aquisição lexical e sintática (estágio de uma palavra) até os estágios mais tardios (entre 4 e 5 anos de idade) em que a compreensão pela criança inclui construções subordinadas, passivas, interrogativas de longa distância, entre outras (Ingram, 1989, Guasti, 2002). Assim, o aprendiz – até por se tratar de um modelo integrado – não conta com um pré-conhecimento lexical, o que implica ter que adquirir palavras e seus significados. Ao mesmo tempo, o IASMIM assume uma visão restrita[1] da GU, visto que o aprendiz parte para a tarefa de aquisição sem o pré-conhecimento de categorias sintáticas ou lexicais.

O IASMIM não assume uma representação paramétrica explícita do conhecimento gramatical, como o fazem alguns dos modelos computacionais baseados na Teoria de Princípios e Parâmetros (Villavicencio, 2002, Yang, 2002, por exemplo). Alternativamente, o formalismo

[1] Quando comparado a outros modelos que, apesar de também assumirem alavancagem semântica, assumem um pré-conhecimento lexical e/ou sintático mais ou menos extenso.

adotado para representação gramatical e lexical é baseado em gramáticas moderadamente sensíveis ao contexto baseadas em *unificação* (Shieber, 1986). Portanto, o presente modelo pode ser visto também como uma proposta inicial de indução de gramática para formalismos dessa família. A base dessas gramáticas está na estrutura de dados chamada de *matriz de atributo-valor* ou *estruturas de atributos* (ou traços), utilizada para representar itens lexicais, categorias e estruturas sintáticas, informações semânticas, etc.

Esta modelagem assume alavancagem semântica[2]. Neste sentido, foi elaborada uma representação semântico-conceitual baseada nas propostas de Jackendoff (1983, 1990) e Pinker (1989), cuja interação com a componente sintática propicia a alvancagem. A premissa por trás da assunção de alavancagem semântica advém das conclusões de Gold (1967), em particular, a de que nenhuma linguagem[3] de cardinalidade infinita – como é o caso das línguas naturais – pode ser induzida apenas com base em dados positivos, compreendidos aí simplesmente como cadeias de símbolos da língua-alvo, portanto, sem informações sobre sua interpretação correta.

1.1 Objetivos da pesquisa

Dois objetivos principais foram perseguidos nesta pesquisa. O primeiro, foi o de implementar parte das propostas feitas em Faria (2009), resultantes da implementação e avaliação do modelo proposto em Berwick (1985). Faria concluiu que algumas mudanças poderiam conferir ao modelo maior abrangência gramatical e universalidade linguística. Em especial, Faria (2009) sugeriu (i) uma abordagem não-transformacional para o conhecimento sintático do aprendiz; (ii) o abandono do arcabouço X-barra, especialmente em relação à posição de especificador; (iii) a exclusão da ordem linear como componente do conhecimento sintático

[2] *Semantic bootstrapping*. Adaptado a partir do termo em inglês *bootstrap* que, literalmente, significa "alça de bota", dispositivo que existe em algumas botas para facilitar sua colocação. "To bootstrap" se tornou, portanto, uma expressão idiomática do inglês para designar situações em que certas tarefas quase impossíveis precisam ser superadas sem (ou como o mínimo de) recursos externos.

[3] Dentre as classes de linguagem definidas em Chomsky (1959).

adquirido; e (iv) o abandono do estoque de categorias sintáticas e papéis temáticos dados de saída ao aprendiz.

Por um lado, (i) e (iv) apontam para caminhos distintos da teoria gerativa vigente, visto que nas últimas décadas tanto o papel da operação Mova-α nas descrições linguísticas tem se acentuado, quanto o estoque de categorias sintáticas consideradas nas análises tem aumentado consideravelmente. De praxe, assume-se que tal maquinaria é dada pela GU. Os objetivos (i) e (iv), entretanto, apontam justamente na direção contrária, isto é, em descrever o conhecimento sintático sem recurso à operação de movimento e verificar se as categorias sintáticas podem ser induzidas a partir dos dados de entrada.

Os itens (ii) e (iii), por outro lado, se aproximam da teoria sintática, especialmente quando se considera direções apontadas pelo Programa Minimalista (Chomsky, 1993, e posteriores), como a que envolve dispensar a Teoria X-barra (Chomsky, 1995a) e a visão da sintaxe como algo mais abstrato, em que apenas a ordem hierárquica tem papel relevante (Uriagereka, 1999). Porém, como uma das balizas do Programa Minimalista é a de diminuir, quando possível, a maquinaria proposta pela teoria linguística, nada impede que (i) e (iv) compartilhem deste espírito e possam ser consideradas coerentes com o viés minimalista, embora distintas do que se tem feito de um modo geral na teoria sintática.

Isto nos leva ao segundo objetivo central desta pesquisa: implementar parte das noções teóricas mais recentes da teoria sintática, em especial aquelas que tiveram advento com o Programa Minimalista, como as citadas acima. Em outras palavras, o objetivo foi o de chegar a um modelo de aquisição que permitisse entrever modos de interação interessantes entre tais noções e, no caso do IASMIM, investigar de que modo a interação entre aquisição lexical e sintática poderia contribuir para atender às propostas (i-iv). Ao mesmo tempo, era importante que o modelo se mantivesse psicológica e, na medida do possível, empiricamente plausível.

1.2 Organização da tese

Esta tese está organizada em duas partes. A primeira compreende a revisão da literatura. Assim, o **Capítulo 2** introduz os pressupostos teóricos envolvidos, discute a natureza do problema da aquisição da linguagem, os resultados dos estudos formais de aprendibilidade[4] e as propriedades formais das línguas naturais. No **Capítulo 3**, são apresentadas e discutidas várias modelagens computacionais relacionadas à aquisição da linguagem, cuja tradição é relativamente longa, remontando a trabalhos como o de Langley (1982), entre outros. Este panorama geral é importante para situar o IASMIM em meio aos estudos deste tipo e para melhor avaliá-lo.

A segunda parte da tese compreende a apresentação e discussão dos resultados do IASMIM e a conclusão. O **Capítulo 4** é dedicado à apresentação do modelo, seus objetivos, arquitetura, estruturas de dados, processadores e procedimentos de aprendizagem, o que inclui exemplos reais de aprendizagem no modelo. O **Capítulo 5** apresenta e avalia os resultados obtidos pelo IASMIM na tarefa de aquisição lexical e sintática, quando submetido a diferentes conjuntos de dados de entrada, tanto para o inglês, quanto para o português. Finalmente, o **Capítulo 6** traz a conclusão e as considerações finais à respeito da presente pesquisa.

[4] *Learnability.*

Parte I

Línguas naturais e modelagens computacionais

2

Introdução

Uma modelagem computacional, seja qual for, demanda a especificação clara e exaustiva dos aspectos que se considera envolvidos no problema e na sua solução. Este seria o "nível computacional" de uma modelagem (cf. Pearl, 2010). Computar algo significa tomar uma *certa entrada* e processá-la segundo procedimentos pré-definidos para produzir *uma saída*. Assim, não é possível partir para a modelagem sem a compreensão da natureza e da estrutura das informações envolvidas na entrada e na saída. Só a partir daí é possível partir para o "nível algorítmico", o nível em que a solução do problema é descrita e em que também é preciso ter clareza sobre os mecanismos de processamento disponíveis e suas propriedades.

Portanto, antes de partir para a análise de modelos computacionais de aquisição (MCAs) e para a apresentação do IASMIM, é necessário estabelecer as bases conceituais e formais dos aspectos envolvidos no problema. Particularmente, é preciso especificar as propriedades formais do problema da aquisição da linguagem, o que implica caracterizar o que significa "saber uma língua" (a "saída" desejada), quais e qual a natureza dos dados de entrada e, finalmente, quais as propriedades e restrições eventualmente impostas sobre os procedimentos de processamento e aprendizagem disponíveis ao aprendiz. O objetivo deste capítulo, portanto, é estabelecer – com base numa revisão dos aspectos teóricos relevantes para este trabalho – uma especificação formal provisória do problema da aquisição, que servirá de base

para a avaliação de alguns MCAs propostos na literatura e do IASMIM.

2.1 A linguagem como um fenômeno mental

Chomsky (1986) ao retomar os fundamentos da gramática gerativa, apresenta a linguagem (e as línguas naturais) como um fenômeno de natureza mental[1]. Deslocando-se das noções de "língua" vinculadas ao senso comum – por exemplo, do tipo que permitem falar em "língua italiana" apesar da enorme variação dialetal (dimensão sócio-política) ou em "certo" e "errado" conforme alguma norma padrão (dimensão normativa ou prescritiva) –, Chomsky questiona a visão de língua como sendo algo independente de propriedades da mente e, portanto, cuja existência seria abstrata.

Nesta visão, uma língua teria o caráter de uma coleção infinita de ações ou comportamentos de um certo tipo, no caso, linguísticos. A gramática, por conseguinte, não seria mais que uma função que enumeraria (estocasticamente, talvez) os elementos dessa coleção. Chomsky (1986) questiona essa concepção de língua, condiderando-a arbitrária, artificial e sem valor teórico significativo, visto que não nos permite ir muito além da mera descrição superficial das expressões da língua. Além disso e fundamentalmente, mais do que armazenar e reproduzir sentenças, o falante se mostra capaz de interpretar e produzir expressões inéditas para ele e até para outros indivíduos de sua comunidade.

Estruturalistas e *behavioristas* (p.e., Skinner, 1957) tentaram explicar a criatividade do falante através de estratégias cognitivas gerais como, por exemplo, a *analogia*, através da qual o falante exploraria similaridades entre expressões para produzir ou interpretar novas. Várias evidências, entretanto, mostram que a analogia por si só não explica os fatos. Como exemplo, tomemos alguns dados apresentados por Chomsky (1986):

(1) a. John ate the apple

[1] Chomsky (1986) muitas vezes se refere à "mente/cérebro", deixando claro que ele não distingue as duas coisas, isto é, falar em "mente" para ele implica falar em "cérebro".

b. John ate

c. John is too stubborn to talk to Bill

d. John is too stubborn to talk to

Em (1b), entende-se que *John* comeu algo, portanto, o complemento omitido em comparação com (1a) é arbitrário. Por analogia, o falante deveria interpretar o complemento omitido em (1d) da mesma forma, isto é, como sendo arbitrário. Porém, não é o caso, visto que aí o elemento omitido se refere ao próprio *John*. Chomsky (1986) então questiona como a criança aprenderia tais distinções sem ensino explícito ou correção, se a analogia falha nestes (e noutros) casos. Este é um típico exemplo do que Chomsky chama "pobreza de evidências" (ou estímulos). Há, segundo ele, algo na mente do aprendiz que o predispõe e que determina certas propriedades da linguagem.

O falante é capaz ainda de *reconhecer* as expressões bem-formadas em sua língua em oposição a outras mal-formadas, mesmo se compreensíveis. Por exemplo, quando ouvimos um estrangeiro dizer algo como "O lua está linda este noite" ou uma criança pequena dizer "Eu fazi" (no lugar de "eu fiz"), interpretamos sem dificuldades o sentido, mas não temos dúvida quanto à sua má-formação gramatical. Por outro lado, reconhecemos a boa-formação de uma expressão como "Ideias verdes sem cor dormem furiosamente" (Chomsky, 1957), mesmo quando parece improvável fazer sentido da mesma. Isso evidencia uma propriedade peculiar da natureza do conhecimento linguístico, a saber, a existência de um nível de análise em que restrições sobre a forma do enunciado estão em operação, à despeito do acesso ao seu sentido global.

Alternativamente, portanto, Chomsky (1986) argumenta que a língua (ou linguagem) é um fenômeno eminentemente interno, mental. Para distinguir estes dois sentidos opostos, Chomsky se refere à "língua-I(nternalizada)" e "língua-E(xternalizada)". A segunda seria na melhor das hipóteses apenas um epifenômeno, enquanto a língua-I seria um "elemento da mente" do falante, um sistema finito, capaz de gerar (e interpretar) infinitas expressões e de

atribuir um estatuto de gramaticalidade para elas. A gramática seria, então, uma teoria da língua-I, enquanto a Gramática Universal seria uma teoria das línguas-I humanas, isto é, o sistema de condições (biologicamente dado) que determinam a classe das línguas naturais.

Assim, Chomsky (op.cit.) assume que o aprendiz parte para a aquisição com a FL em um estado inicial (S_0), comum a toda a espécie, ou seja, o estado descrito pela GU. Dada uma experiência apropriada, a FL passa a um estado mais ou menos estável, S_S, que daí em diante terá apenas mudanças periféricas (p.e., como acréscimo de vocabulário). Esse estado incorpora a língua-I e consiste de dois componentes: um próprio à língua-E em questão (morfologia, por exemplo) e outro relativo à contribuição do estado inicial (a GU). Para Chomsky, a maior parte do que é aprendido diz respeito ao primeiro, ou seja, à contribuição própria à língua-E. No decorrer do texto, usarei o termo *gramática* como referência à língua-I.

2.2 A aquisição da linguagem

2.2.1 Um breve histórico

A história dos estudos de aquisição é relativamente longa e heterogênea, em termos dos métodos adotados e pressupostos assumidos (Ingram, 1989). Iniciou com os estudos diaristas, por volta de 1876, que foram a principal forma de observação e investigação até meados de 1920. Tais estudos não eram sistematizados, sendo em geral resultado da iniciativa dos pais, e se caracterizavam pelo registro da aprendizagem da criança num certo período de tempo. Os diários exibem um significativo grau de aleatoriedade em termos dos eventos registrados no curso da aquisição das crianças, além de grande variação em termos da qualidade e do nível de detalhe dos registros. Ainda assim, constituem uma das fontes mais detalhadas sobre dados de aquisição (cf. Ingram, op.cit.).

Com o advento do *behaviorismo* após a Primeira Guerra Mundial, houve uma mudança importante nos estudos de aquisição. A premissa agora era a de que as mudanças no comportamento (no caso, linguístico) da criança seriam explicadas por condições observáveis –

e, portanto, mensuráveis – do meio-ambiente da criança. O papel da criança era então visto como passivo, sendo seu desenvolvimento controlado e determinado pelo meio-ambiente. Qualquer fator de natureza interna (cognitivo) era considerado não-mensurável e portanto rejeitado. Além disso, havia um maior controle sobre influências do meio na seleção de sujeitos (número similar de meninos e meninas, classe sócio-econômica similar etc.) (Ingram, op.cit.).

Metodologicamente, os estudos passaram a envolver um maior número de crianças, diferentemente dos diaristas, em que normalmente apenas uma criança era acompanhada. Um dos principais objetivos agora era estabelecer o "comportamento normal" da criança nas diferentes fases da aquisição. Ao invés da descrição longitudinal da aquisição de uma criança, os estudos eram transversais, isto é, focavam estágios específicos da aquisição, normalmente marcados pela idade das crianças. Assim, para observar o percurso da aquisição era preciso comparar os resultados dos estudos envolvendo crianças em diferentes estágios.

Uma característica desses estudos era a ênfase na medição, que se refletia na quantificação dos dados e sua tabulação. Em razão dessa ênfase na maior quantidade de crianças, tais estudos forneceram dados normativos importantes para o estabelecimento da tipicidade de crianças em processo de aquisição. Porém, conforme Ingram (op.cit.), havia ainda uma significativa falta de sofisticação linguística nas análises, já que basicamente os estudos atinham-se ao vocabulário, comprimento das sentenças e sons da fala. Assim, Ingram (op.cit.) argumenta que os dados coletados não permitem *insights* mais profundos sobre a aquisição de regras gramaticais e outros aspectos complexos da linguagem.

No nível metodológico, o registro era em geral feito manualmente, sem o uso de gravadores, acompanhando-se a fala da criança o mais rapidamente possível. Isso pode ter afetado a fidelidade dos dados, especialmente em relação à transcrição fonética. Segundo Ingram (op.cit.), os dados são pouco confiáveis para se tirar conclusões definitivas, embora possam dar respostas iniciais para várias questões, contribuindo para a definição do melhor caminho

a seguir numa investigação. Finalmente, os estudos foram majoritariamente descritivos, com algumas tentativas de explicar aspectos da fala (fonologia) e o uso inicial de palavras. Um dos poucos trabalhos em que a aquisição da sintaxe foi abordada foi o de Skinner (1957), como comentado na seção anterior.

O terceiro período, a partir de meados de 1950, chamado por Ingram (op.cit.) de período de "amostras de linguagem longitudinais", se caracterizou por tentar conjugar as virtudes dos estudos anteriores. As amostras agora eram representativas de um longo período de aquisição da criança, compostas por registros feitos em intervalos de tempo pré-determinados (entre cada visita). O registro passou a ser feito por investigadores e não mais pelos pais (apenas) e os estudos envolviam pelo menos três crianças, para que se pudesse ter indicações mínimas de tipicidade.

As sessões eram gravadas e o foco agora estava no desenvolvimento gramatical e não apenas em aspectos pontuais e superficiais da linguagem. Ao acompanhar longitudinalmente uma mesma criança, é possível observar as mudanças na sua gramática, ao mesmo tempo em que os dados das demais crianças permitem a comparação dessas observações. As metodologias de coleta se desenvolveram bastante nas últimas décadas e algumas bases de dados abrangentes foram formadas, tais como a base CHILDES[2] que disponibiliza dados de aquisição de várias crianças e para várias línguas diferentes.

A área experimental também se desenvolveu significativamente nas últimas décadas e provê informações bastante detalhadas sobre o desenvolvimento da percepção, os estágios de aquisição e a aquisição de inúmeros fenômenos gramaticais. Um panorama geral sobre a percepção da linguagem é apresentado em Jusczyk (1997). Sobre a aquisição de aspectos da sintaxe, Guasti (2002), entre vários outros, apresenta um panorama bastante amplo, que vai desde a aquisição lexical e emergência da sintaxe, até a aquisição de quantificação, princípios de ligação e relações de controle, além de apresentar evidências a favor da dissociação entre a

[2] A componente da linguagem da criança, no sistema *TalkBank*. URL (01/jun/2013): http://childes.psy.cmu.edu/.

linguagem e outras habilidades cognitivas advindas de estudos sobre crianças com diferentes tipos de déficit cognitivo.

2.2.2 As principais hipóteses explicativas

Crianças adquirem a linguagem sem esforço aparente. O fazem ainda num tempo limitado, sem ensino explícito por parte de adultos, sem correção ou exemplos agramaticais explicitamente marcados como tal, em circunstâncias variadas e de modo relativamente similar em diferentes línguas. Segundo Guasti (2002), há quatro hipóteses relativamente[3] alternativas para explicar a aquisição, envolvendo os conceitos de imitação, reforço, associação e Gramática Universal.

A hipótese de imitação sugere que crianças aprendem a linguagem ao tentar repetir ou imitar o que ouvem (dos adultos). A previsão dessa hipótese é, portanto, a de que a fala das crianças seria bastante influenciada pela fala dos adultos, algo que não se verifica nas observações. Guasti (op.cit.) aponta estudos que mostram que crianças produzem majoritariamente sentenças declarativas, mesmo sendo os enunciados dos adultos compostos principalmente por interrogativas e imperativas. Além disso, há o fato de que crianças (assim como adultos) conseguem desde muito cedo compreender e produzir novas sentenças, às quais não foram expostas (e, portanto, não poderiam imitá-las).

Outra contra-evidência está nos erros cometidos pelas crianças, por exemplo, na sobregeneralização flexional, ao produzir formas como *goed* (para "went") e *singed* (para "sang"), no inglês, ou *fali* (para "falei") e *canti* (para "cantei"), no português. Certamente a criança não ouviu tais formas na fala dos adultos e não pode, portanto, estar simplesmente imitando ou repetindo o que ouve. Outros exemplos citados por Guasti (op.cit., p.11) incluem interrogativas como *What does he doesn't eat?* e *What do you think what the puppet has eaten?*,

[3] Uma possibilidade é que, ao invés de serem alternativas, as hipóteses sejam complementares, cada uma se aplicando a certos aspectos ou estágios específicos do processo de aquisição como um todo. Acredito que a imitação, o reforço, a associação e a dotação inata têm um papel na aprendizagem da linguagem, embora certamente com relevâncias distintas.

claramente excluídas da gramática dos adultos falantes do inglês (embora sejam possíveis em outras línguas).

A segunda hipótese foi advocada por psicólogos *behavioristas* e sugere que a criança – assim como outros animais – aprende (seja qual for o domínio cognitivo) através de um mecanismo que associa estímulos a certas respostas (Skinner, 1957). Assim, ao produzir enunciados corretos a criança receberia reforço positivo por parte do adulto, recebendo reforço negativo em caso contrário. A aquisição seria, então, o resultado da fixação daquilo que foi reforçado positivamente. Como a hipótese anterior, esta falha especialmente em explicar a criatividade linguística da criança, cujos enunciados inéditos não poderiam ter recebido reforço positivo ou negativo. Além disso, dados de aquisição mostram que, em geral, os pais não corrigem a criança até porque, quando o fazem, a criança ignora as correções sistematicamente (Pinker, 1984, Guasti, 2002).

A hipótese baseada na associação aparece nas abordagens chamadas *conexionistas* (ver Seidenberg, 1997, Kaplan et al., 2008, Frank, 2011, Yang, 2011, para um panorama geral). Utilizando a estrutura do cérebro como metáfora, os modelos implementam redes neurais artificais que consistem de unidades de processamento interconectadas – semelhantes aos neurônios – cuja configuração (níveis de ativação das unidades e pesos vinculados às conexões – sinapses – de entrada e saída) muda à medida que o modelo associa padrões de entrada a padrões de saída. Para isso, as redes são *treinadas*, isto é, um algoritmo de aprendizagem faz ajustes na rede a partir de conjuntos de entrada e as saídas respectivas desejadas, especificados pelo projetista.

A partir daí, o modelo será testado para ver o quanto é capaz de generalizar a aprendizagem para novas entradas. De certo modo, portanto, tais modelos implementam a noção de estímulo-resposta e reforço comentadas anteriormente. Não são explicitamente simbólicos, visto que as informações que circulam na rede são apenas numéricas. As propriedades de interesse (p.e., atributos de tempo verbal) emergem como padrões de ativação, normalmente

distribuídos em duas ou mais unidades de processamento. Guasti (op.cit.) comenta algumas aplicações dessa abordagem para modelagem da aquisição, sendo a aquisição do tempo passado do inglês um dos fenômenos mais investigados. Embora consigam mimificar em parte o comportamento da criança em aquisição, uma inspeção mais criteriosa (Marcus, 1995) mostra limitações importantes que, até o momento, parecem não ter sido superadas (Seidenberg, 1997, Frank, 2011, Berwick et al., 2011).

Por exemplo, os modelos só deixam de sobre-regularizar após uma mudança abrupta nos dados de treinamento, mudança não observada nos dados a que a criança está exposta. Além disso, outros fatores importantes – para além de padrões sonoros – não são capturados por tais modelos, como por exemplo o papel do estatuto gramatical na sobre-regularização dos verbos plenos *have*, *do* e *be*, mas não das formas auxiliares correspondentes (Stromswold, 1990, *apud* Guasti, 2002). Ademais, não há ainda modelos conexionistas abrangentes, seja em relação à morfologia, seja em relação à sintaxe das línguas, que dêem evidências de que a aquisição de modo geral seria possível sem algum tipo de pré-disposição inata – o que incluiria uma manipulação simbólica mais explícita.

Outra importante limitação de tais modelos diz respeito ao modo como lidam com dados de entrada degenerados. Segundo Guasti (op.cit.), sabe-se que crianças expostas a *pidgins* – forma de comunicação rudimentar desenvolvida em ambientes multilíngues, como antigas plantações e colônias de escravos – desenvolvem línguas crioulas, consideradas como línguas naturais por incluir palavras funcionais e estrutura sintática mais elaborada. Modelos conexionitas, por serem completamente determinados pela entrada, não seriam capazes de modelar este fenômeno. Assim, embora seja provável que a associação e informação estocástica estejam envolvidos na aquisição, até o momento desconhece-se se há e qual seria o real potencial de tais modelos para adquirir aspectos mais complexos da linguagem.

Diante da falha das hipóteses comentadas acima em explicar a aquisição de propriedades mais abstratas da linguagem e da impossibilidade de induzir tais propriedades a partir

da experiência disponível – conforme o argumento de Chomsky (1968) sobre a pobreza de estímulos –, Chomsky propõe que parte do conhecimento linguístico seria inato, hipótese que ficou conhecida como Gramática Universal. Segundo essa hipótese, a classe de línguas a que a criança tem acesso é biologicamente determinada e comum a toda a espécie humana. Segundo Berwick et al. (2011), há debates sobre a extensão dessa dotação inata mas, em geral, gerativistas consideram que a GU é ricamente estruturada, incluindo restrições específicas à linguagem.

Desse modo, seria possível explicar a robustez do processo de aquisição, que se reflete na uniformidade do conhecimento exibido pelas crianças no decorrer e ao final do processo, a despeito da heterogeneidade de suas respectivas experiências linguísticas (p.e., em função de diferenças sócio-econômicas, grau de escolaridade dos pais etc.). Tal uniformidade não seria esperada se a aquisição fosse determinada majoritariamente pelos estímulos e condições do meio-ambiente. A GU determina, assim, o espaço de variação possível entre as línguas naturais, delimitando as hipóteses que a criança conjectura para fins de aquisição.

A aprendizagem, desta perspectiva, se restringiria basicamente ao aspecto lexical, morfológico e às relações de ordem nas línguas. Nesse sentido, a aquisição da linguagem pode ser vista também como um processo de maturação ou crescimento, cujo único pré-requisito é a exposição à experiência apropriada em conjunto com outros fatores (ver Chomsky, 2005, sobre o "terceiro fator"). Como exemplo dessa visão, Chomsky (1986, p.2) cita James Harris,

> "O crescimento do conhecimento... [lembra mais exatamente]... o crescimento da Fruta [sic]; embora causas externas possam cooperar em algum grau, é o vigor interno e a virtude da árvore que deve levar os frutos à sua maturidade" (tradução livre)

Ou, como o próprio Chomsky (1965, p.51) coloca,

> "Portanto, a *forma de uma língua*, o esquema para sua gramática, é em grande

> parte dada, embora ela não vá estar disponível para uso sem experiência apropriada para colocar os processos de formação de linguagem em operação." (tradução livre)

Se essa hipótese estiver no caminho certo, a questão passa a ser determinar a extensão da GU, isto é, qual a dotação inata necessária e suficiente para determinar as propriedades da aquisição da linguagem nos seres humanos. Como se trata de propriedades universais, seja qual for a especificação exata da GU, esta certamente não deverá incluir aspectos particulares de alguma língua ou família de línguas, mas sim as propriedades das quais tais aspectos seriam derivados. Para isso, é preciso aprofundar nas particularidades da "pobreza de estímulos", de modo a identificar com exatidão quais propriedades da linguagem de fato não podem ser induzidas[4] a partir da experiência.

2.2.3 A pobreza de estímulos e o problema lógico da aquisição

Chomsky (1968) introduz o conceito de "pobreza de estímulos" (ou "evidência", cf. Chomsky, 1986) apresentando um exemplo envolvendo a inversão sujeito-auxiliar do inglês, necessária para a formação de interrogativas do tipo "sim/não", como mostram os dados em (2) e (3), reproduzidos a seguir:

(2) a. Will the members of the audience who enjoyed the play stand?

b. Has Mary lived in Princeton?

c. Will the subjects who will act as controls be paid?

(3) a. The members of the audience who enjoyed the play will stand.

[4] A aquisição pode ser vista também como um processo dedutivo, em que a criança parte do pré-conhecimento completo e suficiente de princípios e parâmetros para descrever qualquer língua natural, cabendo ao mecanismo de aquisição deduzir qual a configuração ideal para a língua-alvo. Essa dedução implica em testar diferentes configurações, premiando-as ou punindo-as de acordo com seu sucesso em descrever os dados de entrada. Esta não é a abordagem no IASMIM, pois seria como assumir que ela já tem um estoque completo de regras à disposição. O objetivo da modelagem, entretanto, é mostrar que as regras podem ser induzidas e que, portanto, a extensão do aparato inato é menor, restrito a um conjunto de princípios básicos.

b. Mary has lived in Princeton.

c. The subjects who will act as controls will be paid.

Chomsky queria mostrar o que chamou de "dependência estrutural" das regras sintáticas. Fosse a criança livre para conjecturar qualquer tipo de regra, diante de dados como (2a/3a) e (2b/3b), uma regra aparentemente simples e suficiente nestes casos poderia ser descrita como "Mova o primeiro auxiliar da sentença para a primeira posição", baseada apenas na ordem linear (temporal) das palavras. Em ambos os casos, o resultado produzido seria gramatical, mas não para (3c), em que há duas ocorrências do modal "will" e para o qual essa regra moveria o primeiro auxiliar, produzindo a sentença agramatical "*Will the subjects who act as controls will be paid?*".

Portanto, a criança teria que criar outras hipóteses até chegar à hipótese correta que deve fazer referência ao auxiliar da oração principal, o que implica considerar a estrutura sintática da sentença, daí o conceito de "dependência estrutural". Se assumirmos que não há qualquer tipo de restrição sobre a aquisição e que, portanto, qualquer regra pode ser conjecturada, a previsão é de que a criança cometeria erros envolvendo estas construções no decorrer da aquisição, antes de a hipótese correta estar fixada. No entanto, há evidências (Crain & Nakayama, 1987) de que isso não ocorre.

Chomsky extrapola isso para outros fenômenos complexos da gramática (ver, por exemplo, Berwick et al., 2011), argumentando que os dados de entrada por si só, na medida em que são compatíveis com hipóteses cuja natureza é inconsistente com a gramática final, não explicam como a criança adquire tal conhecimento aparentemente sem cometer (certos) erros. Em outras palavras, os dados parecem ser *enganosos* – ao serem compatíveis com hipóteses de *tipo incorreto* – para explicar a aquisição da gramática. Chomsky vai além, segundo Pullum (1996), e afirma que os dados são ainda *incompletos*, isto é, os dados necessários para distinguir *exaustivamente* certas hipóteses corretas das erradas podem, eventualmente, nunca ser apresentados ao aprendiz. Ainda assim, o aprendiz generaliza corretamente para

além da experiência a que foi exposto.

Essa linha de raciocínio ficou conhecida como o "argumento da pobreza de estímulos" (APE) e o estado de coisas apresentado pelo argumento configura, assim, o que ficou conhecido como "problema lógico da aquisição da linguagem", que coloco nos seguintes termos[5]: como explicar a natureza e a riqueza do conhecimento gramatical do falante, se a experiência (assumindo indução) é finita, baseada em dados positivos e, *por si só*, incompleta e até enganosa? A solução proposta por Chomsky está nos princípios e elementos previstos pela Gramática Universal, isto é, no aparato inato que predispõe o aprendiz em relação à classe de gramáticas acessíveis. Tal aparato explicaria a uniformidade e complexidade do conhecimento, bem como a ausência de certos tipos de erros no decorrer da aquisição.

Muito se tem discutido, desde Chomsky (1968), sobre o APE e, consequentemente, sobre a extensão e a natureza (se específico à linguagem ou se de domínio geral) do pré-conhecimento necessário à criança para ter sucesso na aquisição (ver Pullum, 1996, Lidz et al., 2003, Pearl & Lidz, 2009, Berwick et al., 2011, entre outros). Um dos principais meios de questionar a força do argumento – se não absolutamente, pelo menos em parte – tem sido mostrar que os dados não são tão insuficientes quanto afirmado e que é possível induzir regras sintagmáticas a partir deles (Pullum, 1996, Berwick et al., 2011). Tal estratégia põe em evidência *duas dimensões* envolvidas no APE: a dimensão relativa aos dados de entrada (sua pobreza) e a classe das hipóteses conjecturadas (sua restrição).

Para questionar o APE, portanto, é preciso lidar com as duas dimensões. Segundo Berwick et al. (2011), não basta mostrar que regras sintagmáticas podem ser aprendidas, isto é, induzidas a partir dos dados de entrada e que há mais dados disponíveis do que o APE sugere. É preciso ainda explicar por que certas hipóteses parecem nunca ser conjecturadas,

[5] Ver Pinker (2004, p.949) para uma discussão sobre a interpretação correta para esta expressão. Em suas palavras, "*It is the question of how acquisition could work in principle – how a learner can correctly generalize from a finite sample of sentences in context to the infinite set of sentences that define the language from which the sample was drawn.*" No entanto, não consigo ver este problema desvinculado do APE, até por que a infinitude do conjunto de sentenças não é, por si só, o grande desafio para o aprendiz, mas sim as restrições que permitem *certas infinitudes* e não outras. Daí a formulação alternativa que propus aqui.

como indica por exemplo a ausência nas observações de aquisição do tipo de erro que seria esperado se regras de inversão sujeito-auxiliar envolvendo apenas posições na cadeia linear fossem conjecturadas. A força do APE está, portanto, nessa dupla dimensão, especialmente na segunda.[6]

Embora possível, parece pouco provável que a explicação mais apropriada para a restrição das hipóteses (atestada pela ausência de certos tipos de erros) possa ser explicada sem recurso a alguma pré-disposição inata, por mais elementar que seja. Nada impede, entretanto, que tal pré-disposição, chegando a um mínimo, seja de natureza cognitiva geral, o que no fim é uma questão empírica, ou seja, seria necessário dar evidências de que tal pré-disposição também é relevante para outros domínios cognitivos.

2.2.4 Uma caracterização formal do aprendiz

Chomsky (1965) sugere um *dispositivo de aquisição da linguagem* (LAD) que deveria atender às condições em (4), equipando o aprendiz da linguagem com:

(4) (i) uma técnica para representar sinais de entrada;

(ii) uma maneira de representar informação estrutural sobre estes sinais;

(iii) alguma delimitação inicial de uma classe de hipóteses possíveis sobre a estrutura da linguagem;

(iv) um método para determinar o que cada hipótese implica com respeito a cada sentença;

(v) um método para selecionar uma das (presumivelmente, infinitas) hipóteses que são permitidas por (iii) e são compatíveis com os dados linguísticos primários fornecidos.

[6] É sempre bom frisar o sabor "dedutivo" que o APE tem, especialmente em sua formulação original. Esse caráter se opõe em parte à abordagem assumida aqui, que considero mais indutiva, embora não completamente, visto que o aprendiz está pré-disposto em certa medida, o que lhe permite fazer deduções.

Por "dados linguísticos primários" (DLP), Chomsky (1965, p.32) compreende dados que "consistem de sinais classificados como sentenças e não-sentenças e um pareamento parcial e provisório de sinais com descrições estruturais". É interessante notar algumas propriedades dos DLP como definidos por Chomsky. Primeiramente, a inclusão de evidência negativa, na forma de "não-sentenças", isto é, sentenças agramaticais (em relação à língua-alvo) explicitamente marcadas como tal (p.e., correção explícita da fala da criança). Estudos posteriores, entretanto, mostraram que este tipo de dado parece ter pouco ou nenhum[7] papel relevante na aquisição da gramática (Marcus, 1993, Guasti, 2002, Pinker, 2004). Outro aspecto importante dos DLP está relacionado ao que Chomsky chama de "descrições estruturais" que, segundo ele, são

> "um pré-requisito visto que a aquisição da linguagem parece advir da visão amplamente aceita (mas, no momento, pouco suportada) de que deve haver uma base parcialmente semântica para a aquisição da sintaxe ou para a justificação da hipótese sobre a componente sintática da gramática" (Chomsky, 1965, p.32) (tradução livre)

A consequência dessa assunção, para Chomsky, é a de que a criança possui uma teoria inata sobre descrições estruturais potenciais desenvolvida ao ponto de permitir à criança determinar, numa situação real de ocorrência de um "sinal" linguístico, quais descrições são apropriadas. Além disso, ela o faz em parte antes de qualquer assunção sobre a "estrutura linguística" do sinal. Não fica claro o sentido de "estrutura linguística" aí, pois ele não a define. Segundo Chomsky (1959), uma descrição estrutural (DE) consiste da estrutura sintagmática de uma sentença como derivada pela gramática. Vale lembrar que uma mesma sentença pode eventualmente ter mais de uma DE válida. Conclui-se que a criança teria de saída hipóteses sobre a estrutura das sentenças, mesmo antes de processá-las. A condição

[7] Saxton (2000) e Saxton et al. (2005), por outro lado, analisam corpora longitudinais de aquisição e afirmam que seus resultados indicam o uso de evidência negativa pela criança para recorrer de hipóteses incorretas.

para este suporte seria a "base semântica" mencionada por Chomsky.

Uma série de trabalhos posteriores à Chomsky (1965) propuseram definições mais explícitas do pré-conhecimento semântico que estaria na base da aquisição sintática. Chamada de *alavancagem semântica* (ver Anderson, 1978, Pinker, 1984, Gropen et al., 1991, Braine, 1992, Gaylard, 1995, entre outros), essa hipótese prevê que inicialmente os enunciados seriam pareados com uma representação semântica, nalguns modelos na forma de mapeamentos entre categorias semânticas *básicas* (p.e., "paciente-da-ação") e categorias sintáticas (p.e., "objeto"), noutros na forma de grafos acíclicos dirigidos (DAG) codificando os conceitos e relações semânticas envolvidos.

Sobre a plausibilidade dessa assunção, Pinker (1984, p.30) propõe ser suficiente assumir que "[...] *a criança processa, para fins de indução de regras, apenas aquelas sentenças cujo significado esteja disponível contextualmente, ignorando o resto*". Ou seja, apenas quando ela consegue restringir sua atenção a um contexto compatível com o enunciado – isto é, em que padrões recorrentes e co-ocorrentes sejam identificáveis –, o dado é útil para fins de aquisição; do contrário, o dado é descartado. Tal habilidade não necessita, assim, ser absoluta e nem mesmo estar completamente madura desde os primeiros dias de vida.

Em outras palavras, pode ser parte do desenvolvimento da criança, especialmente no período que antecede a emergência das primeiras palavras (i.e., os primeiros 10 a 12 meses de vida). Guasti (2002) cita, por exemplo, uma prediposição inata para estabelecer atenção conjunta com o adulto, no momento da interação, de modo a identificar o foco de atenção. A autora comenta ainda três viéses que guiariam as hipóteses da criança sobre novas palavras: vinculá-las a objetos inteiros (na cena); se for estendê-las, fazê-lo para objetos da mesma categoria ("viés taxonômico"); e assumir que palavras têm um único referente ("viés de exclusividade mútua"), o que auxiliaria a aquisição de palavras para substâncias e partes de objetos. Estes são exemplos de fatores que auxiliariam a criança a delimitar o contexto e ter sucesso no mapeamento entre palavras e significados.

Um aspecto importante do LAD diz respeito ao processamento dos dados de entrada. Chomsky (1986, p.25) propõe a especificação de um analisador sintático (*parser*), que incorporaria as regras da gramática e alguns outros elementos:

> "[...] uma certa organização de memória e acesso (talvez uma estrutura com pilha determinística e uma área temporária de um certo tamanho; ver Marcus, 1980), certas heurísticas e assim por diante. Um analisador[8] não deveria mapear expressões em suas estruturas da maneira que estas são associadas pela gramática. Por exemplo, um parser deveria falhar ao fazer isso no caso das chamadas 'sentenças labirinto'[9] ou sentenças que sobrecarregam a memória numa passagem da esquerda para a direita [...]" (tradução livre)

Basicamente, Chomsky está se referindo a propriedades específicas do processamento humano de sentenças, como investigado em diversos estudos de psicolinguística (Clifton et al., 1991, MacDonald et al., 1994, Kaiser & Trueswell, 1994, entre outros). É importante ressaltar a separação feita por Chomsky entre a gramática e o analisador: a primeira *provê* descrições estruturais para qualquer sentença da língua sem que se coloquem aí questões de custo, complexidade ou dificuldade, pois não haveria processamento de fato. É apenas no âmbito do processamento que tais questões se colocam e dependem tanto das características do analisador, como do problema modelado.

Corrêa (2008) discute o que ela chama de "problema de unificação", isto é, as dificuldades que se colocam para uma completa "identificação entre uma derivação minimalista e uma caracterização funcional da computação *on-line*". Em outras palavras, a separação entre *gramática* e *analisador* não se reduz a uma mera opção teórica; como aponta Corrêa, há pelo menos dois aspectos em que os dois conceitos conflitam, quando se trata da caracterização estrutural (análise) de um enunciado: (i) a direcionalidade da derivação, em

[8] *Parser.*
[9] *Garden-path sentences.*

contra-posição ao caráter incremental do processamento, e (ii) o custo computacional que estaria associado aos diferentes tipos de operações de movimento envolvidas numa derivação. Assim, para se chegar a um modelo verdadeiramente integrado, seria necessário solucionar tais conflitos e Corrêa (2008) aponta alguns caminhos nessa direção.

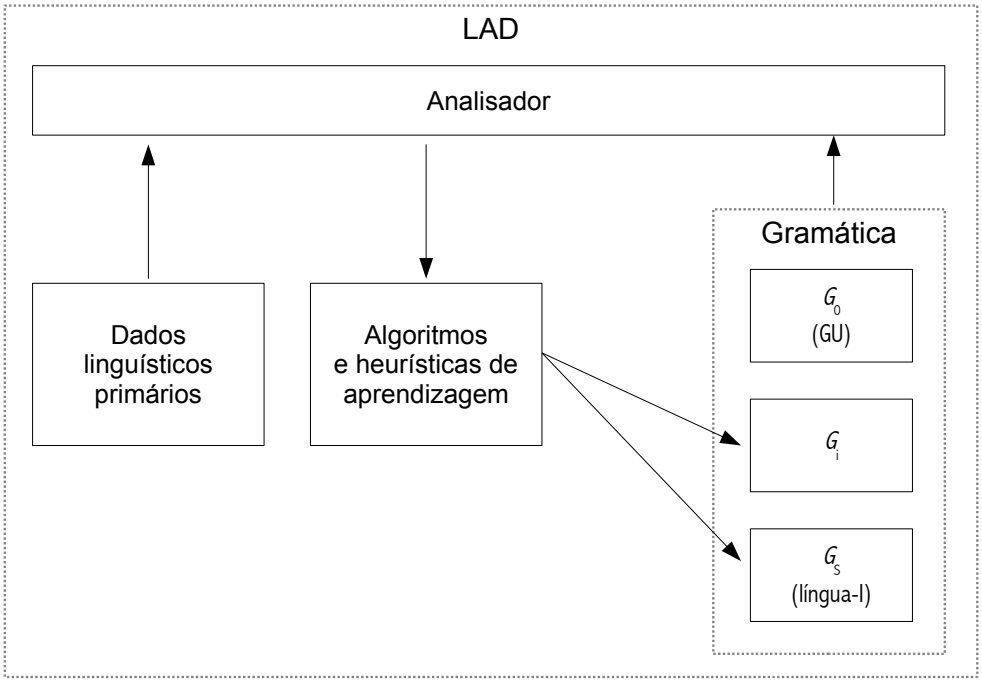


Figura 2.1: Esquema representativo do LAD

Diante de (4) e do que foi discutido acima, é possível propor a especificação esquemática preliminar para o LAD, na Figura 2.1, que seria um dispositivo abstrato em que interagem diferentes componentes da FL: os DLP, o analisador, os procedimentos (algoritmos e heurísticas) de aprendizagem e a gramática (em seus diferentes estágios, sendo o primeiro dado pela Gramática Universal e o último, a gramática adquirida no processo de aquisição). Vale ressaltar ainda a separação estrita entre a GU e os demais componentes envolvidos no LAD, um aspecto decorrente da assunção de uma pré-disposição inata específica à linguagem (pelo menos no que diz respeito à sintaxe) que não necessariamente teria impacto sobre os demais componentes do LAD (embora possa ter).

2.2.5 O percurso geral da criança

Os primeiros passos

Guasti (2002) e Jusczyk (1997) discutem inúmeros experimentos de percepção linguística envolvendo bebês, cujos resultados indicam habilidades bastante precoces para reconhecer e distinguir padrões na fala. A conclusão geral é a de que os seres humanos exibem inicialmente uma habilidade para reconhecer e distinguir padrões rítmicos de línguas (inclusive entre pares de línguas que não incluam sua língua nativa) e que, com o desenvolvimento, essa habilidade vai se especializando nos padrões e distinções relevantes para sua língua nativa (ou línguas, se mais de uma).

Experimentos mostram, por exemplo, que bebês com 4 dias de vida são capazes de discriminar os enunciados em língua nativa de enunciados em línguas estrangeiras.[10] Tais resultados indicam um viés precoce dos seres humanos para certas propriedades sonoras das línguas naturais. Neste caso, os experimentos indicam que informação prosódica, particularmente a característica rítmica das línguas, seria suficiente e necessária para sua discriminação. Por volta dos dois meses de vida, entretanto, experimentos indicam – embora não conclusivamente – que essa habilidade começaria a diminuir, se restringindo à discriminação entre a língua nativa, de um lado, e estrangeiras, de outro.

O próximo passo da criança é aprender as distinções fonêmicas, o que experimentos indicam começar por volta do segundo mês de vida, quando bebês já são capazes de discriminar contrastes consonantais da língua nativa, mas não apenas. Experimentos com bebês de 6 a 8 meses, indicam a percepção de contrastes irrelevantes para sua língua nativa. Esse caráter "universal" da discriminação diminui drasticamente após os 8 meses de vida, praticamente desaparecendo ao final do primeiro ano. A criança parece então focar nas distinções relevantes para sua língua nativa, ou seja, aquelas que vão ajudá-la a distinguir significados

[10] Com base em informações supra-segmentais, visto que nesse estágio elas desconsideram informação segmental.

e, assim, começar a construir um léxico.

A emergência das palavras e os estágios da aquisição

Segundo Guasti (op.cit.), antes da emergência das primeiras palavras, o bebê passa por um período de *balbucio*, em geral iniciando entre 6 e 8 meses de vida, caracterizado por exibir sons com organização silábica, características universais (i.e., não necessariamente sons frequentes na língua nativa) e não associados sistematicamente a significados. Antes dos 6 meses, os sons produzidos pela criança não são ainda considerados balbucio, *nesse sentido*. Só por volta dos 10 meses é que o balbucio passa a apresentar apenas segmentos fonéticos mais comuns à língua nativa e as primeiras palavras (não necessariamente correspondentes às formas adultas) podem começar a emergir.

O balbucio é visto como um passo importante do bebê, pois é quando ele começa a testar suas capacidades articulatórias, descobrindo e praticando os sons e suas combinações na sua língua. Guasti (op.cit.) cita ainda estudos que mostram que crianças surdas expostas a línguas de sinais também passam por um fase similar de "balbucio manual", o que indica que a linguagem não está presa a uma única modalidade de expressão (no caso, a oral). Assim, segundo Guasti, a emergência do balbucio parece depender, antes, da maturação do substrato neural que dá suporte à linguagem, do que da maturação do aparato vocal ou gestual, embora tais desenvolvimentos sejam sem dúvida necessários.

Por volta dos 9 meses, o bebê está bastante sensível às propriedades fonéticas e fonotáticas da língua, o que o coloca na posição ideal para adquirir palavras. Nesse período, a criança parece capaz de armazenar de alguma forma os padrões sonoros das palavras, embora sem associá-los ainda a significados, o que iniciaria entre os 10 e 12 meses. A partir desse período, que poderíamos considerar "pré-gramatical", no sentido em que não envolvem particularidades lexicais, sintáticas ou semânticas, a criança inicia então a aquisição gramatical, que se dá também de modo gradual e relativamente uniforme (entre as crianças de uma

mesma comunidade).

Uma das propostas mais conhecidas, a de Brown (1973), estabelece um conjunto de (cinco) estágios de aquisição do inglês baseados não na idade, mas na co-ocorrência de certos "comportamentos linguísticos" e o comprimento médio do enunciado (MLU[11]), que levava em conta o número de morfemas e não o de palavras. Ingram (op.cit.), a partir da análise dessa proposta e de outras, propõe uma correlação parcial entre os estágios e faixas etárias. Na Tabela 2.1, apresento uma versão unificada e resumida das tabelas apresentadas em Ingram (op.cit., p.50-53), para dar uma visão geral dos estágios propostos.

Estágio	MLU	Descrição
		Período de desenvolvimento pré-linguístico (0–1;0)
		Período de enunciados com uma palavra (1;0–1;6)
		Período das primeiras combinações de palavras (1;6–2;0)
I	1.00–1.99	*Papéis semânticos e relações sintáticas.* Aquisição de relações semânticas básicas (Agente, Paciente) e da ordem das palavras.
		Período de sentenças simples
II	2.00–2.49	*Modulação do significado.* Início da aquisição de flexões e morfemas gramaticais.
III	2.50–2.99	*Modalidade de orações simples.* Aquisição ativa do auxiliar tal como a aparece em questões sim/não, questões-*qu*, imperativas e questões negativas.
		Período de sentenças complexas
IV	3.00–3.99	*Encaixamento de uma sentença na outra.* Orações complexas aparecem com objetos na forma de sintagmas nominais, questões-*qu* e orações relativas.
V	4.00–	*Coordenação de orações simples e relações proposicionais.* O desenvolvimento ativo da coordenação de sentenças, sintagmas nominais e verbais, com o uso de conjunções.

Tabela 2.1: Resumo dos estágios de aquisição do inglês, adaptadas a partir de Ingram (1989).

Ingram não aponta as faixas etárias que estariam vinculadas aos três últimos estágios.

[11] *Mean Length of Utterance.*

Consultando Brown (1973), vemos que as faixas etárias para os estágios variam significativamente entre as três crianças acompanhadas (Adam, Sarah e Eve). Para o Estágio III, enquanto Adam e Sarah aparecem mais próximos (idades de 2;11 e 3;1, respectivamente), Eve foi bem mais precoce (1;11). Para o Estágio 4, o quadro continua similar, que inicia aos 3;2 anos para Adam, 3;8 anos para Sarah e 2;2 anos para Eve. Por fim, o Estágio V inicia aos 3;6 anos para Adam, 4;0 para Sarah e 2;3 para Eve. Note que a diferença de idade entre os estágios III e V é relativamente pequena (5 a 7 meses), razão pela qual Ingram os inclui num mesmo período, chamado por ele de "período de sentenças complexas".

Um quadro geral mais detalhado em termos dos aspectos gramaticais envolvidos pode ser resumido a partir de Guasti (2002). Guasti assume uma visão paramétrica do conhecimento gramatical, segundo a Teoria de Princípios e Parâmetros (cf. Chomsky, 1986), razão pela qual adaptei os dados a seguir, abstraindo dos aspectos particulares consequentes dessa visão (p.e., o parâmetro de movimento de verbo, ordem "V2", entre outros). Vale ressaltar que os dados da Tabela 2.2 se referem primordialmente ao inglês, embora línguas como o francês, italiano, alemão, entre outras, tenham sido consideradas por Guasti e apresentem percursos com características em comum.

Sobre o percurso acima, algumas coisas precisam ser ditas. Primeiramente, Guasti (op.cit.) fez um apanhado geral a partir de vários estudos independentes, o que não permite compreender apenas pela Tabela 2.2 relações de causalidade entre os fenômenos inclusos numa dada faixa-etária. Assim, essa classificação tem também um caráter mais descritivo e exibe alguma coincidência com a classificação anterior, visto que as faixas-etárias na Tabela 2.2 podem ser encaixadas nos dois últimos períodos da Tabela 2.1: os fenômenos listados na faixa de 2–3 anos parecem aprendíveis com base em sentenças simples, enquanto as demais faixas etárias demandam sentenças com encaixamento de orações (período de sentenças complexas, portanto).

Um segundo aspecto importante é que a classificação em faixas etárias não implica a

Idade	Conhecimento (inicial/parcial)
2–3	Ordem relativa entre as palavras Distinção das categorias lexicais e funcionais Propriedades morfológicas dos verbos Distinção de verbos e auxiliares Propriedades de verbos inacusativos Relações de concordância entre elementos da oração Omissão de sujeitos sentenciais Distinção entre queda de tópico e de sujeito Formação de questões
3–4	Compreensão e produção de questões-*qu* de longa distância, como "*What do you think that Ninja Turtles like to eat?* (SR, 3;11) (p.210) Regras recursivas Formação de relativas, através de movimento-*qu* ou de pronomes resumptivos Compreensão e produção de passivas envolvendo verbos de ação e incluindo o agente via preposição *by* Princípios A, B e C da teoria de ligação Distinção entre sintagmas nominais referenciais e quantificados Restrição do quantificador pelo substantivo com o qual se combina Produz estruturas de controle (mas apresentam dificuldades e são mais permissivas em relação à interpretação)
4–5	Mecanismo de alçamento de quantificador

Tabela 2.2: Resumo do percurso da aquisição (baseada em Guasti, 2002).

aquisição plena dos fenômenos respectivos e que a criança não cometa ainda certos erros. Além disso, a ordem de aquisição pode variar entre as línguas, como indicam estudos sobre aquisição de passivas nas línguas inuktitut e quiché.[12] Em geral, a ocorrência de um fenômeno numa dada faixa-etária indica apenas que a criança demonstra algum conhecimento do mesmo, em maior ou menor grau, visto que há fenômenos mais complexos do que outros. As tabelas acima servem, portanto, como um quadro empírico geral contra o qual analisar o percurso de aquisição exibido por modelos de aquisição, tanto em termos do

[12] Guasti (op.cit.) cita a dificuldade das crianças nessa idade para lidar com passivas envolvendo outros verbos (que não de ação), o que indica que a aquisição completa é mais tardia. O PB também parece apresentar uma aquisição mais tardia (cf. Rubin, 2004). Por outro lado, observações para outras línguas, mostram uma aquisição mais precoce. Allen & Crago (1996) mostram que crianças adquirindo a língua inuktitut tem uma aquisição mais precoce de passivas (2;0 a 3;6), um resultado similar ao de Pye & Poz (1988) para a língua quiché. Tais resultados sugerem haver outros fatores envolvidos na aquisição de passivas.

comprimento dos enunciados, quanto em relação à sequência exibida e à dificuldade imposta por alguns aspectos da gramática.

2.3 O caráter formal das línguas naturais

2.3.1 Linguagens formais e autômatos

Nesta seção apresento uma breve introdução sobre a teoria de linguagens formais e autômatos, de modo a preparar o terreno sobre o qual discutiremos certos aspectos das gramáticas formais propostas para as línguas naturais e dos modelos computacionais de aquisição que as assumem. Para uma introdução mais abrangente sobre conceitos e métodos matemáticos e computacionais aplicados aos estudos linguísticos ver, entre outros, Partee et al. (1993), Wintner (2002), Levelt (2008) e Hammond (2010). Sobre linguagens formais, autômatos e teoria da computação, ver Hopcroft et al. (2001).

Conceitos básicos

Uma distinção fundamental para os estudos sobre linguagens formais diz respeito aos conceitos de *linguagem* e *gramática*. Uma linguagem L é definida como um conjunto (eventualmente infinito) de *cadeias* ("strings") construídas sobre algum alfabeto (ou *vocabulário*) finito. Tomando Σ^* como o conjunto de todas as possíveis cadeias sobre o alfabeto Σ, temos que $L \subseteq \Sigma^*$. Uma gramática G seria, em contra-partida, uma maneira de caracterizar a linguagem L de modo a listar quais cadeias pertencem à L, quais não. Pode ser vista, assim, como um procedimento que *aceita* certas cadeias, aquelas pertencentes à L, rejeitando as demais.

Formalmente, uma gramática é definida como $\{V_T, V_N, S, R\}$, em que V_T é o conjunto de elementos *terminais* (elementos do vocabulário), V_N é o conjunto de não-terminais, S é um membro especial de V_N a partir do qual as derivações são iniciadas e R é um conjunto finito de *regras de produção* a partir das quais é possível *derivar* cadeias. Estas regras

consistem em enunciados de equivalência lógica da forma $\alpha \rightarrow \beta$, em que α e β representam cadeias de elementos terminais e não-terminais, representados respectivamente por letras minúsculas e maiúsculas.

Mais particularmente em relação à R, será assumido que o lado esquerdo de uma regra contém pelo menos um elemento não-terminal (i.e., pertecente à V_N). Tomando Σ como $V_T \cup V_N$, isto é, como sendo o conjunto formado por todos os terminais e não-terminais, temos que R é um conjunto finito de pares ordenados gerados a partir de $\Sigma^* V_N \Sigma^* \times \Sigma^*$. Portanto, $\alpha \rightarrow \beta$ seria equivalente a $\langle \alpha, \beta \rangle$. Uma cadeia gerada por G é uma sequência de *terminais* derivada a partir de S através das regras de produção. O conjunto de todas (possivelmente infinitas) cadeias geradas por G é a linguagem $L(G)$. Esta mesma formalização pode ser aplicada, como se vê, para a distinção entre língua-E e língua-I comentada anteriormente.

Classes de linguagens e a hierarquia de Chomsky

Em sua discussão, Chomsky (1959) parte das linguagens recursivamente (ou computavelmente) enumeráveis[13], classificadas por ele como linguagens "tipo 0". Uma gramática correspondente ao tipo 0 teria o poder gerativo de uma máquina de Turing. Assim, poderia gerar cadeias definidas por procedimentos absolutamente arbitrários como, por exemplo, o conjunto de cadeias formadas por um 'a' seguido por um número primo de 'b's ou, mais formalmente, a linguagem $\{ab^m | m$ é um número primo$\}$. Como ressalta Levelt (2008), gramáticas dessa classe – também chamadas de "sistemas de reescritura irrestritos" – são de menor interesse para os estudos linguísticos, dada a falta de restrição sobre os tipos de regras que podem incluir, ou seja, sobre o par $\langle \alpha, \beta \rangle$, como definido acima.

Chomsky propõe então três condições restritoras sobre regras de reescritura e analisa suas consequências em relação às classes de linguagens que delimitam. A primeira condição

[13] Esta classe de linguagens é a classe correspondente às funções computáveis pela máquina de Turing, um dispositivo hipotético capaz de simular a lógica de qualquer algoritmo ou "procedimento efetivo mecânico" (ver Levelt, 2008).

(C1) requer que as regras de produção sejam da forma $\varphi_1 A \varphi_2 \rightarrow \varphi_1 \omega \varphi_2$, onde $\omega \neq \lambda$ (i.e., ω é não-nula). Com isso, garante-se que $A \rightarrow \omega$ no contexto $\varphi_1 \_ \varphi_2$ (que podem ser nulos). A segunda condição (C2) determina que (i) A deve consistir apenas de um não-terminal, isto é, $A \in V_N$, e (ii) que $\omega \neq \lambda$ e $A \rightarrow \omega$, isto é, que o contexto (φ_1 e φ_2) seja nulo. Por fim, a condição (C3) determina que (i) A deve consistir apenas de um não-terminal, isto é, $A \in V_N$, e (ii) que $\omega \neq \lambda$, $A \rightarrow \omega$ (contexto nulo) e que ω seja da forma a ou aB.

Partindo de gramáticas tipo 0 cujas regras de produção tem a forma $\alpha \rightarrow \beta$, temos por aplicação de C1 a classe de gramáticas "tipo 1", chamada de gramáticas *sensíveis ao contexto* (CSGs), cuja única restrição diz respeito ao comprimento dos lados esquerdo e direito. Chomsky (1959, p.148) afirma que gramáticas tipo 1 tem o poder de incorporar permutações através de regras do tipo $AB \rightarrow BA$ que podem ser problemáticas para modelar as línguas naturais. O exempo dado por Chomsky é uma gramática tipo 1 em que tanto *John will come* quanto *will John come* são derivadas a partir de uma linha (passo derivacional) anterior como *Noun Phrase-Modal-Verb* e em que *will John come* é produzida por uma permutação; neste caso, *will* seria dominado pelo nó *Noun phrase*, enquanto *John* seria dominado pelo nó *Modal*, na árvore[14] correspondente à derivação, levando a conclusões inconsistentes com os fatos linguísticos do inglês.[15]

Gramáticas tipo 2 representam a classe obtida a partir da aplicação de C2 sobre as gramáticas tipo 0 e são chamadas de gramáticas *livres de contexto* (CFGs). Todas as regras

[14] Assumindo algumas restrições, Chomsky (1959) mostra que é possível associar derivações a árvores rotuladas, que podem ser tomadas como *descrições estruturais*. Assim, uma subcadeia x na cadeia terminal de uma dada derivação poderia ser chamada de *sintagma do tipo A*, caso pudesse ser rastreada a um ponto rotulado como A na árvore associada.

[15] No entanto, Chomsky (1959) não exclui absolutamente as CSGs como descrições possíveis para as línguas naturais, o que fica evidente quando ele diz

> "Contextual restrictions of this type are often found necessary in construction of phrase structure descriptions for natural languages. Consequently the extra flexibility permitted in type 1 grammars is important. It seems clear, then, that neither Restriction 1 nor Restriction 2 is exactly what is required for the complete reconstruction of immediate constituent analysis." (p.148)

são da forma $A \rightarrow \omega$, em que A é um não-terminal e ω uma cadeia de terminais e não-terminais. Assim, A dominaria a cadeia ω na árvore correspondente. Como o contexto não é considerado nestas regras, gramáticas tipo 2 são menos poderosas que as do tipo 1. CFGs representam a classe de gramáticas de *estrutura sintagmática* (PSGs) tradicionalmente empregadas pela teoria linguística para a análise da estrutura sintática das sentenças das línguas naturais.

Conforme Chomsky, gramáticas desse tipo inibem certas anomalias, tais como as advindas de regras do tipo $ab \rightarrow cd$, em que há a substituição de uma cadeia de itens terminais por outra. Além disso, Chomsky demonstra que qualquer linguagem tipo 2 pode ser gerada por gramáticas tipo 2 cujas regras são da forma $A \rightarrow a$ ou $A \rightarrow BC$, isto é, possuem no máximo dois elementos do lado direito (ou, em relação à árvore, não mais que duas ramificações por nó). Outra característica desse tipo de gramáticas é seu potencial para "auto-encaixamento" ("*self-embedding*"), isto é, derivações do tipo $A \stackrel{\star}{\Rightarrow} \varphi A \psi$ (com φ e ψ não-nulos).

Finalmente, a aplicação de C3 sobre gramáticas tipo 0 determina a classe das gramáticas "tipo 3", menos poderosas que as duas anteriores, visto que as regras são da forma $A \rightarrow aB$ ou $A \rightarrow a$. Tais gramáticas são chamadas de "regulares" ou "lineares à direita" (ou à esquerda, no caso de regras do tipo $A \rightarrow Ba$). Segundo Chomsky, linguagens como $L_2 = xy$ (em que x é uma cadeia de 'a's e 'b's e y é o espelho de x) não podem ser geradas por gramáticas tipo 3. Porém, segundo ele, as línguas humanas exibem propriedades de L_2[16] e, portanto, gramáticas regulares não poderiam caracterizar adequadamente as línguas naturais. Voltarei a este ponto mais adiante.

Uma importante decorrência de Chomsky (1959), é que as quatro classes especificadas podem ser ordenadas pela relação de subconjunto. Gramáticas tipo 1 são um subconjunto estrito das gramáticas de tipo 0; gramáticas tipo 2 são um subconjunto estrito das gramáticas tipo 1; e as gramáticas tipo 3 são um subconjunto estrito das gramáticas tipo 2. Isso implica

[16] Um exemplo de construção desse tipo citado por Gazdar (1988) seria "*Jude is* $|less|_b$ *obviously* $|as|_a$ *nice* $|as|_a$ *Kim* $|than|_b$ *Chris is*", portanto incluindo um padrão do tipo *baab*.

que o *conjunto* de linguagens do tipo n ($n > 1$) é também um subconjunto estrito do conjunto das linguagens de tipo $n - 1$. Em outras palavras, sempre há uma gramática de um tipo mais alto capaz de gerar uma linguagem de um tipo mais baixo, mas não o contrário. Essa ordem ficou conhecida como "hierarquia de Chomsky". Vale ressaltar que a hierarquia não é exaustiva e cada classe definida por Chomsky contém inúmeras subclasses.

Uma última noção envolvendo gramáticas diz respeito à equivalência entre elas. Duas gramáticas G_1 e G_2 são *fracamente equivalentes*, se elas geram linguagens idênticas, isto é, se $L(G_1) = L(G_2)$. Nesse sentido, dada uma linguagem L, pode haver inúmeras (talvez infinitas) gramáticas capazes de gerá-la. Por outro lado, se duas gramáticas geram o mesmo conjunto de árvores de derivação (ou, nos termos de Chomsky (1959), de "descrições estruturais"), então elas são *fortemente equivalentes*. Uma alternativa menos estrita seria assumir que uma relação isomórfica entre as descrições estruturais produzidas por duas gramáticas já seria suficiente para considerá-las fortemente equivalentes (Kornai & Pullum, 1990).

Autômatos finitos

A teoria de autômatos está intimamente ligada à de linguagens formais. Um *autômato* é um dispositivo hipotético concebido como um modelo simples de um computador. Esse dispositivo inicia em um *estado inicial* e, à medida em que lê uma cadeia de símbolos de entrada, pode alternar diferentes *estados* baseado nos símbolos lidos a cada passo. Ao terminar de ler a cadeia, o dispositivo estará num certo estado, que poderá ser de um estado pré-determinado de *aceitação* (chamado "estado final") ou não.

Há dois tipos de autômatos finitos (FSAs): *determinísticos* (DFAs) e *não-determinísticos* (NDFAs). Um DFA é um dispositivo composto por um conjunto finito de estados, vinculados a *arcos* rotulados com símbolos de um alfabeto finito. Cada símbolo lido a partir da cadeia de entrada determina se o dispositivo vai permanecer no mesmo estado ou se vai mudar para outro. Essa transição é unicamente determinada pelo símbolo lido e pelos arcos

rotulados vinculados ao estado. Um DFA pode ser formalmente definido como a quíntupla $\langle K, \Sigma, q_0, F, \delta \rangle$, onde K é um número finito de estados, Σ um alfabeto finito, $q_0 \in K$ é um estado inicial único e δ é uma função de $K \times \Sigma$ para K, que especifica o conjunto de arcos.

Na Figura 2.2, temos um DFA em que $K = \{q_0, q_1\}$, $\Sigma = \{a, b\}$, q_0 é o estado inicial, q_1 o estado final e os arcos são $\langle q_0, a, q_0 \rangle$, $\langle q_0, b, q_1 \rangle$, $\langle q_1, a, q_1 \rangle$ e $\langle q_1, b, q_0 \rangle$. Este DFA reconhece qualquer cadeia formada por um número ímpar de 'b's e qualquer número de 'a's (inclusive nenhum), em qualquer ordem.

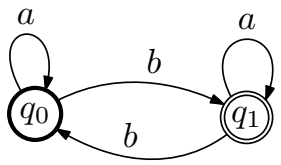


Figura 2.2: Um exemplo de DFA (Hammond, 2010).

Autômatos finitos não-determinísticos são semelhantes aos DFAs, exceto pelo fato de que (i) arcos podem ser rotulados com o símbolo nulo ϵ, e (ii) um estado pode estar vinculado a mais de um arco rotulado pelo mesmo símbolo. Portanto, em termos matemáticos, δ é uma relação (um subconjunto finito de $K \times (\Sigma \cup \epsilon) \times K$), não uma função. Conforme Hammond (2010), assim definido, fica claro que DFAs são um subcaso dos NDFAs e, portanto, pode-se demonstrar trivialmente que NDFAs podem gerar todas as linguagens geradas por DFAs. Mas a equivalência também vai na direção contrária, ou seja, pode-se demonstrar que ambos são equivalentes (ver Hopcroft et al., 2001).

Por ser não-determinístico, um NDFA permite vários caminhos na análise de uma cadeia. Alguns caminhos serão parciais (sem-saída), outros serão completos, porém poderão terminar em estados finais (de aceitação) ou não. Para que uma cadeia seja considerada aceita, basta que haja pelo menos um caminho que leve a um estado final. A Figura 2.3 exemplifica um NDFA para a linguagem cujas cadeias começam com pelo menos um 'b' e para qualquer instância de 'a', deve haver pelo menos um 'b' de cada lado:

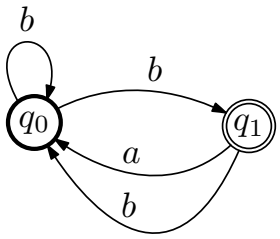


Figura 2.3: Um exemplo de NDFA (Hammond, 2010).

Autômatos finitos podem gerar precisamente a classe das linguagens regulares (tipo 3), particularmente, as gramáticas lineares à direita. Note que é possível emparelhar as regras da gramática com os estados e os arcos de um autômato: cada regra da forma $A \rightarrow xB$ equivale a um arco rotulado com x que vai do estado A para o estado B e regras da forma $A \rightarrow x$ equivalem a arcos rotulados com x que vão do estado A para um estado final (F) designado. Desse modo, para qualquer gramática G_i tipo 3 é possível especificar um autômato finito correspondente (DFA ou NDFA).

Gramáticas livres de contexto também são equivalentes a um dispositivo particular, chamado de "autômato com pilha"[17] (PDA). A diferença de um PDA para um FSA é que o primeiro inclui uma *pilha*, uma espécie de memória utilizada para armazenar certos elementos na medida em que a cadeia é consumida. A única restrição imposta é a de que elementos só podem ser acessados, inseridos ou retirados do topo da pilha (estratégia "último a entrar, primeiro a sair" ou LIFO[18]). Uma cadeia é aceita por um PDA mediante três condições: (i) a entrada foi completamente consumida; (ii) a pilha está vazia; e (iii) o PDA está em um estado final.

Formalmente, um PDA é a sêxtupla $\langle K, \Sigma, \Gamma, s, F, \Delta \rangle$, onde K é um número finito de estados, Σ um alfabeto finito (de entrada), Γ um alfabeto finito da pilha, $s \in K$ é um estado inicial único, $F \subseteq K$ é o conjunto de estados finais e Δ é o conjunto de transições, um subconjunto finito de $K \times (\Sigma \cup \epsilon) \times (\Gamma \cup \epsilon) \times K \times (\Gamma \cup \epsilon)$. Algumas especificações

[17] *Pushdown automaton.*
[18] *Last in first out.*

formais de PDAs incluem também o elemento $Z \in \Gamma$, o símbolo inicial da pilha (quando do estado inicial s). PDAs, assim como FSAs, podem ser determinísticos, gerando a classe de linguagens livres de contexto (CFLs) determinísticas, ou não-determinísticos, gerando a classe de CFLs não-determinísticas.

Dispositivos para linguagens do tipo 0 (máquinas de Turing) e 1 (autômatos linearmente limitados) também são investigados, mas não serão considerados aqui, visto que o foco do presente trabalho recai sobre gramáticas de estrutura sintagmática e, portanto, especialmente sobre os PDAs. Para mais informações sobre estes e outros dispositivos, ver Hopcroft et al. (2001), Levelt (2008) e Hammond (2010).

2.3.2 PSGs como descrições das línguas naturais

Desde Chomsky (1956), pelo menos, considera-se que as regularidades da sintaxe das línguas naturais, pelo menos em sua *base*, isto é, as construções canônicas da língua (p.e., em que não há elipses, elementos omitidos ou aparentemente deslocados), pode ser descrita por gramáticas livres de contexto, mais particularmente as PSGs mencionadas anteriormente. Conceitualmente, as palavras da língua compõem o conjunto de terminais (V_T) de uma gramática, enquanto as categorias gramaticais compõem o conjunto de não-terminais (V_N). Como se sabe, PSGs estão na base da maior parte dos formalismos gramaticais já propostos e os atualmente utilizados (ver Partee et al., 1993, Vijay-Shanker & Weir, 1994, Steedman & Baldridge, 2011, entre outros).

PSGs estritamente livres de contexto, entretanto, apresentam limitações. De um lado, temos limitações que as impedem, por exemplo, de capturar de um modo elegante relações de concordância entre elementos da sentença ou generalizações envolvendo relações entre construções, como por exemplo a relação entre sentenças ativas e passivas ou entre sentenças declarativas e interrogativas formadas a partir delas (Chomsky, 1957, 1959). Por outro lado, há argumentos defendendo – como o próprio Chomsky (1959) sinalizou – um caráter

parcialmente sensível ao contexto nas línguas naturais (Shieber, 1985, Joshi et al., 1990). Se tais argumentos estiverem corretos, as PSGs se tornam por definição insuficientes para expressar tais propriedades.

Vale ressaltar que a afirmação sobre a inadequação das PSGs para descrever generalizações relevantes e sobre o caráter sensível ao contexto das línguas naturais foram e continuam sendo questionados e debatidos (Pullum & Gazdar, 1982, Gazdar, 1983, Mohri & Sproat, 2006, entre outros). De todo modo, parece haver uma concordância geral de que PSGs precisam de extensões – por menores que sejam – para que possam lidar mais adequadamente com as generalizações linguísticas mais significativas. Nesse sentido, duas abordagens alternativas se destacam: a transformacional e a não-transformacional. A principal característica da abordagem transformacional – originalmente proposta por Chomsky (1957) – é a assunção de duas componentes na gramática das línguas: a componente de *base* e a componente *transformacional*.

A primeira seria essencialmente uma PSG, responsável por gerar as estruturas de base das sentenças, chamada *estrutura profunda*. A segunda seria a responsável por derivar novas estruturas a partir de estruturas de base, chamadas de *estruturas superficiais*, através de operações e de restrições gerais sobre estas operações. Embora simplificassem bastante a expressão de regularidades linguísticas as mais diversas, os modelos transformacionais iniciais exibiam um poder expressivo demasido grande para a descrição das línguas naturais (Peters & Ritchie, 1971, 1973), razão pela qual passaram por diversas revisões até serem substituídos por um modelo mais restrito (Chomsky, 1986, 1995b), em que há apenas uma operação transformacional, *Mova-*α, cuja ação seria restringida por princípios da GU.

Alternativamente, outras propostas de extensão da PSG surgiram, propondo uma abordagem não-transformacional. A ideia era estender o mínimo possível, de modo a não aumentar demasiadamente o poder expressivo do formalismo. Entre os vários formalismos propostos, há aqueles baseados em gramáticas de unificação, tais como a *gramática léxico-*

funcional (LFG) (Bresnan, 2001) e a HPSG[19] (Pollard & Sag, 1994). Outros formalismos incluem a *gramática de adjunção de árvores* (TAG) (Joshi et al., 1990) e a *gramática categorial combinatória* (CCG) (Steedman & Baldridge, 2011).

Uma virtude destes formalismos face à presente pesquisa é sua maior adequação à modelagem computacional, por apresentarem uma formalização mais rígida. Além disso, ao contrário do modelo transformacional, há classes de máquinas de estado (extensões dos PDAs) bem definidas para tais formalismos (Vijay-Shanker & Weir, 1994), o que permite uma melhor compreensão de aspectos envolvendo eficiência e complexidade computacional. É interessante mencionar a observação de Steedman & Baldridge (2011) à respeito das similaridades entre os formalismos não-transformacionais mencionados. Segundo os autores, todos eles assumem essencialmente a mesma teoria de ligação e relações de escopo semântico.

A consequência disso, segundo Steedman & Baldridge, é que a análise de fenômenos envolvendo dependências vinculadas ao domínio verbal flexionado, tais como alçamento, controle, passivização, reflexivização, entre outros, seria relativamente similar em qualquer dos formalismos. As diferenças mais significativas entre eles emergem quando se consideram dependências que extrapolam os limites da oração flexionada, tais como orações relativas ou estruturas coordenadas envolvendo lacunas. Em termos de poder expressivo, Vijay-Shanker & Weir (1994) demonstraram que os formalismos em questão são fracamente equivalentes, ou seja, geram a mesma classe de linguagens (lembrando que linguagens se referem ao conjunto das sentenças e não das descrições estruturais).

Um último ponto a ressaltar é o fato de que a classe das línguas naturais não corresponde estritamente a um dos tipos da hierarquia de Chomsky. Como destacam Berwick & Weinberg (1984), é possível delimitar subclasses na hierarquia, exatamente o que fazem os formalismos indicados nesta seção. Em geral, acredita-se que a classe das línguas naturais está em algum ponto entre as linguagens livres de contexto e as sensíveis ao contexto, daí a

[19] *Head-driven phrase structure grammar.*

expressão "gramática moderadamente sensível ao contexto" (Joshi et al., 1990). Vários fatores – para além dos estritamente linguísticos – podem contar para a determinação da classe de linguagens mais adequada, inclusive fatores de ordem computacional, tais como complexidade e analisabilidade eficiente (ver Berwick & Weinberg, 1984, para uma discussão destes aspectos). Todos tem sua importância e devem ser levados em conta na análise comparativa dos formalismos.

2.4 Aprendibilidade

Até aqui, vimos que o conhecimento de uma língua, doravante *competência* do falante, se caracteriza pela propriedade de gerar infinitas cadeias simbólicas (a língua-E) a partir de meios finitos (a gramática). Vimos ainda que a competência exibe propriedades formais que a identificam em grande parte com as gramáticas de estrutura sintagmática e que mínimas extensões à PSG poderiam dar conta das demais generalizações linguísticas que escapam a uma descrição significativa por PSGs estritas. Esta hipótese surge como resposta à primeira das três perguntas basilares para os estudos em gramática gerativa Chomsky (cf. 1986, p.3), isto é, "*O que constitui o conhecimento da linguagem?*".

Assumindo que esta seja a natureza formal do conhecimento gramatical, podemos então dar o segundo passo e enfrentar a questão relativa às origens de tal conhecimento, isto é, "*Como este conhecimento da linguagem é adquirido?*". Responder a esta pergunta implica, segundo Chomsky, atingir o grau de "adequação explicativa" nos estudos da linguagem, superando o grau de "adequação descritiva" em que a hipótese sobre a natureza da gramática das línguas humanas nos localiza. Um caminho para abordar o problema da aquisição surgiu com os estudos sobre *aprendibilidade*. Como explica Yang (2002), *aprendibilidade* é o estudo matemático da aprendizagem de linguagem, vista aí como a descoberta de uma função computável que produza uma gramática[20] a partir de um conjunto de exemplos.

[20] Ou produza todo o conjunto de cadeias pertencentes a uma linguagem.

A pergunta básica para as teorias de aprendibilidade é: dado um conjunto de exemplos de uma língua (ou linguagem formal) qualquer, seria possível chegar a um gramática aceitável para a linguagem? Segundo Levelt (2008), não há até os dias atuais uma resposta definitiva para esta questão, da forma como está colocada. No que concerne às línguas naturais e ao processo de aquisição, as evidências indicam que sim, que tem que haver um procedimento capaz de levar a criança a adquirir uma língua, com base na experiência a que está exposta. Qualquer resposta à pergunta colocada acima depende, segundo Levelt, (1) do que se sabe sobre a gramática, (2) da composição dos dados de amostragem, e (3) o que se entende por "aceitável".

Yang (2002) também particiona o problema em vários componentes, tais como a apresentação dos dados de entrada, a composição do espaço de hipóteses, o mecanismo e a complexidade do algoritmo de aprendizagem, a condição de convergência, entre outros. Segundo Nowak et al. (2002), a aprendizagem da linguagem se caracteriza como uma *inferência indutiva*[21]: crianças generalizam para além de sua experiência, visto que a experiência é finita e o conhecimento adquirido se caracteriza pela infinitude. Cabe à teoria de aprendibilidade descrever a matemática desse processo, determinando as condições para que apenas as generalizações corretas sejam feitas. Para um panorama amplo do problema da aprendibilidade envolvendo as línguas naturais, ver Pinker (1979), Bertolo (2001), Nowak et al. (2002), Kaplan et al. (2008), Heinz (2010), Pearl (2010) e Yang (2011), entre outros.

Um trabalho basilar nessa área é o de Gold (1967) (ver Heinz, 2010, para uma discussão crítica deste trabalho), em que o autor demonstrou alguns fatos paradigmáticos a partir do estudo dos aspectos formais envolvidos na aprendibilidade das classes de gramáticas definidas na hierarquia de Chomsky. Em primeiro lugar, Gold mostrou que a aprendizagem é, em geral, impossível, se for assumido que (i) a classe de linguagens é mais abrangente do que a classe

[21] Berwick et al. (1992) apresenta investigações computacionais sobre abordagens *dedutivas* para a análise sintática. É possível que tais abordagens, se desenvolvidas nessa direção, levassem a algum modelo de aquisição também de caráter dedutivo. Entretanto, até onde sei, tais estudos não tiveram desenvolvimentos posteriores relevantes, pelo menos no que diz respeito ao processamento e aquisição das línguas naturais.

de linguagens de cardinalidade finita, (ii) que o aprendiz tem acesso apenas a dados positivos (i.e., apenas sentenças pertencentes à língua-alvo), e (iii) que o aprendiz parte para a tarefa de indução sem quaisquer pré-assunções sobre as propriedades específicas da classe a que pertence a linguagem indicada pelos dados.

Em outras palavras, apenas línguas[22] definidas por um conjunto finito de sentenças poderiam ser aprendidas com base apenas em dados positivos. Neste caso, evidentemente, bastaria que o aprendiz fosse apresentado a todas elas. Segundo esta conclusão e diante do que discutimos acima sobre os dados de entrada disponíveis à criança, a aquisição das línguas naturais seria impossível, o que não é o caso, evidentemente. Por outro lado, Gold mostrou que a aprendizagem seria possível se para classes mais altas o aprendiz for apresentado também à informação *negativa*, isto é, sentenças não pertencentes à língua-alvo explicitamente marcadas como agramaticais.

Além disso, o aprendiz teria que ter memória infinita (lembrança total de sentenças já recebidas), o que o permitiria enumerar todas as hipóteses possíveis e testar uma por vez contra o conjunto de sentenças recebido até ali, para determinar a hipótese mais correta para descrição dos dados. Outros estudos, como o de Valiant (1984), abordaram a questão flexibilizando o critério de sucesso, de modo que o aprendiz não precisaria necessariamente identifcar a gramática/língua exata, bastando se "aproximar" dela. Os resultados, no entanto, continuaram bastante negativos em face das particularidades da aquisição da linguagem humana.

Tais resultados, especialmente os de Gold, são particularmente significativos se considerarmos o fato de que crianças aprendem línguas naturais em um tempo finito e relativamente curto, com dados restritos (no sentido chomskiano da "pobreza de estímulos"), sem (aparentemente) acesso a dados negativos e com recursos cognitivos limitados, como a memória, por exemplo. Além disso, línguas naturais são conjuntos infinitos de sentenças, que apresen-

[22] A partir daqui, com referência também às demais classes de linguagens formais e não apenas às línguas naturais.

tam tanto propriedades livres de contexto, como propriedades sensíveis ao contexto, como já vimos (ver Anderson, 1978, Pinker, 1979, Partee et al., 1993, Joshi, 2003, Heinz, 2010, entre outros).

Note, entretanto, que Gold (1967) não estava descrevendo uma criança adquirindo uma língua natural, mas sim a aprendibilidade dada das características das grandes classes formais da hierarquia de Chomsky. Por essa razão, suas conclusões não devem ser estendidas de qualquer maneira ao problema da aquisição. Ao contrário, elas apontam para componentes específicos do problema que precisam ser pensados, tais como a condição de sucesso, as propriedades dos dados de entrada e, particularmente, as propriedades específicas da classe das línguas naturais. Nada impede que as restrições inerentes à classe das línguas naturais em conjunto com propriedades dos dados de entrada disponíveis à criança possam delimitar o espaço de hipóteses da criança de modo a tornar a aquisição possível com base em dados positivos.

Essa é exatamente a linha de raciocínio que tem orientado as investigações formais e computacionais sobre aquisição, desde a década de 1970 (Pinker, 1979). O estudo sobre a natureza das línguas naturais visa exatamente fornecer subsídios que limitem o espaço de hipóteses considerado pela criança na tarefa de determinar a gramátiva-alvo adequada. Em outras palavras, é preciso identificar os aspectos da linguagem que a criança não precisa aprender, ou seja, que seriam dados *a priori*, determinados por restrições inatas exclusivas à espécie. Em relação aos dados de entrada, hipóteses como a de alavancagem semântica ou, ainda, a de alavancagem prosódica (Jusczyk, 1997, Mazuka, 1998, entre outros), indicam que os dados contém mais informações do que a mera cadeia não-analisada de palavras.

Quanto à alavancagem prosódica, Mazuka (1998) argumenta em favor de um parâmetro prosódico, anterior à aquisição sintática, que forneceria à criança a ordem básica de ramificação da língua, isto é, se sua língua é essencialmente núcleo-inicial (como o inglês ou o português, por exemplo) ou núcleo-final (como o japonês). Além disso, Jusczyk (1997)

cita estudos segundo os quais a criança teria pistas para segmentar alguns pontos da cadeia linguística, que lhe permitiriam isolar, por exemplo, sintagmas nonimais ou verbais, pelo menos. Não que isso seja absolutamente necessário, visto que os corpora de dados de aquisição indicam que a criança é bastante exposta a sintagmas isolados, especialmente nonimais. Porém, com o auxílio das pistas prosódicas, a criança teria facilitada sua tarefa de aprender a segmentar o fluxo da fala em unidades mínimas semanticamente significativas.

Langley (1987) classifica o conjunto de abordagens que assumem alavancagem semântica como "paradigma de mapeamento gramatical", visto que envolve um mapeamento que relaciona os símbolos que compõem a sentença às informações semânticas. Langley ressalta um aspecto importante sobre a plausibilidade dessa abordagem: embora nos primeiros estágios da aquisição os enunciados dirigidos à criança tendam a estar fortemente ancorados no contexto discursivo imediato, os mais tardios envolvem aspectos semânticos mais abstratos e não poderiam ser explicados pela alavancagem semântica. O autor conclui, então, que embora a inferência gramatical estrita (que considera apenas as sentenças) possa não ser o modelo ideal para os primeiros estágios, parece um modelo viável para os estágios mais tardios.

Outra hipótese importante relacionada à natureza dos dados de entrada, conhecida como hipótese da "aprendibilidade grau-2" (Culicover & Wexler, 1980), prevê que a criança não necessita de todos os dados de entrada, mas apenas do conjunto formado por sentenças com *até* dois níveis de subordinação, como em [$_0$ *O João disse que* [$_1$ *a Maria pensa que* [$_2$ *o Pedro gosta dela*]]]. Os autores apresentam um estudo formal sobre detectabilidade de erros nas línguas naturais, isto é, detectabilidade de distinções entre contextos sintáticos, argumentando que todas as regularidades gramaticais das línguas naturais teriam expressão nesse conjunto de sentenças. Ao mesmo tempo que esta hipótese restringe o conjunto de dados de entrada, ela também restringe a classe das línguas naturais, ou seja, restringe o espaço de hipóteses da criança.

Finalmente, uma terceira origem possível para as restrições, embora pouco considerada

em modelos formais de aquisição, estaria nas propriedades do processamento humano da linguagem, tais como a limitação de memória e outras propriedades do analisador sintático mental (Frazier & Fodor, 1978, Altmann & Steedman, 1988, Clifton et al., 1991, Kaiser & Trueswell, 1994, entre outros). Todos esses fatores tem papel potencialmente restritor no processo de aquisição, cabendo às investigações formais e computacionais fornecer subsídios para uma melhor delimitação da relevância de cada um. Por ser esta uma discussão muito extensa e em função de que o presente estudo está focado na aquisição da competência (e não em aspectos do processamento), estas questões tiveram que ser deixadas de lado, restando como tópicos para futuras investigações.

2.5 Conclusão: o LAD revisado

O que foi discutido aqui nos permite tirar algumas conclusões sobre possíveis características a serem exibidas pelos componentes do LAD, detalhando-os um pouco mais, como mostra a Figura 2.4:

Primeiramente, a GU predispõe o aprendiz sobre a natureza livre de contexto (estendida, "e-CFG") ou moderadamente sensível ao contexto ("m-CSG") do conhecimento gramatical. Com relação aos dados de entrada, assume-se que eles compõem uma classe bem definida (hipótese de aprendibilidade grau-2) e que contém informação para além da cadeia de palavras, se assumirmos as alavancagens semânticas e prosódica. Com relação ao analisador, vimos que é importante limitar seu poder de processamento, o que vai variar de acordo com sua arquitetura. Mas se adotarmos o PDA como o modelo de processamento do aprendiz, tais limitações podem ser especificadas sobre duas estruturas do autômanto: a "fita" (ou "área temporária") e a pilha.

Com relação à primeira, pode-se limitar a capacidade de antecipação[23] de elementos da cadeia para além do elemento momentâneamente em análise a, por exemplo, mais 2

[23] *Lookahead*.

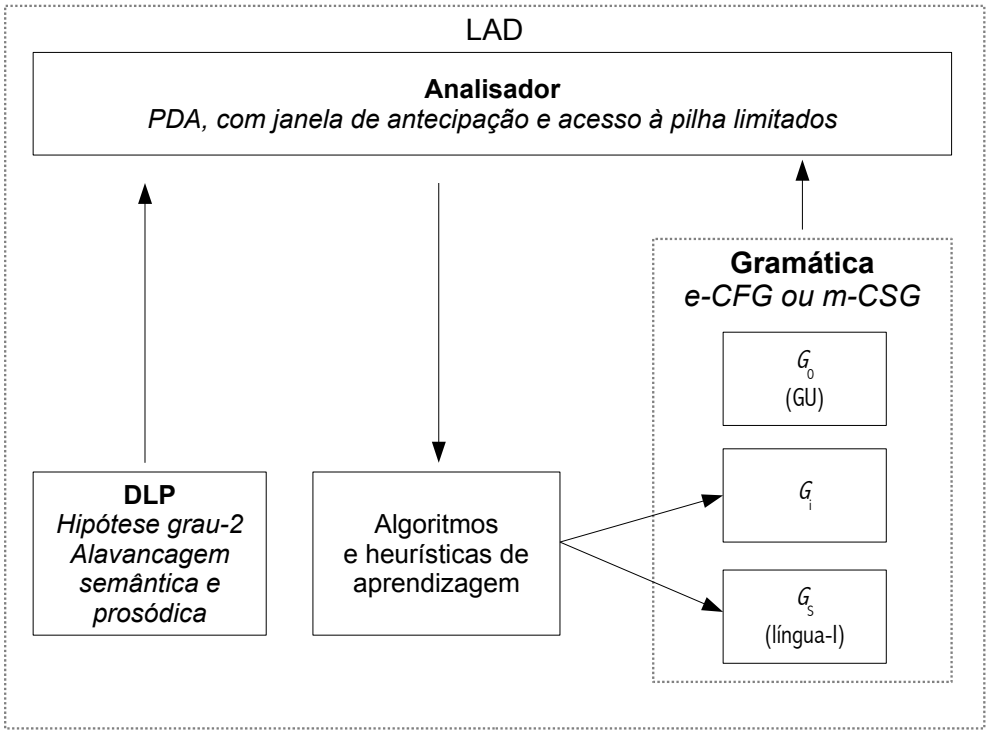


Figura 2.4: Esquema atualizado do LAD

elementos, o que comporia uma "janela" de análise de 3 elementos. Com relação à pilha, pode-se limitar o número de elementos visíveis num determinado momento da análise, restrição que está relacionada a aspectos de *localidade* das relações sintáticas e que pode ser também relacionada à hipótese de aprendibilidade grau-2. Tais restrições são consideradas psicologicamente plausíveis (Marcus, 1980, Pinker, 1984, Berwick, 1985, entre outros).

Como veremos no próximo capítulo, esta é a arquitetura básica dos modelos computacionais de aquisição, no que tange especificamente à aprendizagem (alguns modelam também a produção de sentenças). Na segunda parte do texto, veremos como o IASMIM aborda e implementa cada uma destas componentes. Enfim, espera-se que a modelagem lance luz sobre a adequação desta arquitetura (p.e., quanto aos módulos assumidos e o tipo de interação entre eles) como modelo do dispositivo de aquisição da linguagem.

3

Modelagens computacionais de aquisição

3.1 Visão geral

Quando falamos em aquisição de primeira língua, estamos nos referindo não a um processo específico e homogêneo, mas a um processo complexo, heterogêneo, em que diferentes aspectos linguísticos estão em jogo. Como se pode depreender a partir do capítulo anterior, para que uma criança atinja o conhecimento de um adulto, é preciso que ela resolva uma série de problemas, dentre os quais podemos destacar a identificação do padrão prosódico de sua língua e dos fonemas relevantes, em que pontos do fluxo da fala se encontram fronteiras de palavras, o que estas palavras significam, quais são as categorias sintáticas e como combiná-las.

Cada um destes problemas é complexo por si só e impõe desafios significativos para modelagens computacionais. Em uma modelagem computacional, Marr (1982, apud Pearl, 2010) explica que é preciso lidar com três níveis de processamento da informação: o nível *computacional* (relativo à descrição do problema), o nível *algorítmico* (relativo aos passos necessários para sua solução) e o nível *implementacional* (de que forma o algoritmo é implementado). Do ponto de vista linguístico, os níveis de principal interesse são os dois primeiros, o computacional e o algorítmico, enquanto o último nível diz respeito à "engenharia"

da modelagem.

Em relação ao nível computacional, Pearl (2010) e Yang (2011) enfatizam a necessidade de um suporte teórico adequado, que permita a descrição do conhecimento a ser adquirido e de como será representado. Vê-se, assim, que este nível é o que mais dialoga com a teoria (psico)linguística, pois é aí que se encontram os modelos teóricos para os diversos aspectos da linguagem envolvidos no problema. Por exemplo, o caráter transformacional ou não do conhecimento gramatical a ser adquirido depende fundamentalmente das teorias assumidas. Este nível também dialoga fortemente com as teorias de aprendibilidade, para especificação dos critérios de aprendizagem e sucesso, no que tange à aquisição.

Pearl (2010) ressalta ainda a necessidade de um suporte empírico que permita especificar dados de entrada plausíveis, bem como das informações necessárias para a interpretação dos mesmos. Este suporte deve ser suficiente para que sejam estabelecidos parâmetros de avaliação do modelo, que vão desde parâmetros que determinam o sucesso ou insucesso na aquisição, até parâmetros desenvolvimentais, que permitam comparar o comportamento do modelo ao do aprendiz durante o processo de aquisição (tipos de erros cometidos, ordem na aquisição etc.).

O nível algorítmico, por sua vez, encarna a teoria de aprendizagem assumida ou proposta pelo modelo e é aí que são especificados os procedimentos de aquisição responsáveis por induzir (ou deduzir) a gramática a partir dos dados de entrada. Outra questão importante nesse nível é definir, por exemplo, qual mecanismo desencadeará os procedimentos de aquisição. Parte dos modelos assume um paradigma de aprendibilidade semelhante ao estudado em Gold (1967), em que a aprendizagem se configura como a indução de gramáticas baseada na exposição apenas a exemplos positivos da língua. Neste caso, desconsidera-se evidência negativa (i.e., exemplos agramaticais marcados como tal).

Outros modelos, alternativamente, assumem evidência negativa através da figura de um "oráculo", isto é, um componente do modelo que simula um falante adulto, capaz de

fornecer juízos de gramaticalidade sobre expressões hipotetizadas e produzidas pelo aprendiz. São, portanto, modelos interativos. Com base nesta interação, o aprendiz corrige, descarta ou reforça hipóteses sobre a gramática. Tais modelos tem também como objetivo modelar a produção de sentenças e não apenas a compreensão. Os modelos, de modo geral, podem ser concebidos para fazer uso de evidência (negativa) indireta, isto é, tomar a ausência de certos exemplos (hipoteticamente possíveis) como evidência de sua agramaticalidade.

Em modelos não-interativos, é a análise dos dados de entrada que permite disparar os procedimentos de aquisição. Nestes, uma falha na análise de um dado – assumido como sendo sempre positivo – é interpretada como evidência de que a gramática é ainda insuficiente, sendo necessário disparar procedimentos de aquisição. Tais modelos são conhecidos como modelos "baseados em erro". Nalguns desses modelos, o analisador sintático tem um papel fundamental na determinação de propriedades do que é aprendido, além de ser também um *locus* de garantia de plausibilidade psicológica, em função das restrições que podem ser colocadas sobre sua capacidade de processamento.

Finalmente, há duas grandes abordagens para lidar com os níveis computacional e algorítmico da modelagem: a abordagem simbólica (determinística ou probabilística) e a abordagem não-simbólica, baseada noutros mecanismos de aprendizagem, tais como as redes neurais artificiais, as abordagens genéticas ou evolutivas, entre outras (ver Jain et al., 1999). De um modo geral, a abordagem simbólica reflete uma visão modular da cognição humana, enquanto as não-simbólicas assumem mecanismos gerais de aprendizagem e de representação do conhecimento. Neste capítulo, a atenção está voltada especialmente para modelos simbólicos, isto é, MCAs que assumem um formalismo gramatical explícito. Para uma visão geral sobre MCAs não-simbólicos, ver Seidenberg (1997), Kaplan et al. (2008), Frank (2011) e Yang (2011).

3.1.1 Avaliação de modelos

Os modelos e modelagens considerados neste capítulo serão discutidas de modo mais qualitativo, de modo a indicar pontos fortes e fracos de cada um. Vários fatores dificultam uma avaliação comparativa e quantitativa de MCAs, a saber:

- *Corpora particulares.* É comum os modelos serem acompanhados de um corpus particular, preparado especialmente para as necessidades da modelagem em questão. Assim, por exemplo, enquanto num modelo o corpus é um texto artificial, gerado a partir de palavras combinadas aleatoriamente e sem nenhum tipo de pré-segmentação (Wolff, 1975), noutro o corpus é composto por sentenças pré-segmentadas em conjunto com sua forma lógica e o aprendiz dispõe de um pré-conhecimento sobre categorias sintáticas básicas, tais como S (sentença), N (núcleo nominal), PP (sintagma preposicional) etc. (Villavicencio, 2002). A razão para estas diferenças é que as assunções teóricas determinam, inclusive, a natureza dos dados de entrada. Assim, dificilmente é possível submeter um mesmo corpus a diferentes modelos, de modo a comparar sua performance.

- *Formalismo gramatical.* Modelos variam em relação ao formalismo assumido para o conhecimento gramatical. Alguns articulam regras transformacionais e a representação X-barra (Berwick, 1985), enquanto outros assumem um formalismo não-transformacional, como a LFG (Gaylard, 1995). Há modelos semi-paramétricos, como o de Villavicencio (2002), e os que abstraem completamente de formalismos gramaticais, optando por representar sentenças e gramáticas através de sequências de parâmetros binários (Yang, 2002). Há, ainda, modelos que assumem uma representação própria, particular do conhecimento linguístico do aprendiz (Langley, 1982, Selfridge, 1986). Formalismos nem sempre são diretamente comparáveis, visto que muitas vezes partem de pressupostos diferentes sobre a natureza do conhecimento linguístico e não é adequado, portanto, reduzi-los a meras opções formais.

- *Abordagens de aprendizagem.* Os modelos também variam bastante em relação às abordagens. Alguns assumem evidência negativa explícita (exemplos agramaticais marcados como tal), outros evidência negativa indireta (ausência de certas construções), outros apenas exemplos positivos. Além disso, alguns envolvem uma aquisição determinística (Berwick, 1985, Gaylard, 1995), enquanto outros a modelam como um processo probabilístico (Villavicencio, 2002) que visa a melhor descrição possível dos dados.

- *Escopo.* O escopo do problema de aprendizagem modelado também varia consideravelmente entre os modelos. Alguns estão focados na segmentação do enunciado em unidades linguísticas (palavras ou morfemas) (Wolff, 1975), enquanto outros focam a aquisição lexical (Siskind, 1996). Há modelos que abstraem questões de segmentação e aquisição lexical (Berwick, 1985), enquanto outros integram aquisição lexical e sintática (Gaylard, 1995, Villavicencio, 2002). Novamente, a comparação direta entre os modelos é bastante difícil, visto que os modelos assumem pontos de partida distintos.

Portanto, opto aqui por uma análise de cunho qualitativo, em que os modelos são discutidos em face dos objetivos a que se propõem, de certos critérios formais de aprendizagem e de plausibilidade psicológica e desenvolvimental. De um modo geral, os modelos concebidos devem ser formalmente *suficientes*, isto é, devem aprender o que se propõem; devem apresentar *compatibilidade desenvolvimental*, isto é, ser psicologicamente plausíveis em termos das capacidades de processamento pressupostas; devem ter poder *explanatório*, isto é, em que medida o aspecto crucial do modelo impacta a teoria; além de serem *abrangentes* e *universais* (Pearl, 2010, Yang, 2011).

Em outros termos, Pinker (1979) também define critérios de avaliação, que em sua visão seriam seis: as condições (i) de *aprendibilidade*, (ii) de *equipotencialidade*, (iii) de *entrada*, (iv) de *tempo*, (v) *desenvolvimental* e (vi) *cognitiva*. Segundo Pinker, o aprendiz deve aprender o que se espera (aprendibilidade) em qualquer língua (equipotencialidade), com dados de

entrada equivalentes aos da criança (entrada), no tempo que ela leva (tempo), cometendo os mesmos erros da criança ao longo do caminho (desenvolvimental) e com os recursos cognitivos que a criança tem à disposição (cognitiva). Entretanto, apesar de pertinentes, nem todos estes critérios são possíveis de avaliar na prática.

Frank (2011), por exemplo, questiona as condições de tempo e a cognitiva, argumentando que não é óbvio de que modo se pode avaliá-las num modelo computacional, o que é de fato uma questão. Embora existam limitações de processamento na criança para as quais há uma relativa concordância, tal como a de que a memória é imperfeita (incapaz de relembrar dados vistos anteriormente, para reanálise), há outras para as quais ainda não há consenso (por exemplo, se a criança é capaz de utilizar informações sobre probabilidades transicionais) (cf. Frank, op.cit.). Ainda assim, em geral a condição cognitiva é interpretada como uma orientação: se for possível supor um menor poder de processamento, melhor.

A condição de tempo, por sua vez, é problemática de ser medida, visto que nenhuma simulação computacional ficará executando por cinco ou seis anos. Talvez, um meio de simular o tempo seja estabelecê-lo na forma do número de sentenças apresentadas, desde que possamos ter uma estatística confiável do número de sentenças médio (ou mínimo) que uma criança ouve nesse período de tempo. A partir dessa informação, seria possível por exemplo submeter o corpus repetidamente de modo a atingir o mesmo número de sentenças. Porém, tal estratégia seria equivalente a assumir que a distribuição e a qualidade dos dados de entrada para a criança não mudam no decorrer da aquisição, o que não parece plausível.

Assim, a avaliação dos modelos acaba incidindo mais sobre os critérios de aprendibilidade, de entrada, compatibilidade desenvolvimental e cognitiva. O critério de equipotencialidade é considerado importante, mas até o momento nenhum dos modelos de aquisição sintática demonstrou performance translinguística significativa. Os modelos foram testados basicamente sobre dados do inglês. Modelos paramétricos tem o potencial de serem universais, dado o alto grau de abstração que assumem na representação dos dados e da gramática.

O problema de tais modelos é a atual ausência de propostas concretas sobre o conjunto de parâmetros necessários para modelar as línguas naturais.

3.2 Modelagens computacionais

3.2.1 Modelos interativos

Langley (1982) apresenta o AMBER, um MCA em que o aprendiz melhora sua performance através de procedimentos para recorrer de erros. O AMBER se caracteriza por ser um sistema de produção adaptativo, que inicia com a capacidade de produzir enunciados de uma palavra e vai combinando mais palavras na medida em que novas regras são criadas. O objetivo é combinar várias palavras e morfemas na ordem correta e sem erros de omissão ou comissão[1]. A aprendizagem se dá a partir da comparação entre sentenças previstas (produzidas pelo aprendiz) e observadas (produzidas pelo adulto). Sempre que um erro é encontrado, o aprendiz produz regras mais conservadoras, contendo mais condições.

Uma regra não passa a ser utilizada logo após criada, mas sim após ser hipotetizada um certo número de vezes, o que garante à regra a "força" necessária para entrar na produção de sentenças. Com isso, o AMBER aprende gradualmente, passando por estágios contendo cada vez mais palavras e morfemas e consegue mimificar em parte a ordem de aquisição de alguns morfemas do inglês (por exemplo, adquirindo o progressivo antes dos auxiliares). O modelo distingue entre palavras de "conteúdo" e funcionais, utilizando informações sobre pausa (codificadas nos dados de entrada) para determinar os limites dos sintagmas e, assim, identificar a quais palavras correspondem certos morfemas, simulando o papel da informação prosódica na aquisição.

As limitações do modelo incluem a ausência de aquisição lexical (palavras e significados)

[1] Erros de omissão são aqueles em que o aprendiz deixa de responder (ou produzir) algo que deveria, enquanto erros de comissão, ao contrário, são aquelas situações em que o aprendiz produz algo que não deveria produzir.

e a incapacidade de lidar com formas irregulares. O conhecimento adquirido consiste de sentenças declarativas, contendo ou não auxiliares e com algumas variações, como verbos no progressivo. O modelo converge com um corpus relativamente pequeno de sentenças, adquirindo os auxiliares (o último estágio do modelo) por volta dos 300 exemplos. Antes do estágio final, o aprendiz atinge uma MLU igual à 7 após 70 exemplos e a taxa de erros de omissão e comissão cai para zero após 80 exemplos.

Embora consiga mimificar aspectos da aquisição, o AMBER carece de uma base formal linguisticamente relevante para a representação do conhecimento gramatical, que no modelo consiste basicamente em ordenar (linearmente) sintagmas e os morfemas contidos nestes. As sentenças não apresentam estrutura hierárquica e os sintagmas (que aí são apenas sequências de palavras e morfemas) estão correlacionados com papéis temáticos básicos, como *agente*, *objeto* e *ação*). Portanto, a aquisição no AMBER está mais para aprender a repetir o que ouve, do que adquirir de fato um sistema linguístico mais rico e capaz de interpretar e produzir exemplos inéditos.

Selfridge (1986) apresenta o CHILD, um modelo que também pressupõe a interação entre um adulto e a criança e é capaz tanto de interpretar, quanto de produzir enunciados. A partir do retorno do adulto para as expressões produzidas por ele, o aprendiz faz ajustes na sua gramática até atingir a forma correta. O CHILD tenta dar conta de seis fatos da aquisição pela criança, a saber: (i) que a compreensão precede a produção, (ii) que a taxa de crescimento do vocabulário cresce inicialmente e depois diminui, (iii) que o tamanho dos enunciados aumenta, (iv) que palavras irregulares são regularizadas, (v) que enunciados semanticamente improváveis são mal-compreendidos, e (vi) que passivas reversíveis são mal-compreendidas.

Para isso, o CHILD se configura também como uma teoria tanto da aprendizagem quanto do conhecimento gramatical da criança, conhecimento este que – em relação à sintaxe – se distingue fortemente das teorias propostas pela gramática gerativa e assumidas

nesta pesquisa, pois não assume nem categorias, nem regras para construção da estrutura hierárquica. No CHILD, a sintaxe se restringe à especificação de relações de ordem linear envolvendo os elementos correlacionados a papéis temáticos (5a), semelhante ao AMBER, portanto. Parte dos enunciados são acompanhados da representação semântica (a entrada "visual") que especifica predicados, papéis temáticos (no caso de verbos) e modificadores envolvidos, como em (5b-d).

(5) a. Sintaxe: (ACTOR): precedes "give," (OBJECT), (TO VAL)
b. **give**, (ATRANS ACTOR (NIL) OBJECT (NIL) TO (POSS VAL (NIL)))
c. **ball**, (BALL1 REF (NIL))
d. **the**, (DEF)

Quando o enunciado não vem acompanhado da representação semântica, o aprendiz conta com um procedimento para inferir o sentido, que escolhe o melhor candidato (em caso de haver ambiguidade) e o submete ao "usuário" (do programa) para que dê um retorno sobre a escolha. Se negativa, o aprendiz repete o processo e escolhe o "próximo melhor" candidato. A aquisição lexical no CHILD é trans-situacional, isto é, o sentido de uma palavra é obtido através da comparação dos contextos em que ela ocorre, de modo a obter a parte do sentido que lhe correspondente unicamente (sem tratamento de homonímia ou polissemia). A aquisição sintática ocorre na medida em que mais de uma palavra é reconhecida e relações de ordem podem ser estabelecidas entre elas. Durante a aquisição, o aprendiz atravessa oito estágios.

No estágio um, sem nenhum conhecimento de palavras e sintaxe, o aprendiz apenas ouve os enunciados e responde (ao adulto) com "hum". No estágio dois, ele começa a convergir para os sentidos de algumas palavras e a produzir enunciados de uma palavra. No estágio três, começa a aprender a ordem básica entre sujeito, objeto e verbo. No estágio quatro, palavras começam a ser adquiridas numa única exposição, enunciados começam a apresentar de 3 a 5 palavras e a interpretação dos papéis dos elementos do enunciado é baseada em

preferências semânticas. No estágio cinco, o aprendiz começa a regularizar verbos irregulares, processo que termina no estágio seis. No estágio sete, sentenças ativas já são compreendidas com base no conhecimento sintático, sendo as passivas reversíveis ainda mal-compreendidas. No estágio oito, enfim, passivas começam a ser interpretadas corretamente com base no conhecimento sintático.

Não há informações quantitativas sobre a modelagem, tais como a extensão do corpus ou o ritmo do aprendizado. Em termos gerais, o modelo parece mimificar relativamente bem certos aspectos da aquisição. Porém, assim como no AMBER, o CHILD carece de uma representação mais adequada do conhecimento sintático, que aqui também é visto basicamente como ordenação linear de palavras e sequências de palavras (sintagmas). Como dito acima com respeito ao AMBER, isto impede que o aprendiz seja capaz de refletir a criatividade de um falante típico.

No que tange à interpretação de ativas e passivas, o modelo me pareceu bastante problemático, pois depende de duas propriedades bastante artificiais: um pré-conhecimento do aprendiz sobre preferências semânticas (p.e., para cada verbo, que tipo de elemento costuma ser o agente ou o paciente) e a supressão artificial de sua capacidade de inferência por um certo período de tempo. Portanto, o CHILD é um modelo que mimifica certos estágios da aquisição, porém o faz com base em mecanismos extra-linguísticos específicos e não em função de possíveis propriedades do próprio conhecimento sendo adquirido. Isto o torna menos interessante do ponto de vista teórico.

3.2.2 Modelos mais abrangentes de aquisição sintática

Nesta subseção, apresento mais detalhadamente dois MCAs cujas características serviram de base e motivaram vários aspectos do IASMIM. Estes modelos são considerados aqui como mais abrangentes, visto que buscam modelar a aquisição de uma gama maior de construções, pelo menos em tese. Outra característica importante destes dois modelos é

o fato de assumirem formalismos explícitos para representação do conhecimento sintático e semântico-lexical do aprendiz, o que permite levantar questões mais interessantes do ponto de vista (psico)linguístico.

Berwick (1985)

Arquitetura

O modelo apresentado em Berwick (1985) parte da premissa de que a aprendizagem é disparada sempre que o aprendiz é incapaz de analisar um dado de entrada. Em outras palavras, quando o analisador sintático (*parser*) não consegue evoluir na análise de uma sentença (o que configura um estado de erro), significa que a gramática do aprendiz é insuficiente e precisa ser atualizada através de procedimentos de aquisição. É proposto um modelo com três componentes, que reflete a caracterização formal do processo de aquisição em Chomsky (1965):

1. Um estado inicial de conhecimento.
2. Um estado alvo de conhecimento.
3. Um procedimento de aquisição que dirige o sistema do estado inicial para o estado final. O procedimento se subdivide em:
   (a) Os dados de entrada que o sistema de aprendizagem utiliza.
   (b) O algoritmo de aquisição em si.

Berwick propõe uma versão revisada do analisador sintático proposto em Marcus (1980). O analisador é um PDA construído em torno de duas estruturas de dados principais, motivadas pela meta teórica de análise determinística: a *pilha de nós constituintes* e a *área temporária de três células*. A pilha armazena a estrutura da árvore em construção, isto é, nós (sintagmas) que ainda estão incompletos. A ordem dos nós em construção na

pilha, no caso do inglês, é inversa à ordem hierárquica na árvore. A área temporária, por sua vez, armazena tanto as palavras da sentença (na ordem de apresentação), quanto nós completos retirados da pilha (que por sua vez podem conter uma ou mais palavras).

A tarefa do analisador é mapear cadeias de palavras em árvores sintáticas. No processamento de uma sentença, o analisador percorre a área temporária num passo único da esquerda pra direita (restrição que reflete as exigências do processamento de sentenças humano) e o interpretador executa as regras compatíveis com o estado momentâneo do analisador (determinado pelos elementos na pilha e na área temporária). Eventualmente – para decidir sobre o próximo passo – o analisador pode processar material à sua direita (antecipação), até uma distância finita e pré-fixada.

Berwick mantém a característica geral do analisador de Marcus, isto é, a análise segue um esquema *padrão-ação*, em que padrões são predicados relativos aos símbolos na pilha e na área temporária e as ações são as operações que constroem a árvore em si. A análise é *bottom-up*, ou seja, constrói a árvore a partir de seus níveis mais baixos, com predição *top-down*, na medida em que itens mais à esquerda disparam projeções. A gramática "madura" terá duas partes: regras gramaticais de base ou estrutura sintagmática[2] (a maior parte fixada pela Teoria $\overline{\text{X}}$) e regras gramaticais transformacionais.

A única ação possível para regras de base é ANEXE, que retira o primeiro item da área temporária e anexa ao primeiro nó da pilha, de acordo com as restrições impostas pelo protocolo $\overline{\text{X}}$. Não há necessidade de regras de projeção, já que a criação e a finalização de sintagmas é feita pelo próprio protocolo embutido. Qualquer outra parte da construção do sintagma terá que envolver uma regra separada. A figura 3.1 exemplifica algumas regras de base para análise de sintagmas nominais:

Os símbolos C e CYC dizem respeito, respectivamente, ao primeiro e segundo nós em construção na pilha (contexto esquerdo), interpretados por Berwick como níveis cíclicos, no

[2] *Phrase structure rules*. No decorrer do texto usamos as duas formas como sinônimas.

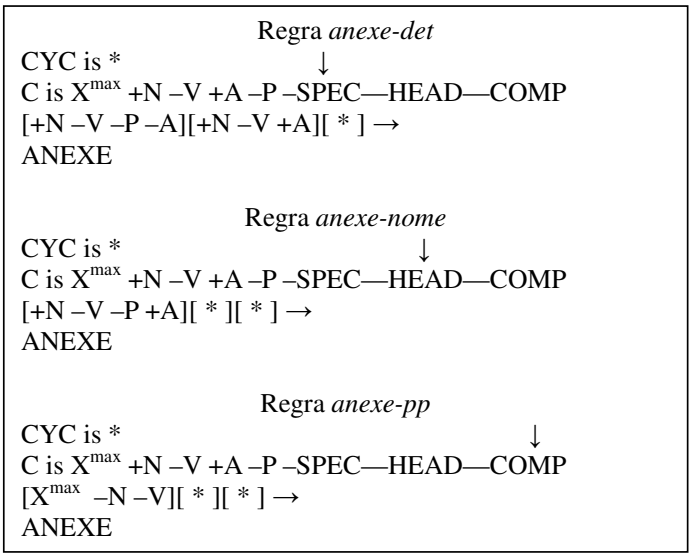


Figura 3.1: Regras para analisar NPs (Berwick, 1985).

sentido da aprendibilidade grau-2 (cf. Culicover & Wexler, 1980). O asterisco (*) indica um "coringa"[3], isto é, combina com qualquer conjunto de atributos. Note na segunda linha o esquema de pacotes baseado na Teoria $\overline{\text{X}}$: um ponteiro controla que posição do sintagma – "spec", "head" ou "comp" – está sendo considerada.

A ordem relativa dessas posições é configurada com base nos dados de entrada. Na terceira linha de cada regra, temos o estado da área temporária: cada célula faz referência ao tipo de elemento (item lexical ou sintagma) e aos atributos relevantes para a regra. Na quarta linha é expressa a ação correspondente. Outras três ações se destinam às regras transformacionais: INVERTA, INSIRA ITEM LEXICAL e INSIRA VESTÍGIO. Regras do tipo INVERTA são consideradas transformações locais simples e capturam fenômenos como a inversão sujeito-auxiliar, no inglês.

As regras de inserção de item lexical e vestígio lidam com construções envolvendo elementos ausentes (no caso, imperativas) ou movidos (no caso, interrogativas). Todas estas regras afetam a estrutura gerada, invertendo posições de elementos ou inserindo outros. Seu formato é idêntico ao das regras de base apresentadas acima. O modelo assume algum conhe-

[3] *Wild card.*

cimento lexical já estabelecido, que propicia ao aprendiz os atributos categoriais das palavras (lexicais e funcionais). Berwick sugere que o aprendiz é capaz de adquirir novas palavras com base em sua distribuição sintática, mas não apresenta os procedimentos respectivos.

Além do enunciado em si, os dados de entrada informam a categoria sintática das palavras, os papéis temáticos dos argumentos e a grade temática dos elementos predicadores. Como assume também uma correspondência entre posições sintáticas e papéis temáticos[4], Berwick embutiu no modelo o pré-conhecimento dessa correspondência e é isto que guiará o aprendiz no momento de decidir onde anexar os itens no sintagma.

Aprendizagem

Para Berwick, o curso da aquisição pode ser visto como o desenvolvimento de uma sequência de analisadores, cada vez mais poderosos, sendo que inicialmente o analisador não contém regra alguma. Há cinco passos principais, no procedimento de aquisição. A figura 3.2 mostra o fluxo básico de funcionamento do analisador e do procedimento de aquisição. O primeiro passo é tentar analisar a sentença com as regras existentes, passando à próxima sentença, caso a análise tenha sucesso (ou seja, quando não há o que aprender).

Em caso de falha, entra-se no procedimento de aquisição. No **passo 2**, o procedimento registra o estado momentâneo do analisador. No **passo 3**, o procedimento checa se a nova sentença demanda a configuração do parâmetro de ordem, a projeção de um núcleo ou a anexação de um item a um sintagma em construção. No último caso, uma nova regra de base é criada. A informação semântica é fundamental neste passo. Se a regra criada tornar possível seguir com a análise, o analisador vai para o **passo 5**, a generalização de regras. Caso contrário, vai para o **passo 4**, em que tenta-se criar uma regra transformacional. Se funcionar, vai para o **passo 5** e tenta a generalização. Após o **passo 5**, volta-se para o **passo 1**.

[4] Por exemplo, no inglês o argumento no papel de *paciente* normalmente ocupa a posição de complemento de VP.

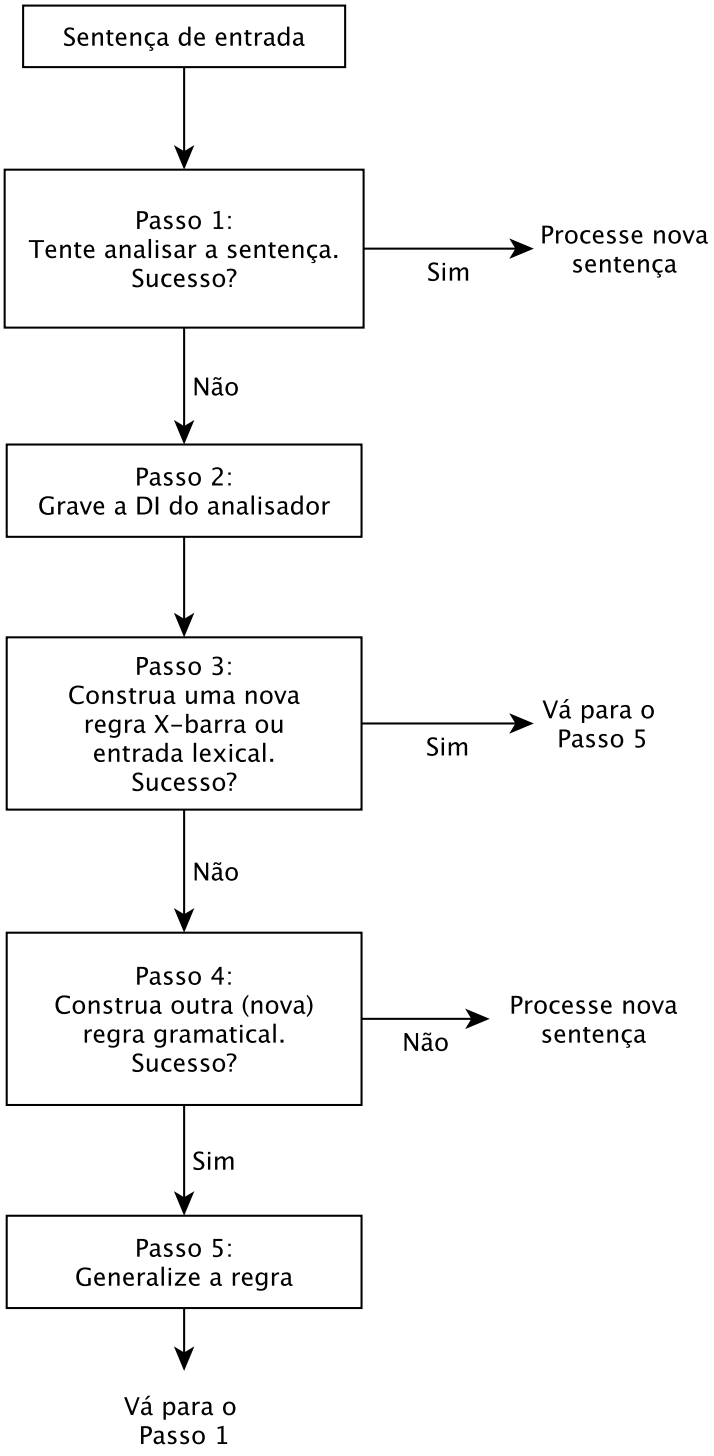


Figura 3.2: Fluxo do processamento e do procedimento de aquisição (Berwick, 1985).

Novas regras são sempre enviadas para a generalização. O primeiro passo, então, é comparar a nova regra com todas as demais de *mesma ação*. Se as primeiras células da área

temporária das regras (nova e existentes) são idênticas entre si, são então verificados seus contextos esquerdo (pilha) e direito (células seguintes à primeira célula). Se nenhum contexto é idêntico, a nova regra é incluída como está. Quando apenas um dos contextos (esquerdo ou direito) é idêntico entre as regras, funde-se o outro por intersecção de atributos e a regra generalizada é incluída na gramática. Quando ambos os contextos são iguais, significa que a regra é redundante e deve ser descartada.

Se as primeiras células da área temporária das regras (nova e existentes) é distinto, há ainda dois casos: quando os contextos direito e esquerdo são iguais, significa que são dois itens lexicais sintaticamente equivalentes e que devem formar uma classe lexical; se, por outro lado, os contextos diferem e os itens fazem parte de uma classe lexical previamente conjecturada, então eles devem ser separados em classes distintas. Este procedimento permite, portanto, algum grau de discriminação entre itens lexicais. Por fim, através da generalização, o modelo é capaz de capturar a recursividade da gramática.

Uma característica bastante importante do modelo de Berwick é a aplicação do Princípio do Subconjunto tanto na execução de regras, quanto na aquisição. Este princípio estabelece que o aprendiz deve conjecturar as hipóteses mais restritas a cada passo, de modo que evidências positivas invalidem as hipóteses, se for o caso. Se o aprendiz fizer o contrário e assumir uma hipótese muito geral, pode nunca encontrar um dado que a invalide, fracassando no aprendizado. O mesmo vale para a execução de regras: diante de um contexto momentâneo da análise, regras mais específicas devem ser testadas antes das mais gerais, sob pena de nunca serem utilizadas, caso contrário.

Com isso, Berwick consegue descartar o esquema de prioridades vinculadas às regras do analisador de Marcus, que tem como objetivo garantir que todas as regras possam ser testadas. No caso da aquisição, para que o aprendiz obedeça ao princípio, novas regras são ordenadas por tipo de ação, na ordem ANEXE, INVERTA, INSIRA ITEM LEXICAL, INSIRA VESTÍGIO. Com isso, operações transformacionais serão conjecturadas apenas na

impossibilidade de aplicação de regras de base.[5]

Finalmente, uma importante restrição do modelo de Berwick é a de que o procedimento de aquisição não pode ser disparado recursivamente, isto é, que durante a aquisição de uma regra, se esta ainda se mostrar insuficiente, o aprendiz dispare os procedimentos de aquisição novamente. Com essa restrição, somada às limitações impostas ao processamento do analisador e às restrições para atender ao Princípio do Subconjunto, Berwick obtém um comportamento incremental do modelo, que aprende gradualmente, impondo uma ordem intrínseca sobre a aquisição, que vai das construções mais simples para as mais complexas.

Escopo

O modelo de Berwick reflete muito claramente os pressupostos gerativistas em relação à FL: há uma GU que determina a natureza do conhecimento linguístico e restringe significativamente a classe de gramáticas aprendíveis. A tabela 3.1 resume as características do modelo quanto ao conhecimento dado inicialmente (a GU) e o conhecimento adquirido pelo aprendiz, ao final do processo de aquisição.

Dado	Adquirido
Laço de execução de regras	Ordem básica livre de contexto
Estruturas de dados	Regras transformacionais
Classificação rudimentar de palavras	Regras de base
Primitivos de atributos	
Restrições $\overline{\text{X}}$	
Estrutura temática simples	

Tabela 3.1: Conhecimento do aprendiz (Berwick, 1985)

Como mostra a tabela, a aquisição lexical não faz parte do escopo de aprendizagem do modelo. Da perspectiva sintática, o modelo aprende a ordem básica da língua, isto é, a ordem relativa entre "spec", "head" e "comp" e as regras gramaticais. A configuração da

[5] Gaylard (1995) afirma, incorretamente em minha opinião, que não há justificativa independente para o ordenamento dos tipos de regras no modelo de Berwick. A justificativa é exatamente o Princípio do Subconjunto.

ordem é absoluta, ou seja, o modelo assume que a ordem básica nas línguas se aplica de modo pervasivo, a despeito da categoria sintática envolvida. Uma vez configurado o parâmetro, todas as projeções $\overline{\text{X}}$ criadas irão instanciá-lo.

O modelo foi testado apenas para o inglês e os tipos de construções sintáticas que o aprendiz pode adquirir incluem NPs (incluindo determinantes, adjetivos e adjuntos), VPs, PPs, o sistema de auxiliares, orações principais e orações subordinadas. O aprendiz é capaz de adquirir sequências de especificadores, cujos elementos não projetam, sendo anexados diretamente na posição de especificador na ordem em que são processados pelo analisador.

Além disso, o sistema deverá processar transformações locais simples, como a inversão sujeito-auxiliar, imperativas e orações interrogativas. O corpus é composto apenas por orações completas, exemplificando todos estes tipos de construção, e as sentenças são apresentadas ao aprendiz em ordem aleatória. Ademais, em relação ao corpus, Berwick assume as seguintes restrições:

- O aprendiz é exposto apenas a sentenças gramaticais (dados positivos).
- Sentenças com no máximo dois níveis de subordinação (aprendibilidade grau-2).
- As palavras são recebidas pré-segmentadas e associadas a atributos.

Discussão

O gramática de uma língua natural em Berwick (1985) é formalmente vista como uma gramática transformacional. Esta assunção somada a outras restrições, tais como o uso de informação semântica, a limitação da antecipação e do número de nós da pilha considerados (aprendibilidade grau-2), garante a eficiência do analisador sintático, o que – segundo a hipótese Berwick – tem como consequência tornar a aprendizagem mais fácil. Berwick propõe que condições para uma análise sintática eficiente também contribuem para uma aprendizagem eficiente. Por ter a componente transformacional, o modelo é capaz de

lidar com alguma variação na ordem, em particular, a inversão sujeito-auxiliar e o movimento-Qu em interrogativas.

Por outro lado, com a opção por uma componente transformacional, Berwick se compromete com um modelo de gramática (Chomsky, 1965) que, por um lado, se mostrou poderoso para capturar regularidades no interior de uma língua, mas, por outro lado, é bastante limitado para capturar regularidades entre línguas distintas, razão pela qual foi sucedido por outros modelos teóricos (Chomsky, 1986, 1993, Gazdar et al., 1985, Steedman, 2000, entre outros). Essa aspecto do formalismo se reflete no caráter idiossincrático das regras transformacionais e de propriedades embutidas no aprendiz, cujas características acabam sendo determinadas por propriedades específicas da língua modelada, no caso, o inglês. Por exemplo, regras do tipo INVERTA são motivadas particularmente pelas construções de inversão sujeito-auxiliar do inglês, tendo pouca ou nenhuma utilidade noutras línguas.

Outro problema é a sobregeração de estrutura sintática do modelo. Isso se deve ao protocolo $\overline{\text{X}}$ embutido que para cada núcleo lexical encontrado projeta um molde $\overline{\text{X}}$ completo – na forma `[XP [Spec] [X' [X] [Comp]]]` – independentemente da presença ou não na sentença dos elementos que preencheriam as posições de especificador e complemento. Em Berwick (1985) isso não parece ser um problema, provavelmente porque itens de certas classes não são considerados como núcleos capazes de projetar estrutura sintática, sendo anexados diretamente à árvore, como é o caso de determinantes. Mazuka (1998) observa, ainda, que o modelo é incapaz de processar línguas como o japonês cuja ordem seja total ou parcialmente inversa à do inglês. Ou seja, seriam necessárias adaptações ao processador no modelo para ampliar sua universalidade.

Porém, se levarmos em conta o desenvolvimento da teoria sintática que sucedeu o modelo, esse problema passa a ser relevante, visto que atualmente entende-se que categorias funcionais são também núcleos sintáticos (ver Pollock, 1989, Cinque, 1999, entre outros). A consequência disso é que o modelo de aquisição produziria estruturas sintáticas contendo

inúmeras posições de especificador não utilizadas, algo que não soa econômico nem eficiente. Vale ressaltar que a própria teria sintática se desenvolveu na direção de assumir menos estrutura, como indica a proposta da "estrutura sintagmática nua" (Chomsky, 1995a), cujo objetivo foi diminuir ao máximo o uso de rótulos e projeções.

Alguns aspectos da informação semântica também devem ser considerados. As relações de predicação no modelo são expressadas basicamente através de papéis temáticos dos argumentos e grades temáticas dos elementos predicadores, informação esta que acompanha o enunciado nos dados de entrada. Embora informação temática pareça ter de fato um papel importante nos estágios iniciais da aquisição (Pinker, 1984), o estatuto teórico dos papéis temáticos em geral não é totalmente claro (Dowty, 1991), havendo discussão sobre a existência de instâncias puras de papéis como "agente", por exemplo, ou se papéis temáticos podem ser melhor descritos como agrupamentos prototípicos de propriedades.

Ademais, em função desse mecanismo e da correspondência assumida pelo aprendiz entre posições estruturais e papéis temáticos, acaba sendo necessário recorrer a pseudo-papéis temáticos para relações não-argumentais, tais como as que envolvem núcleos funcionais e seus complementos, as que envolvem adjuntos, entre outras. O sistema de atributos utilizados pelo aprendiz também é questionável, na medida em que lhe fornece, de saída, especificações para várias categorias sintáticas. Em razão disto, o modelo me parece mais bem compreendido como o de aquisição mais tardia (estágio de três ou mais palavras), altura em que aprendiz já teria fixado as principais categorias lexicais da língua, contando assim com um estoque de atributos semelhantes aos assumidos por Berwick.

Note que os questionamentos feitos até aqui sobre o modelo são de caráter linguístico e formal. Não vejo outro objetivo mais importante para uma modelagem computacional deste tipo do que suscitar questões dessa natureza. Isso evidencia que Berwick (1985) dá importantes passos em direção à plausibilidade psicológica e a um formalismo gramatical mais apropriado, obtendo um modelo relativamente plausível de aquisição de primeira língua.

Apesar dos aspectos considerados problemáticos, o modelo alcançou alguns resultados interessantes:

(i) O percurso e o sucesso da aprendizagem são relativamente independentes da ordem de apresentação dos dados de entrada: a ordem de aquisição é intrínseca, ou seja, é imposta aos dados de entrada em função das características do analisador e do procedimento de aquisição.

(ii) O modelo dá um passo na direção de uma sintaxe relativamente desvinculada da ordem linear, ao localizar essa informação em um nível mais abstrato da gramática (o parâmetro da ordem).

(iii) O modelo permite observar o impacto do Princípio do Subconjunto sobre a aquisição e sobre o uso de regras da gramática.

(iv) A aprendizagem é incremental, prevendo uma ordem intrínseca de aquisição, que começaria por declarativas simples, seguidas de passivas, questões sim ou não, sentenças encaixadas, questões-*qu* e sentenças encaixadas sem o sujeito.

Gaylard (1995)

Arquitetura

Gaylard argumenta em favor de uma abordagem empiricista para a modelagem computacional de aquisição da linguagem, em oposição a abordagens que ela define como inatistas, tais como a de Berwick (1985). A autora reconhece as virtudes do modelo de Berwick, mas critica o mecanismo de aquisição que considera demasiado poderoso, em função da assunção da Teoria $\overline{\text{X}}$ (que fornece de saída o molde geral de qualquer sintagma adquirido) e das ações das regras do analisador, em particular, as regras transformacionais.

Devido a estas assunções, argumenta Gaylard, a aquisição no modelo de Berwick precisa ser restringida por meios que ela considera arbitrários e injustificáveis, a saber, a de

ordenar as tentativas de regras por tipo de ação e a de não permitir o acesso recursivo aos procedimentos de aquisição. Um outro problema geral de abordagens inatistas, segundo a autora, é o de que qualquer restrição inata que se proponha deveria ser verdadeiramente universal (entre as línguas), sob pena de sacrificar a meta de universalidade do modelo. Como foi mencionado anteriormente, regras transformacionais são altamente idiossincráticas e, portanto, vão na direção contrária à universalidade.

Assim, Gaylard propõe um modelo alternativo, tendo como meta assumir o mínimo de pré-conhecimento disponível ao aprendiz e, ainda assim, mantê-lo capaz de adquirir a sintaxe, em especial, regras recursivas. A arquitetura geral do modelo proposto é a seguinte:

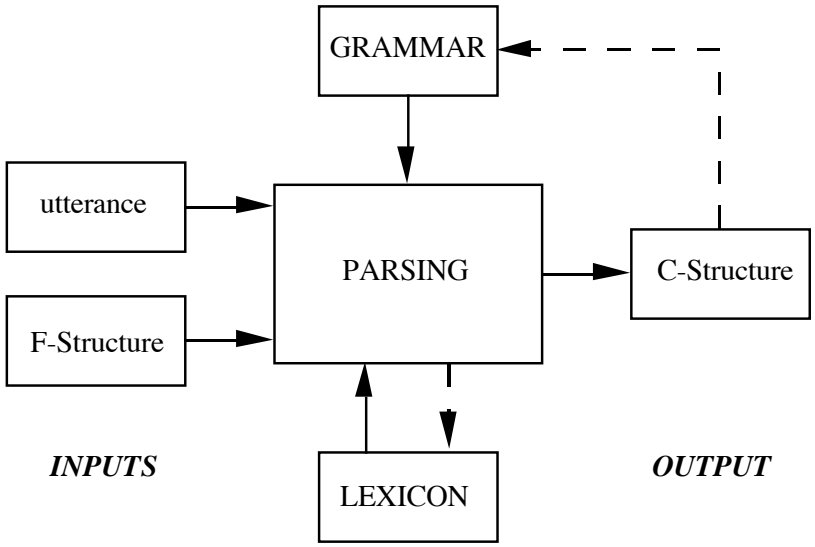


Figura 3.3: Arquitetura geral do modelo em Gaylard (1995).

Como em Berwick (1985), o analisador tem um papel central na arquitetura, utilizando a gramática e o léxico para construir uma árvore sintática para o enunciado. O método de análise implementado no modelo é o de *análise tabular ativa*[6] (ver Jurafsky & Martin, 2008), um método de programação dinâmica. As estruturas de dados são a *Tabela*, que armazena informacões parciais durante a análise, a *Lista de Palavras* a serem consumidas e a *Agenda*,

[6] *Active chart parsing.*

que é uma lista de constituintes a serem processados.

Por ser um método de análise não-determinístico, algumas informações são utilizadas para atender ao critério de determinismo das análises, entre elas, a antecipação do próximo item a processar (quanto à categoria sintática) e informação argumental (para resolver a ambiguidade entre anexar como argumento ou adjunto). O modelo pode ainda recorrer às condições de *Minimal Atachment* (anexar ao constituinte mais próximo) e *Right Association* (sintagmas mais longos são preferidos), estratégias consideradas plausíveis no processamento sintático humano.

Por implementar um outro tipo de analisador sintático, o modelo difere do de Berwick em relação às ações possíveis, que aqui são apenas duas: *anexação* e *invocação* (de regras). Por serem consideradas simultaneamente e dado que a informação semântica de entrada é que determina a ação apropriada, Gaylard afirma eliminar a necessidade da restrição sobre a aquisição recursiva, visto que a invocação de uma nova regra só seria licenciada na impossibilidade de anexação.

Apesar de apresentar também uma primeira versão do modelo que assume dados de entrada pré-segmentados e um léxico já adquirido, o foco de Gaylard está no modelo integrado, em que o aprendiz, além de adquirir a sintaxe, deve também adquirir o léxico, na medida em que aprende a segmentar os enunciados. Portanto, o modelo inclui procedimentos de reconhecimento e aquisição lexical, bem como procedimentos de aquisição sintática. A autora argumenta que certos aspectos da aquisição sintática, como a aquisição de itens funcionais da língua, só é possível em função da interação entre esses procedimentos.

O formalismo gramatical adotado para representação do conhecimento adquirido é o da Gramática Léxico-Funcional. Em lugar das condições de completude e coerência[7] da LFG foi proposta uma notação de regras aumentada responsável por ajustar itens lexicais com base

[7] Uma estrutura-F é completa se ela contém todas as funções gramaticais regidas pelo seu predicado e é coerente se apenas as funções gramaticais regíveis pelo predicado estão presentes.

nas informações de subcategorização e de categoria sintática. Há dois níveis de representação na LFG, a estrutura constituinte (estrutura-C) e a estrutura funcional (estrutura-F), exemplificadas nas figuras 3.4 e 3.5.

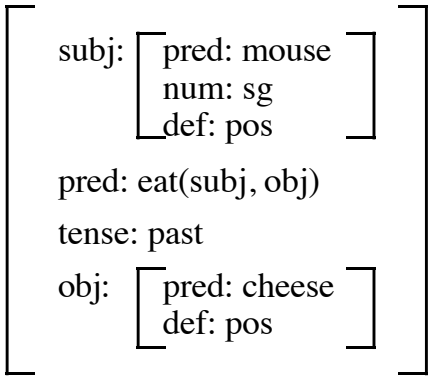


Figura 3.4: Exemplo de estrutura-F para "*the mouse ate the cheese*".

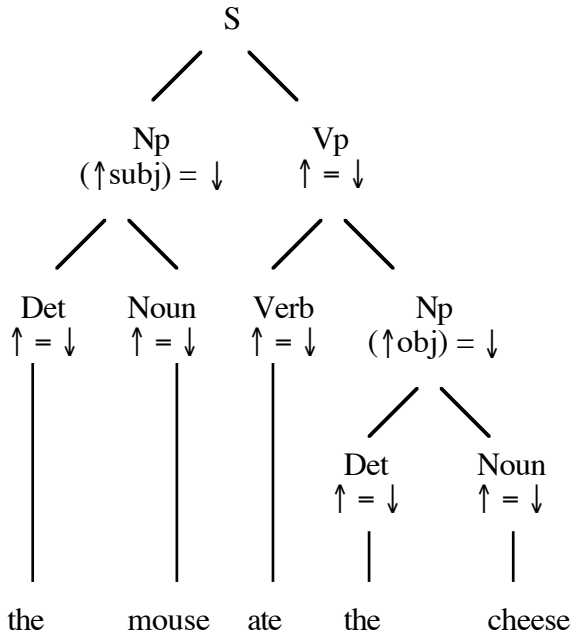


Figura 3.5: Exemplo de estrutura-C para "*the mouse ate the cheese*".

O formalismo da LFG, especialmente em comparação ao assumido em Berwick, se distingue por não ser derivacional, isto é, não assumir qualquer tipo de transformação envolvendo a sentença. Alternativamente, através do relacionamento entre a estrutura-C e

a estrutura-F, a LFG é capaz de representar qualquer mapeamento envolvendo estrutura argumental, funções sintáticas e estrutura sintagmática (ver Bresnan, 2001). Assim, nessa perspectiva, todos os elementos da sentença são analisados como se estivessem *in situ*.

Aprendizagem

Segundo Gaylard, seu modelo adquire uma gramática de estrutura sintagmática (doravante, PSG[8]) recursiva, sem a necessidade de assumir que tal conhecimento é inato. Inicialmente, o modelo adquire uma gramática de estado-finito[9] e um léxico de porções não-analisadas de linguagem, tais como sintagmas e sentenças. Na medida em que as unidades do léxico forem sendo segmentadas em suas partes constituintes, uma PSG passa a ser adquirida, substituindo a gramática de estado-finito inicial. Categorias sintáticas são induzidas através da generalização das regras adquiridas e esse mecanismo é que induz a recursividade na gramática, o que depende, vale ressaltar, de contextos sintáticos envolvendo dois níveis de encaixamento, da mesma forma que em Berwick (1985).

O primeiro passo na análise de um novo enunciado é o reconhecimento lexical. A estratégia assumida por Gaylard é *top-down*, visto que a aquisicão de novos itens é guiada pelo conhecimento de itens lexicais existentes, definidos como "unidades funcionais de sentido". Em outras palavras, a concepção de item lexical no modelo é flexibilizada para incluir porções não-analisadas de linguagem, tais como "ogato" ou "acartachegou". Não apenas isso, mas para que a estratégia funcione é preciso assumir que o aprendiz está exposto não apenas a sentenças completas, mas também a sintagmas e até palavras isoladas, tais como "o gato", "meu livro", "cachorro" etc.[10]

O processamento de um enunciado, para fins de reconhecimento, é feito da esquerda

[8] *Phrase-structure grammar*.
[9] Basicamente restrita a relações de contiguidade linear entre os itens lexicais, sem acesso às relações hierárquicas. Isso se deve aos itens lexicais não-analisados, que tornam a estrutura hierárquica opaca ao analisador.
[10] Assunção plausível, como mostram as tabelas exibidas no Capítulo 5, relativas as tipos de construção encontrados na fala dirigida à criança.

para a direita[11]. À medida que é percorrido, o procedimento de reconhecimento verifica se a porção do enunciado momentaneamente considerada diz respeito à um ou mais itens lexicais já adquiridos pelo aprendiz. Quando apenas um item é reconhecido, a segmentação ocorre, retirando o item reconhecido e reiniciando o reconhecimento para o restante do enunciado. Se a porção em análise for compatível com mais de um item, uma antecipação de até dois itens consecutivos é utilizada para desfazer a ambiguidade. Uma falha ocorre quando nenhum item lexical é reconhecido (seja para o enunciado completo ou para o que resta a ser processado).

Nesse momento, a aquisição lexical é disparada e um item lexical relativo à porção correspondente é conjecturado. Quando todo o enunciado é adquirido como um item lexical, a estrutura-F completa é também vinculada a ele. No entanto, quando uma parte do enunciado tiver sido reconhecida, a aquisição envolve deduzir qual é a porção da estrutura-F do enunciado relativa à porção sendo adquirida. Dada a natureza das gramáticas de unificação (como a LFG), Gaylard propõe um mecanismo de "unificação reversa" que permite encontrar atributos comuns entre duas estruturas (Gaylard, 1995, p.95-98), como exemplifica a figura 3.6.

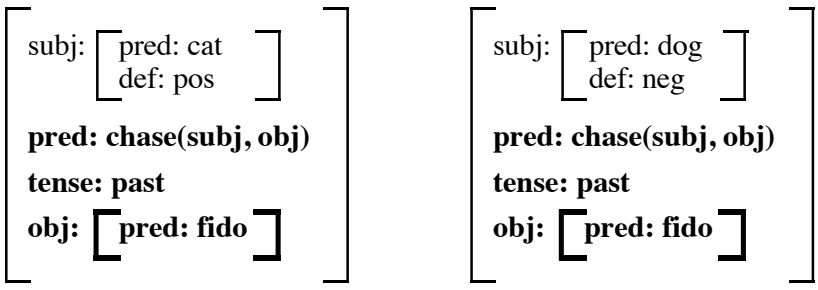


Figura 3.6: Unificação reversa para obtenção de atributos comuns e aquisição de "*chasedfido*".

Eventualmente, a dedução de atributos comuns pode resultar em estruturas contendo mais atributos do que o correto. Neste caso, considera-se que o acúmulo de mais experiência irá permitir ao aprendiz chegar ao conjunto exato para o item lexical em questão. Nesse

[11] Não estritamente, segundo Gaylard (1995, p.95).

sentido, Gaylard propõe um mecanismo nomeado como "consulta lexical flexível"[12], que permite refinar a estrutura-F de entradas lexicais existentes, durante o reconhecimento lexical, a partir da detecção de uma incompatibilidade entre sua estrutura-F e a estrutura-F do dado de entrada.

Uma vez reconhecidos os itens do enunciado, estes são enviados para o analisador. Gaylard assume duas operações básicas de análise que visam estender constituintes existentes ou propor novos. Assim, considerando-se tanto a fase de análise quanto a de aquisição, há quatro ações que o modelo pode entreter, ao processar um dado de entrada: (i) anexação a partir de regra existente; (ii) proposição de novo constituinte a partir de regra existente; (iii) anexação guiada pela aquisição; e (iv) proposição de novo constituinte guiada pela aquisição. As operações estão apresentadas na ordem em que se espera que elas sejam aplicadas: a falha de (i) dispara (ii), cuja falha, por sua vez, dispara a aquisição – itens (iii) e (iv).

Segundo Gaylard, o modelo requer dados de entrada com dois níveis de encaixamento (aprendizagem grau-2) para adquirir estruturas recursivas, resultado similar ao de Berwick (1985). O sucesso da aquisição sintática envolve o mapeamento da estrutura-F numa estrutura-C compatível, a partir da ordem linear do enunciado. Ou seja, Gaylard assume que a representação semântica pode ser vista como uma versão não-ordenada da estrutura-C. Portanto, os efeitos das assunções de Gaylard são bastante similares aos da "Condição de Deformação de Grafo" (Anderson, 1978) o que, em consequência, limita o modelo em termos de variações de ordem (constituintes descontínuos não poderiam ser processados).

Ao processar um item lexical, a primeira ação do analisador sintático é criar um rótulo sintático para o item, como mostra (6). Rótulos[13] são a base para a indução de categorias lexicais. Assim, enquanto o aprendiz não for capaz de segmentar o enunciado, o processamento sintático irá apenas aumentar seu estoque de rótulos sintáticos, visto que um item

[12] *Flexible lexical lookup.*

[13] Note que Gaylard não assume categorias sintáticas pré-definidas, daí o uso de "Node" + numeração incremental para dar nomes aos rótulos. As categorias sintáticas se formam, assim, a partir dos atributos semânticos dos itens que desempenham uma mesma função sintática.

isolado não pode ser combinado. Quando o número de itens lexicais reconhecidos é maior que um, regras de constituência, como mostra (7), começam a ser propostas, permitindo a indução de categorias sintagmáticas ("Node15").

(6)
Node1	→	dog	[pred:dog]
Node2	→	thecat	[pred:cat, def:pos]

(7)
Node1	→	dog	[pred:dog]
Node2	→	thecat	[pred:cat, def:pos]
Node3	→	the	[def:pos]
Node4	→	cat	[pred:cat]
...			
Node15	→	Node3	Node4

Na medida em que as regras vão sendo adquiridas, um procedimento de generalização é responsável por compará-las, criando novas regras, mais gerais, para aquelas que apresentam alguma intersecção no contexto direito. Por exemplo, suponha que o aprendiz tenha adquirido regras para processar "the cat" e "the dog", como mostra (8a). Note que os lados direitos das regras indicam que "Node1" e "Node4" aparecem num mesmo contexto sintático ("Node3"). Neste caso, como mostra (8b), um novo rótulo ("Node12") é proposto para representar "Node1" e "Node4" e uma nova regra mais geral ("Node13") é criada com base neste rótulo.

(8) a.
Node1	→	dog	[pred:dog]
Node4	→	cat	[pred:cat]
...			
Node10	→	Node3	Node1
Node11	→	Node3	Node4

b. Node12 → Node1
→ Node4
Node13 → Node3 Node12

Gaylard descreve o desenvolvimento no modelo como sendo razoavelmente consistente com as propriedades observadas no desenvolvimento da linguagem na criança, prevendo uma progressão que vai da reprodução mecânica de enunciados para a aquisição de construções gramaticais cada vez mais complexas, num percurso suave e contínuo. A aquisição do léxico e da gramática-alvo se dão necessariamente sobre representações inicialmente mais simples do conhecimento. As unidades iniciais não são unidades linguísticas tradicionais, porém seriam consistentes (cf. Peters, 1983, 1985, *apud* Gaylard, 1995, p.6) com os achados sobre sub-segmentação na linguagem da criança.

Finalmente, Gaylard ressalta que há um certo espaço de não-determinismo no modelo, visto que conjuntos de dados de entrada distintos (por ordem de apresentação ou por variedade de construções) vão implicar percursos relativamente distintos, embora o aprendiz vá gradualmente convergir para um sistema mais uniforme de representação. Aparentemente há alguns casos específicos em que a ordem de apresentação dos dados de entrada parece inibir a aquisição de certas estruturas (Gaylard, 1995, p.121).

Escopo

O objetivo principal do modelo de Gaylard é demonstrar a aprendibilidade de estrutura sintagmática, incluindo estruturas recursivas. Um objetivo secundário é demosntrar a interação entre os processos de aquisição lexical e sintática, especialmente para a aquisição de itens funcionais. Tal conhecimento seria a base mínima sobre a qual a criança poderia adquirir as várias estruturas da gramática. Assim, a autora identifica o conhecimento adquirido pelo aprendiz em seu modelo com o conhecimento descrito pelos Estágios I e II propostos em Brown (1973). Gaylard sugere ainda que a aquisição de regras recursivas seria pré-requisito para que o aprendiz ultrapasse estes estágios.

Resumidamente, o Estágio I diz respeito ao período que se inicia com as primeiras combinações de palavras produtivas pela criança. É nesse estágio que ela aprende as relações semânticas básicas (agente, paciente, predicado, dativo e locativo), relacionando-as (no caso do inglês) com a ordem canônica da língua. Outra característica desse período é a ausência geral de morfemas gramaticais e a omissão de constituintes normalmente obrigatórios.

No Estágio II a criança adquirindo o inglês começa a adquirir morfemas gramaticais, tais como as flexões nominais e verbais, preposições, artigos e cópulas. Há uma ordem significativamente consistente na aquisição destes itens, caracterizado como a "curva em U" (Marcus et al., 1992): crianças começam usando morfemas funcionais corretamente (ao que tudo indica, um uso não-analisado), para então passar por uma fase de sobregeneralização e sobrerregularização das regras envolvendo alguns morfemas funcionais e, por fim, voltar a usá-los corretamente, indicando um sistema de regras maduro.

O escopo da aprendizagem é mais restrito que a do modelo de Berwick, embora sejam adquiridas regras para NPs, VPs, PPs, orações principais e orações subordinadas. Para o sintagma nominal, por exemplo, não há indicações de que o modelo consiga lidar com outros elementos, tais como adjetivos e itens pré-nominais para além do determinante. Também não há indicações de como seriam tratados os auxiliares do inglês e nem as relações de concordância. Ademais, apenas sentenças declarativas canônicas são consideradas, ficando de fora construções imperativas, interrogativas, etc.

Discussão

Em alguns aspectos, o modelo de Gaylard (1995) parece ter dado um passo à frente, em relação ao de Berwick (1985), em especial quanto ao descarte do protocolo X-barra. No entanto, apesar de se auto-definir como "empiricista", Gaylard assume diversas propriedades de domínio específico à linguagem, tais como os atributos semânticos que constituem a estrutura-F dos dados de entrada, informações de subcategorização e função sintática, além da própria natureza do conhecimento gramatical, assumido como uma PSG (em oposição, por

exemplo, a modelos não-simbólicos da gramática também (auto-)denominados empiricistas).

O modelo de Gaylard tem o mérito de adquirir aspectos importantes do conhecimento gramatical, tais como categorias sintáticas, classes de equivalência lexical e regras recursivas, assumindo menos restrições inatas (protocolo $\overline{\text{X}}$, categorias sintáticas e atributos categoriais). Um ponto fraco do modelo, por outro lado, é lidar com uma variedade menor de construções, pois seria interessante se a autora pudesse oferecer alternativas concretas para as construções que demandam regras transformacionais em Berwick (1985).

A chave para o sucesso do modelo de Gaylard em adquirir regras sintagmáticas e recursivas está no uso de rótulos sintáticos que encarnam, de um modo mais econômico que a Teoria $\overline{\text{X}}$, o princípio de projeção. Vale ressaltar que a Teoria $\overline{\text{X}}$ é uma generalização sobre as estruturas sintáticas, de modo a capturar todos os tipos de sintagmas possíveis. Porém, não necessariamente a fôrma X-barra é instanciada para toda e qualquer categoria, daí a virtude do modelo de Gaylard, ao propor apenas um mecanismo básico de rotulação.

Um resultado limitado, mas bastante interessante do modelo, é a aquisição de itens funcionais. Ao conceber um modelo integrado, Gaylard sugere uma solução para a aquisição de itens funcionais, que seriam mais difíceis de serem adquiridos pelo aprendiz por não aparecerem de modo isolado ou saliente nos enunciados. Sua aquisição no modelo é possível a partir da comparação de itens não analisados como “adog” e “amouse”, de onde é possível extrair o item comum “a”. A solução parece promissora, porém faltou à Gaylard trazer dados de entrada mais complexos e realistas, isto é, contendo um maior conjunto de atributos, de modo a permitir uma avaliação mais adequada do modelo.

Tomemos o sintagma nominal, por exemplo. Os dados discutidos por Gaylard se encaixam perfeitamente à solução proposta, pois diante de “adog” com uma estrutura-F `[def:neg, pred:dog]` e “amouse” com estrutura-F `[def:neg, pred:cat]`, a comparação é trivial, já que há apenas um atributo distinto em ambas. Diferentemente de Berwick, Gaylard não assume atributos categoriais sintáticos, portanto seria necessário propor um sistema

de atributos semânticos mais refinado, que permitisse a indução das categorias sintáticas representadas no modelo de Berwick através de um sistema de atributos categoriais.

Por exemplo, tal sistema de atributos teria que considerar outros atributos definidores dos predicados, tais como *animacidade*, distinção massa/contável, para substantivos, e atributos distintivos para os diferentes tipos de eventualidades envolvidas nos predicados verbais (Filip, 1999). Sem falar nos atributos normalmente envolvidos em relação de concordância. É muito provável que um sistema mais refinado certamente imporia ao aprendiz um problema menos trivial que o exemplificado por Gaylard. Enfim, não é possível ter uma percepção clara dos limites do modelo, visto que os exemplos apresentados não impõem obstáculos significativos aos procedimentos de aquisição propostos.

Em relação ao método de análise empregado no analisador, dada a importância deste para o modelo, Gaylard reconhece (em retrospecto) que a opção pelo método de análise tabular ativa não foi uma escolha ótima, visto que a imposição de limitações de memória e outras restrições no poder de processamento é mais facilmente implementada em estruturas de dados como a pilha no modelo de Berwick. Portanto, nesse aspecto Gaylard reconhece que seu modelo é menos plausível que o de Berwick.

Gaylard critica a ordenação das ações de regras no modelo de Berwick, porém uma ordenação também é assumida em seu modelo, na medida em que este tem preferência pela anexação sobre a proposição de constituintes. No entanto, não há menção explícita ao princípio por trás dessa ordenação, como há em Berwick, quando este associa a ordem escolhida ao Princípio do Subconjunto. É possível e até provável que a ordem das ações no modelo de Gaylard derive do mesmo princípio, visto que a anexação é uma opção mais econômica (quanto à criação de estrutura) do que a de proposição de novos constituintes.

Algumas características da PSG do modelo são a não-binariedade estrita das árvores e a assunção de que a ordem linear tem papel nas regras sintáticas. Tais características, como já foi discutido em relação ao modelo de Berwick (1985), distanciam o modelo de propos-

tas mais recentes envolvendo a sintaxe das línguas naturais. Ademais, assumir uma sintaxe menos presa à ordem linear parece ser um requisito necessário (mas talvez não suficiente) para que o aprendiz seja capaz de lidar com constituintes descontínuos. O modelo de Gaylard muito provavelmente seria incapaz de lidar com ordens conflitantes e com constituintes descontínuos.

Em relação ao de Berwick, o modelo de Gaylard avança ao mostrar que aquilo que o modelo de Berwick adquire com base em assunções inatistas mais fortes pode ser aprendido assumindo-se restrições mais gerais, possivelmente universais, sobre a natureza do conhecimento linguístico. Faltou à Gaylard indicar um caminho alternativo para o tratamento das construções que, no modelo de Berwick, demandam regras transformacionais. A alternativa parece existir, dado que a LFG oferece maquinaria para a representação de tais construções. A questão – não trivial, diga-se – é mostrar como tais representações poderiam ser induzidas sem recorrer a mecanismos demasiado *ad hoc*.

Um avanço particularmente importante em Gaylard (1995) é a integração dos mecanismos de segmentação, reconhecimento e aquisição lexical ao modelo de aprendizagem. Com isso, além de obter um comportamento mais suave e gradual no desenvolvimento da gramática do aprendiz, Gaylard consegue oferecer uma alternativa interessante para o tratamento de itens funcionais, que poderiam ser adquiridos mesmo não estando salientes nos enunciados. O problema, em relação a este aspecto do modelo, é que a autora não apresenta resultados quantificados sobre o grau de sucesso da segmentação e da aquisição lexical, quanto ao que seria esperado e o que o modelo atingiu. Portanto, não dá para saber até que ponto a proposta de fato funciona.

3.3 Outras modelagens

3.3.1 Aquisição lexical

Siskind (1996) apresenta um estudo computacional da tarefa de aquisição lexical, isto

é, a tarefa de fazer o mapeamento palavra/significado. O autor apresenta um algoritmo para resolver este problema, no qual é implementada uma interpretação precisa da aprendizagem trans-situacional e o princípio de contraste aplicado sobre palavras de um enunciado. O algoritmo consiste de uma série de heurísticas concebidas para que a aquisição tenha sucesso e seja eficiente, mesmo sob variadas condições, tais como ruídos (enunciados eventualmente pareados com sentidos incorretos), incerteza referencial (enunciados pareados com mais de um sentido, sendo alguns apenas parcialmente corretos) e homonímia e/ou polissemia.

Os enunciados são apresentados como listas não-ordenadas de palavras (portanto, assume pré-segmentação) e a representação do sentido é baseada na representação semântico-conceitual proposta em Jackendoff (1983). Para as simulações do modelo, um corpus foi gerado automaticamente (e aleatoriamente), a partir de uma pequena gramática livre de contexto, de modo que os enunciados eram compostos de sintagmas nominais ou verbais, sendo os sintagmas nominais compostos de um substantivo e uma palavra funcional (opcional) e os verbais de um verbo, um modificador opcional e a sequência de sintagmas nominais relativa aos argumentos do verbo. Portanto, um corpus composto apenas de sentenças declarativas simples, sem adjetivos ou adjuntos.

Palavras funcionais (tais como determinantes) não apresentam conteúdo semântico no modelo e, portanto, não contribuíam símbolos para a interpretação dos enunciados. Enunciados gerados compostos de 1 palavra, mais de 30 palavras ou envolvendo mais que 30 símbolos conceituais foram descartados. A MLU das sentenças variou de 4.99 a 6.29, entre as simulações. Os parâmetros controlados nas simulações foram (i) o tamanho do vocabulário (1000 a 10000 palavras), (ii) o grau de incerteza referencial (10 a 100 sentidos eventualmente pareados com um enunciado), (iii) a taxa de ruído (0 a 20% de enunciados com sentidos incorretos), (iv) o tamanho do inventário de símbolos conceituais (250 a 2000) e (v) a taxa (média) de homonímia do corpus (de 1 a 2).

Os resultados mostraram que, para o parâmetro (i), um corpus de 10 mil palavras são

suficientes para um vocabulário de 1000 palavras e até 150 mil palavras são necessárias para atingir 95% de convergência, dado um vocabulário de 10 mil palavras. Para o parâmetro (ii), o tamanho do corpus não afeta significativamente a convergência, bastando em torno de 12 mil palavras, para qualquer valor do parâmetro. Para (iii), 10 mil palavras são suficientes para um corpus sem ruído e até 80 mil palavras são necessárias para processar uma taxa de 20% de ruído. Para (iv), o tamanho do corpus necessário também não é afetado pelo tamanho do inventário de símbolos conceituais, sendo necessárias em torno de 11 mil palavras para qualquer valor.

Finalmente, para o parâmetro (v) 10 mil palavras são suficientes para um corpus sem homonímia, enquanto uma taxa de 2 demandou um corpus de 900 mil palavras para convergir, o que evidencia o enorme impacto desta variável sobre a aquisição no modelo. A taxa de aquisição de palavras começa lenta para as primeiras 25 palavras em média e então aumenta significativamente até que a maior parte do vocabulário tenha sido adquirido, quando começa então a cair e a se estabilizar em patamares mais baixos. A quantidade de exposições necessárias para adquirir uma palavra também diminui à medida em que o vocabulário cresce, chegando ao ponto de precisar apenas de uma exposição para adquirir novas palavras, em estágios mais tardios.

É interessante notar que o grau de incerteza referencial e o tamanho do inventário de símbolos conceituais parecem não afetar a aquisição. Se o resultado estiver na direção correta, indica que a hipotética infinidade de elementos contextuais presentes numa cena não seria problema para a criança. Da mesma forma, mesmo que o inventário de símbolos conceituais seja potencialmente infinito, isto parece não ter impacto sobre a aquisição. Dois resultados até surpreendentes, mas que fazem todo o sentido, dado o sucesso da criança na aquisição. Por outro lado, os resultados mostram que homonímia, ruído e tamanho do vocabulário afetam a aquisição. A verdade é que todos estes parâmetros afetam a esparsidade do problema, isto é, quanto maior o vocabulário, a taxa de homonímia ou a de ruído, mais enunciados são necessários para capturar a distribuição das palavras.

Apesar dos resultados positivos, Siskind (1996) ressalta algumas limitações do algoritmo. Uma delas é o fato de que o algoritmo assume homonímia estrita, isto é, que algumas palavras podem ter mais que um sentido e que tais sentidos são totalmente distintos. Porém, línguas exibem também polissemia e é possível que em algumas (ou várias) situações, palavras exibam um conjunto de sentidos que compartilham parte dos símbolos conceituais. Tais casos, se tratados como homonímia, elevariam demasiadamente a taxa de homonímia, possivelmente impedindo a convergência do algoritmo.

Outra limitação diz respeito à representação semântico-conceitual, que é simplificada no modelo, visto que não assume propriedades próprias às palavras funcionais e trata a composição semântica apenas como substituição de argumentos. O autor ressalta ainda que o algoritmo não é capaz de lidar com enunciados em que certos elementos estão omitidos, tais como construções com elipses ou argumentos nulos. No modelo, fragmentos de enunciados devem estar emparelhados com fragmentos da representação semântica. Por fim, o fato de que a simulação contempla apenas sentenças declarativas simples é também uma limitação importante, ainda mais quando categorias funcionais e lexicais são desconsideradas, tais como flexão verbal, adjetivos, advérbios e adjuntos preposicionados.

Apesar destas limitações, o algoritmo proposto em Siskind (1996) apresenta propriedades bastante interessantes do ponto de vista linguístico e aquisicional. Primeiro, por implementar a aprendizagem trans-situacional, considerada uma estratégia plausível de aprendizagem (Pinker, 1989, Fisher et al., 1994). Em segundo lugar, por refletir certos aspectos da aquisição lexical, tais como a taxa de aquisição de novas palavras, baseando-se em heurísticas simples e gerais, sem a necessidade de mecanismos específicos e extra-linguísticos. Finalmente, o modelo tem sucesso no que se propõe, parece ser relativamente independente da língua-alvo e é relativamente robusto a circunstâncias mais ou menos difíceis de aprendizagem. Por estas razões, este algoritmo foi adaptado e acoplado ao IASMIM e seus detalhes de funcionamento são apresentados no próximo capítulo.

3.3.2 O modelo paramétrico em Villavicencio (2002)

Neste estudo, Villavicencio (2002) apresenta um MCA composto por uma GU associada a parâmetros e um algoritmo de aprendizagem, de acordo com a Teoria de Princípios de Parâmetros. O estudo se concentra na aquisição de grades de subcategorização e da ordem de palavras. A abordagem é probabilística e o aprendiz é concebido para ser robusto à ruídos e ambiguidade nos dados de entrada, composto no modelo por sentenças de um corpus de fala espontânea dirigida à criança, cujos enunciados foram anotados com as respectivas formas lógicas. A GU é implementada como uma Gramática Categorial Generalizada Baseada em Unificação, um formalismo lexicalizado, de modo que a gramática (categorias e regras) está codificada diretamente no léxico, embutida numa rede de herança de padrões.

Em tais redes, generalizações abrangentes da gramática são capturadas de modo econômico, através da especificação de tipos gerais, cujas características são herdadas por seus sub-tipos que, por sua vez, acrescentam características próprias e podem também apresentar sub-tipos próprios, que herdam suas características e assim sucessivamente, de acordo com a necessidade. A GU pode conter ainda regras lexicais, que permitem derivar itens lexicais a partir de outros (como, por exemplo, deriver verbos flexionados em terceira pessoa do singular, a partir da forma base). Assim, o léxico se constitui de uma parte básica e de outra derivada.

Os parâmetros assumidos no modelo também são codificados como tipos na rede de herança, de modo que possam ter valores finitos especificados como *não-configurados*, *padrão* e *não-padrão* (absolutos). Tais propriedades podem, então, ser herdadas por sub-parâmetros vinculados a um dado parâmetro. Há dois conjuntos de parâmetros no modelo de Villavicencio: *categoriais* e *de ordem de palavras*. Os primeiros, definem as categorias permitidas pela gramática num dado momento, no curso da aquisição. São 89 parâmetros, agrupados de acordo com o tipo sintático das respectivas categorias e ordenados de acordo com sua valência.

Por exemplo, o parâmetro para verbos transitivos é um sub-tipo do parâmetro para verbos intransitivos, visto que no modelo verbos transitivos são definidos com base em verbos intransitivos (i.e., herdam a característica de ter um sujeito). Em relação à ordem, são 18 parâmetros, também implementados como tipos na hierarquia da rede. O parâmetro inicial é o relativo à ordem geral da língua, havendo sub-parâmetros para elementos específicos, tais como o sujeito ou outros argumentos do verbo. Todos os sub-parâmetros herdam o valor dos parâmetros mais altos na hierarquia, mas podem, de acordo com a experiência, alterar o valor.

O modelo inclui aquisição lexical, modelada conforme a adaptação de Waldron (1999, *apud* Villavicencio, 2002) do algoritmo proposto em Siskind (1996). A principal diferença, segundo a autora, é que na adaptação de Waldron os itens funcionais possuem conteúdo semântico, ao contrário de Siskind, que os assume como elementos semanticamente vazios. Em relação ao corpus utilizado, o algoritmo de aquisição lexical conseguiu processar 63,6% dos enunciados (num total de 1517 que compõem o corpus), dos quais 95,23% eram corretos, ou seja, um total de 965 enunciados. Esta saída da aquisição lexical é então enviada para o procedimento responsável por identificar as categorias sintáticas apropriadas a cada palavra.

O procedimento para aquisição sintática, numa execução típica, atingiu 52,6% de enunciados processados (dos 965). Porém, do total de 508 enunciados processados, apenas 4,7% eram atribuições de categorias sintáticas corretas, com o restante contendo pelo menos uma categoria incorreta. Para contornar este problema, o modelo conta com um procedimento para descartar atribuições inválidas, aumentando o número de enunciados aproveitados dentre os 508.[14] Após obter uma atribuição válida, a sintaxe envia as categorias para o módulo de detecção e processamento de gatilhos, responsável por identificar os gatilhos providos pelo dado e configurar os parâmetros respectivos.

A natureza desta etapa da aprendizagem é probabilística, concebida para determinar

[14] O número exato não é informado pela autora e, no fim, não fica claro de fato quantos enunciados foram utilizados na configuração paramétrica.

a configuração paramétrica da gramática mais adequada para descrever os gatilhos contidos no dado de entrada. Villavicencio (2002) conduz então uma série de simulações, controlando algumas variáveis como nível de ruído ou ambiguidade (envolvendo PPs locativos). A autora mostra que o modelo converge para a gramática-alvo nas simulações, mostrando-se, assim, robusto a ruídos e ambiguidade, alterando os parâmetros apenas quando encontra fortes evidências para isso.

Entre os pontos fracos da modelagem, dentro daquilo que idealmente se espera, vale ressaltar a assunção de uma GU específica para o inglês, o que indica que o modelo é restrito em termos translinguísticos, e o baixo número de enunciados de fato utilizados para aquisição paramétrica, o que pode ser indício – dado que esta convergiu mesmo assim – de algum viés próprio aos procedimentos de aquisição ou aos dados de entrada, de modo a propiciar a convergência com base em uma baixa quantidade de dados (especialmente dado o caráter probabilístico do modelo).

Por outro lado, o estudo de Villavicencio possui várias propriedades interessantes. Primeiramente, é uma tentativa concreta de modelar uma visão paramétrica do conhecimento linguístico, ao mesmo tempo em que a autora não idealiza demasiadamente a tarefa, visto que cabe ao aprendiz adquirir o léxico, identificar as categorias sintáticas das palavras para, só então, extrair os gatilhos dos enunciados e configurar os parâmetros. Em segundo lugar, o fato de mesclar uma abordagem probabilística (para configuração paramétrica) com uma relativamente determinística (para aquisição lexical e sintática), é basicamente inédito no que diz respeito a MCAs.

Uma terceira virtude é o uso de um corpus de sentenças dirigidas à criança, o que confere maior plausibilidade ao modelo, embora corpora artificiais possam ser adaptados de modo a refletir certas propriedades distribucionais observadas em corpus de fala dirigida à criança. Outro aspecto interessante desta proposta é que ela implementa (através da noção de hierarquia) a ideia de que certos parâmetros (quando configurados) devem ter impacto

noutros aspectos da gramática e não apenas um efeito isolado. Por fim, a disponibilização de medidas quantitativas dos resultados obtidos é importante, na medida em que permite alguma comparação mais direta com resultados obtidos noutras modelagens.

3.3.3 Modelos para aspectos mais pontuais da gramática

Berwick & Pilato (1987)

Neste trabalho, os autores mostram como o modelo de Angluin (1982) – um algoritmo para inferência indutiva de uma classe particular de autômatos finitos determinísticos – pode ser aplicado com sucesso para domínios restritos da gramática de uma língua, no caso, as relações de ordem envolvidas no subsistema de especificadores e no de auxiliares do inglês. Dois corpus foram construídos, um para cada domínio, contendo sequências de itens tais como "*could be taking*", "*will have been taking*" e "*the very old big deer*". Os corpus não são exaustivos quanto às combinações possíveis, mas cobrem uma parte significativa dos subsistemas em questão. Os autores aplicam o modelo diretamente à cadeia superficial de elementos, visto que estão focados apenas na ordem linear.

A abordagem do modelo é de domínio geral. Portanto, basta que o sistema a ser induzido seja regular, isto é, que seja possível especificar um autômato finito determinístico para ele. Assim, seria possível aplicar a proposta a estruturas sintáticas, desde que as relações em foco exibam características regulares. Os autores concluem que procedimentos gerais de indução, como o que discutem, podem ser parte das estratégias de aquisição da linguagem, senão para aspectos centrais desta, pelo menos para alguns de seus subsistemas. Aliás, os autores afirmam que a inferência indutiva só é computacionalmente tratável se aplicada a domínios restritos do conhecimento linguístico.[15]

[15] Entendo que esta afirmação se relaciona com os resultados de Gold (1967) já citados e que derivam do fato de que os autores partem da premissa de que não há informação semântica disponível, juntamente com as sentenças.

Pearl & Sprouse (2011)

Neste estudo, os autores argumentam em favor de uma aprendizagem com base em informação probabilística distribucional, sem a necessidade de postular um viés específico à linguagem, no caso a GU. Na modelagem em questão, assume-se que o aprendiz já teria conhecimento das categorias sintáticas (tais como IP, CP, VP etc.) e o corpus utilizado foi constituído por sentenças dirigidas à criança, obtido na base CHILDES e anotadas sintaticamente pelos autores e sua equipe. O foco do modelo é a aprendizagem de restrições envolvendo ilhas sintáticas, tais como nos exemplos abaixo:

(9) a. I like the car that you think [that John bought __].

b. *I like the car that you wonder [whether John bought __].

(10) a. I don't know who bought most of these cars, but that car, I think [that John bought __].

b. *I know who bought most of these cars, but that car, I wonder [whether John bought __]?

(11) a. Smart though I think [that John is __], I don't trust him to do simple math.

b. *Smart though I wonder [whether John is __], I trust him to do simple math.

Pearl & Sprouse propõem que o aprendiz pode utilizar *trigramas* para calcular a probabilidade de certas cadeias de dominância intervenientes entre um item-*qu* e sua lacuna correspondente. Por exemplo, a probabilidade de uma cadeia como *qu*-XP-YP-ZP-LP-MP-*lacuna* é calculada como o produto das probabilidades dos trigramas formados a partir da cadeia: p(start-XP-YP) * p(XP-YP-ZP) * p(YP-ZP-LP) * ... * p(LP-MP-end). O interessante dessa proposta é a aplicação dos trigramas não a uma sequência de palavras, como em geral ocorre em etiquetadores morfológicos, mas à sequência linear de nós dominantes que ligam o item-*qu* à lacuna correspondente.

Segundo os autores, os resultados mostram que o aprendiz modelado foi capaz de adquirir o padrão superaditivo de interação[16] observado nos experimentos de julgamento de aceitabilidade feitos com adultos para as construções envolvendo ilhas. Porém, ainda houve diferenças importantes entre os julgamentos de aceitabilidade observados e as preferências de gramaticalidade inferidas pelo aprendiz, diferenças que os autores sugerem advir de outros fatores que afetariam julgamentos de aceitabilidade, tais como questões de acesso lexical, probabilidade semântica e dificuldade de processamento.

Apesar da limitação dos resultados, destaco esta modelagem aqui como exemplo de um uso mais sofisticado – quando comparado, por exemplo, à proposta em Berwick & Pilato (1987) – da informação contida na representação sintática das sentenças, no caso, as cadeias de dominância contidas na árvore. Embora os autores estejam, inclusive, argumetando contra a GU, não se pode deixar de notar que a própria assunção de uma estrutura sintática e categorias sintagmáticas depõe contra este argumento. Mas deixando esta questão de lado, fica aberta a possibilidade de que os procedimentos de aprendizagem e de processamento da linguagem façam um uso mais sofisticado das estruturas disponibilizadas pela gramática.

3.4 Sumário

A análise dos modelos citados permite destacar aspectos das modelagens computacionais de aquisição que serviram de base para a especificação de várias propriedades do IASMIM. Em primeiro lugar, em relação às características do analisador sintático, o modelo de Berwick se destaca em termos do poder de processamento. A pilha de sintagmas em construção, em especial, permite o controle direto sobre o número de nós cíclicos considerados como contexto sintático do elemento em análise e, com isso, testar a hipótese de aprendibilidade grau-2 de Culicover & Wexler (1980).

Quanto ao formalismo gramatical, os dois modelos analisados assumem uma repre-

[16] Ver Sprouse et al. (2012), em que os autores propõem que o efeito de ilha resulta da interação "superaditiva" de dois fatores: (i) a dependência de longa distância, e (ii) a presença de uma ilha sintática.

sentação sintática n-ária, isto é, a ramificação de nós pode ter mais que dois ramos. É possível que isto facilite a vida do analisador na tarefa de construir a árvore, pois implica menos estrutura sendo gerada. Porém, perde-se em aproximação com os modelos de análise mais recentes da teoria sintática (Chomsky, 1993, e posteriores), em que não apenas as árvores são estritamente binárias – seguindo Kayne (1981) – mas itens funcionais (inclusive flexionais) são considerados de um modo geral como núcleos sintáticos (p.e., Cinque, 1999). Seria interessante investigar, portanto, quais implicações sobre a aquisição advém da assunção de estrutura estritamente binária.

Outra questão envolve a assunção de um caráter transformacional para a gramática, aspecto que distingue os modelos de Berwick e de Gaylard. Embora Gaylard (1995) não tenha enfrentado o problema da aquisição das classes de construções linguísticas para as quais a abordagem transformacional foi proposta em Berwick (1985), sua sugestão de abordagem não-transformacional se mostra mais propensa a atingir objetivos de universalidade, ainda mais se levarmos em conta que o formalismo da LFG prevê mecanismos apropriados para tais construções. Seria, portanto, uma questão de superar o ponto atingido por Gaylard, capacitando o modelo para lidar com variações na ordem linear, elementos nulos, etc.

Um terceiro aspecto envolve o sistema de atributos assumido no modelo. Enquanto Berwick (1985) assume atributos categoriais sintáticos, tais como $\pm$N (nominal) e $\pm$V (verbal), Gaylard (1995) assume atributos semânticos a partir dos quais categorias sintáticas seriam induzidas no processo de aquisição sintática. Embora a assunção de Berwick seja razoavelmente plausível para estágios mais tardios da aquisição, a possibilidade de induzir categorias sintáticas a partir de uma representação semântica compatível com os estágios iniciais da aquisição é atrativa, se viável. A modelagem de Gaylard (1995) indica que sim.

Gaylard (1995) mostrou ainda que o protocolo $\overline{\text{X}}$ embutido no modelo de Berwick pode ser substituído por uma versão mais econômica do princípio de projeção que prevê apenas a projeção do rótulo imediato dos itens lexicais, sem no entanto perder (em relação ao modelo

de Berwick) a capacidade de induzir regras sintagmáticas recursivas. A ausência do protocolo tem ainda o benefício de simplificar a anexação de itens à árvore, visto que não seria mais necessário embutir no modelo o tipo de mapeamento entre papéis temáticos (ou funções sintáticas) e as posições de especificador e complemento de uma projeção $\overline{\text{X}}$ completa.

Quanto à integração com a aquisição lexical, a proposta de Siskind (1996) aparece como uma opção interessante para acoplar em um modelo de aquisição sintática, visto que seus resultados parecem estabelecidos e, pelo menos em números, melhores que outros modelos que implementaram algum tipo de aquisição lexical, tais como os de Gaylard (1995) e Villavicencio (2002). Os demais modelos e modelagens considerados neste capítulo apontam para possíveis extensões futuras do IASMIM, por exemplo, quanto a aspectos pontuais do conhecimento gramatical passíveis de serem abordados probabilisticamente (Villavicencio, 2002, Pearl & Sprouse, 2011) ou mesmo através de máquinas de estado-finito, como proposto em Berwick & Pilato (1987).

Finalmente, vale ressaltar que inúmeras modelagens foram deixadas de fora do levantamento apresentado neste capítulo. Seria inviável fazer justiça a todas elas, dado seu grande número. Foi dada preferência aqui às modelagens que foram mais relevantes para a presente pesquisa, especialmente aquelas que serviram de base para o desenvolvimento do IASMIM. Enfim, parte dessa intensa e extensa linha de pesquisa pode ser conferida em Pinker (1984), Seidenberg (1997), Christiansen & Chater (1999), Broeder & Murre (2000), Kaplan et al. (2008), Frank (2011) e Yang (2011), entre outros.

Parte II

IASMIM

4

O modelo de aquisição

4.1 Visão geral

4.1.1 Objetivos da modelagem

A motivação inicial para o desenvolvimento de um modelo computacional de aquisição veio a partir do estudo em Faria (2009), em que foi feita uma implementação experimental do modelo descrito em Berwick (1985). Neste estudo, Faria conclui que – para atingir maior universalidade e abrangência gramatical – seria necessário fazer uma série de alterações no modelo, sugerindo (i) uma abordagem não-transformacional para o conhecimento sintático do aprendiz; (ii) o abandono do arcabouço X-barra, especialmente em relação à posição de especificador; (iii) a exclusão da ordem linear como componente do conhecimento sintático adquirido; e (iv) o abandono do estoque de categorias sintáticas e papéis temáticos dados de saída ao aprendiz. Note a confluência entre estas sugestões e as conclusões do capítulo anterior.

A análise de outras modelagens veio corroborar as conclusões em Faria (op.cit.). O primeiro objetivo desta pesquisa foi, assim, o de ir ao encontro das sugestões (i) a (iv) acima. Ao mesmo tempo, o intuito foi o de atingir isso assumindo um conjunto restrito de mecanismos pressupostos, que possam ser motivados de maneira convincente. Desse

modo, partindo de assunções mais restritas (i.e., dispositivos mais abstratos em relação ao conhecimento alvo) é possível investigar se tais restrições se mostram suficientes para a aquisição ou se é preciso robustecer o aprendiz com pré-disposições mais específicas.

Gaylard (1995) (cf. o capítulo anterior) considera esta uma abordagem "empiricista". Entendo que não seja necessariamente o caso, a não ser que o termo "empiricista" possa ser equacionado em algum sentido com o termo "minimalista". Por exemplo, ao dispensar a Teoria X-barra, Gaylard – afirmando guiar-se por considerações empiricistas – dá um passo em termos práticos equivalente ao de Chomsky (1995a), quando – por considerações minimalistas – propõe dispensar essa teoria enquanto componente da gramática. Excluir a ordem linear enquanto componente do conhecimento sintático do falante também tem sido considerado na teoria (ver Uriagereka, 1999). Portanto, o que distinguiria "empiricista" de "minimalista", neste caso, seria determinar se este "mínimo de pré-conhecimento" é de caráter geral ou específico à FL.

4.1.2 Assunções sobre a Faculdade da Linguagem

O modelo assume uma divisão mais estrita entre processos sintáticos, de um lado, e o que chamo de "processos de codificação", responsáveis pela linearização da estrutura sintática, incluindo aí o ordenamento linear das palavras. Dessa forma, tais processos se dariam na interface PF – que aqui assume um escopo mais amplo que o tradicionalmente conferido a ela – ou após ela. Essa assunção tem como objetivo se aproximar de Chomsky (1968), quando sugere que grande parte do que é aprendido na aquisição se deve à contribuição própria à língua, o que interpreto como sendo basicamente as informações morfológicas (flexões, concordância etc.), lexicais (aquisição de palavras) e as relações de ordem. A aprendizagem sintática, por outro lado, estaria direcionada às relações de constituência, às categorias sintáticas, lacunas, relações de ligação, entre outros.

Palavras ou mesmo morfemas, nessa perspectiva, não seriam exatamente os elementos

primitivos com os quais a sintaxe opera. Em seu lugar, assume-se aqui que a sintaxe opera diretamente – e tão-somente – com conjuntos de atributos[1]. Outro aspecto importante é a ausência de atributos estritamente formais ou 'sintáticos', tais como os atributos de *Caso*, EPP[2] etc. Apenas atributos que possam ser motivados semanticamente são considerados no modelo (o que inclui atributos normalmente considerados em análises sintáticas, tais como número, pessoa e *wh*) e, assim, caberia ao processo de aquisição determinar quais destes atributos entram – e como entram – na determinação de propriedades sintáticas.

Vale ressaltar que a assunção de que a ordem linear seja relevante apenas em PF, não tendo papel relevante na sintaxe, torna possível lidar com as diferentes ordens lineares encontradas nas línguas naturais sem a necessidade de recorrer a operações de movimento.[3] Isto porque a ordem linear poderia ser determinada, por exemplo, por parâmetros de ordem mapeados na interface PF, a partir de estruturas sintáticas subjacentes relativamente fixas, isto é, com variações estruturais mínimas, embora exibindo variações nos valores dos atributos dos elementos envolvidos (este aspecto é ilustrado no decorrer do capítulo). Assim, as distintas ordens lineares derivariam de diferentes configurações paramétricas em PF.

4.1.3 O modelo

O IASMIM é um *modelo minimalista integrado do sistema de aquisição da língua-I*, ou seja, da gramática. Embora inclua um processador, é um modelo voltado especialmente para a aquisição lexical e da *competência* sintática do falante e não para aspectos da performance, isto é, aspectos envolvendo o processamento psicolinguístico de modo mais amplo ou o uso da língua, em termos de preferências lexicais etc. Pode ser considerado *minimalista* (cf. Chomsky, 1995b), na medida em que busca modelar a aquisição a partir de um

[1] Utilizo 'atributo' como alternativa ao termo 'traço', até em função de adotar gramáticas de unificação, em que o termo 'atributo' é mais comum. Ambos seriam equivalentes ao termo em inglês '*feature*'.

[2] O atributo EPP advém do *Princípio de Projeção Estendida* (Chomsky, 1981), que prevê que toda oração deve conter um sintagma nominal em posição de sujeito. Portanto, seria um atributo estritamente formal, por ser particular à sintaxe.

[3] Na verdade, há uma leve influência da ordem linear sobre a análise sintática no modelo, relativa à composição de adjuntos, como veremos mais adiante.

conjunto restrito de assunções tanto sobre o que predispõe o aprendiz em relação ao conhecimento gramatical, quanto sobre a maquinaria disponível (símbolos, operações e interfaces) à componente sintática.

O IASMIM incorpora ideias e noções minimalistas, tais como as operações *set-Merge* e *pair-Merge*, como veremos mais adiante. Mas se distingue um tanto fortemente de tais modelos, por outro lado, quando não assume a operação de movimento. Esta decisão, entretanto, não pode ser considerada não-minimalista a priori, visto que a única operação virtualmente necessária é a de combinação.[4] Embora os resultados do modelo não sejam suficientes para afirmar que a operação de movimento seja desnecessária, eles indicam que para as variações de ordem envolvidas em interrogativas e nas diferentes ordens canônicas das línguas, é possível prescindir desta operação.

É *integrado*, pois a aquisição sintática está integrada ao processo de aquisição lexical, de modo que a aquisição sintática só se inicia de fato, quando o aprendiz passa a ser capaz (o que ocorre gradativamente) de reconhecer completamente os itens lexicais de um enunciado e então enviá-los para o processamento sintático. Ao não assumir um léxico dado previamente, obtém-se uma transição mais gradual na aquisição, visto que o aprendiz – em função das características da estratégia de aprendizagem – vai em geral iniciar a aquisição pelos enunciados mais simples para só mais tardiamente processar os mais complexos.

O modelo é ao mesmo tempo de *processamento* e de *aprendizagem*, visto que novas sentenças submetidas ao aprendiz são primeiramente enviadas ao analisador. Se, num dado ponto da análise, não houver regra capaz de processar a sentença, ou seja, se a gramática for momentaneamente insuficiente, o analisador dispara os procedimentos de aquisição. Por outro lado, como está orientado apenas para a aquisição em situações em que a alavancagem semântica é assumida como plausível, o analisador embutido no modelo não dispõe de meca-

[4] Suponhamos, por exemplo, que não houvesse distinções de ordem linear intra e entre línguas e que todos os elementos aparecessem em suas posições de base; neste caso, não haveria necessidade de explicar aparentes deslocamentos de constituintes e, portanto, operações de movimento deixariam de ser necessárias. Entretanto, ainda assim, não se poderia abrir mão da operação *merge*.

nismos para lidar com ambiguidade, visto que a informação semântica é em geral suficiente para desambiguizar os contextos de análise.

O formalismo assumido no modelo é não-transformacional e, portanto, o analisador tem regras para construir estruturas, mas não para transformá-las por movimento, apagamento, substituição ou outras operações. O modelo conta com uma representação de informações semânticas e sintáticas em matrizes de atributo-valor (cf. Shieber, 1986) e com regras moderadamente sensíveis ao contexto que fazem menção direta às informações semânticas e também a um *contexto local* de tamanho pré-fixado (nós cíclicos e antecipação).

Outra propriedade do modelo é a de não utilizar papéis temáticos explícitos, de modo a evitar as complicações advindas das controvérsias envolvendo esse tema, tais como, por exemplo, sobre quais seriam os papéis temáticos disponíveis e se assumi-los como um conjunto discreto (e não como elementos num contínuo ou agrupamentos prototípicos) é suficiente para a adequada descrição das línguas (ver Dowty, 1991). A estratégia adotada foi conceber um sistema de representação semântica baseado na articulação dos atributos propostos em Filip (1999) para a distinção entre tipos de eventualidades com as estruturas conceituais propostas em Jackendoff (1983, 1990) e Pinker (1989).

O modelo simula aspectos gerais da aquisição no período compreendido entre os 10 meses, quando surgem as primeiras palavras, e os 5 anos, o que inclui, por exemplo, a emergência de construções que evidenciam a recursividade da gramática, interrogativas sim/não e com elementos-Qu, entre outros construções (ver Tabela 2.2, do Cap. 2). Tais aspectos permitem testar o modelo em termos do poder de expressão da gramática, conforme assumida aqui. Entre os aspectos fora do alcance do modelo, vale citar os fenômenos de sobregeneralização flexional (e outros próprios à performance), ligação (princípios A, B e C) e quantificação.

Outro aspecto que também está fora do escopo da modelagem é o da mudança qualitativa e quantitativa dos dados de entrada apresentados ao aprendiz no decorrer da aquisição. É muito provável que certas construções inicialmente descartadas (em função de sua com-

plexidade) pelos procedimentos de aprendizagem passem a ser úteis (pois analisáveis) em estágios mais tardios, seja pelo acúmulo de experiência, seja por desenvolvimentos cognitivos. Ao mesmo tempo, é plausível supor que a distribuição estatística das construções apresentadas à criança varie nos diferentes estágios do percurso. Apenas o primeiro aspecto é parcialmente controlado nesta simulação, através de limitações impostas pelos procedimentos de aquisição. No mais, os dados são apresentados de modo aleatório e o mesmo aparato "cognitivo" está disponível ao aprendiz em qualquer estágio do processo.

Finalmente, a arquitetura do LAD – como implementado no IASMIM – é ilustrada na Figura 4.1 a seguir. Seus componentes são basicamente os mesmos apresentados no Capítulo 2, ou seja, os dados de entrada, os procedimentos de aquisição, a gramática e o analisador, que passará a ser designado como *processador* para fazer menção tanto ao mecanismo de reconhecimento lexical quanto ao analisador sintático. Cada componente é apresentado em detalhes nas próximas seções.

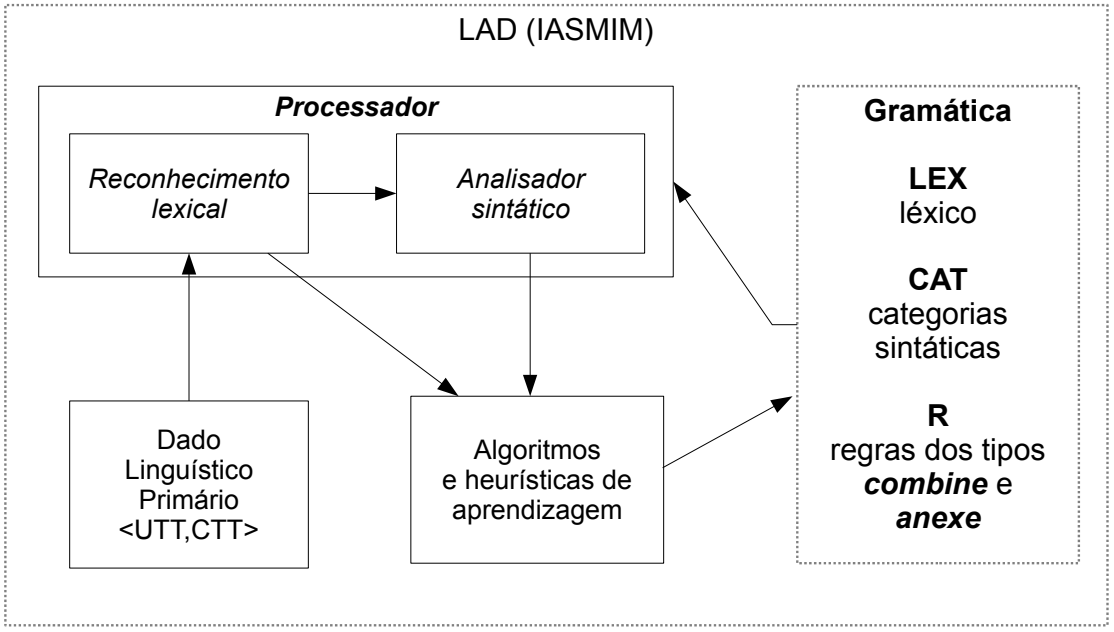


Figura 4.1: Esquema do LAD no IASMIM

4.2 Os dados de entrada

O papel dos DLP no IASMIM é fundamental. Isso se deve ao fato de que nesta simulação as propriedades das categorias sintáticas são determinadas pelo atributos semânticos. Além disso, a ordem linear dos itens lexicais também tem influência sobre as regras gramaticais adquiridas, embora não influa sobre as árvores sintáticas geradas pelas análises. Dado que um dos objetivos do modelo é induzir categorias gramaticais a partir dos DLP, a pergunta que orientou a definição das propriedades estruturais da representação semântica e dos atributos que a compõem é: *o que seria plausível assumir como estando disponível para a criança através da percepção?*

Esta pergunta é de extrema importância, visto que não é plausível fazer certas suposições como, por exemplo, a de que a representação semântica do verbo contenha (no caso do inglês ou do português) atributos de concordância com o sujeito, o mesmo valendo para adjetivos. Talvez essa suposição seja plausível em estágios mais tardios da aquisição, quando a morfologia pode ser explorada mais ativamente pela criança como meio de obter pistas sobre o significado de novas palavras. Nesse sentido, os dados disponíveis são mais restritos do que uma abordagem mais ingênua sugeriria. Por outro lado, os dados são ricos, visto que a representação semântica é altamente informativa (em termos de relações estruturais) para a aquisição sintática.

Por outro lado, na tarefa de determinar a semântica de predicados, é muito fácil se perder em considerações intermináveis e até circulares, especialmente quando se trata de construções gramaticais mais complexas, como as sentenças contendo subordinações ou orações com elementos topicalizados. Portanto, é preciso estabelecer certos limites e fazer algumas escolhas em certa medida arbitrárias. Novamente, essas decisões terão impacto direto sobre o desempenho e mesmo sobre a própria possibilidade de sucesso do aprendiz em certos casos.

4.2.1 Características formais

Cada dado de entrada consiste do par ⟨UTT,CTT⟩ em que UTT é o *enunciado* (que pode ser um sintagma isolado ou uma sentença completa) e CTT é o *contexto*, uma representação semântica do enunciado. Seja I o conjunto dos DLP, então $I = \{\langle UTT_1, CTT_1\rangle$, $\langle UTT_2, CTT_2\rangle$, ..., $\langle UTT_n, CTT_n\rangle\}$, n finito. UTT consiste na sequência de itens lexicais do enunciado, assumidos como previamente segmentados e disponíveis para reconhecimento lexical.

Por sua vez, a informação semântica é representada através de matrizes de atributo-valor, que são também chamadas de "estruturas de atributos". Como o nome explicita, tais matrizes são compostas por atributos (p.e., ±definido, ±animado etc.) e valores, que podem ser números, cadeias de caracteres (p.e., 'SG', 'gato' etc.), sub-matrizes ou, ainda, um valor *subespecificado* (indeterminado). Segundo Shieber (1986), AVMs podem ser vistas (matematicamente) como funções parciais de atributos para seus valores. Em (12), temos um exemplo de AVM representando os atributos `[+definido, pessoa=3]`.

(12)
$$\begin{bmatrix} \text{DEFINIDO} & + \\ \text{PESSOA} & 3 \end{bmatrix}$$

Além de conter valores estruturados (sub-AVMs), estas estruturas podem ainda conter *reentrâncias*, isto é, atributos que compartilham o *mesmo* valor. É importante ressaltar que *compartilhar* um valor é diferente de simplesmente ter valores iguais: no primeiro caso há apenas *um único valor* compartilhado por dois atributos, enquanto no segundo há de fato dois valores, que ocorre de serem iguais. Reentrâncias são identificadas através de índices dentro de pequenas caixas, como por exemplo $\boxed{1}$. As duas propriedades (valores complexos e reentrância) são exemplificadas respectivamente em (13) e (14) a seguir:

(13)
$$\begin{bmatrix} \textsc{definido} & + \\ \textsc{concordância} & \begin{bmatrix} \textsc{número} & \text{singular} \\ \textsc{pessoa} & 3 \end{bmatrix} \end{bmatrix}$$

(14)
$$\begin{bmatrix} \textsc{atributo1} & \boxed{1}\ \text{a} \\ \textsc{atributo2} & \boxed{1} \end{bmatrix}$$

AVMs podem estar relacionadas por *subsunção*, isto é, uma relação que permite ordenar, por ordem de generalidade, AVMs cujos atributos em comum tenham os mesmos valores. Uma AVM a subsume uma AVM b se a contém um subconjunto da informação em b. Em outras palavras, é possível interpretar a como sendo uma superclasse de b ou, inversamente, b como uma subclasse (mais específica) de a. Por exemplo, a AVM em (15) subsume a AVM em (13), visto que a primeira contém um subconjunto dos atributos da segunda. O subconjunto não precisa ser próprio, entretanto; assim, AVMs iguais também subsumem uma à outra.

(15)
$$\begin{bmatrix} \textsc{definido} & + \end{bmatrix}$$

Finalmente, temos a operação de *unificação* (representada por $\sqcup$) que pode ser definida como uma função de combinação de AVMs, de modo a obter a estrutura mais geral que contenha as informações das AVMs combinadas. Para que duas AVMs sejam unificáveis, é necessário que elas não tenham atributos conflitantes, isto é, atributos de mesmo nome cujos valores difiram. Nestes casos, tais conflitos levam a uma falha na unificação. Excepcionalmente, os valores podem diferir desde que um deles seja um valor subespecificado (representado aqui por "@"): neste caso, a unificação favorece o valor especificado. Em (16) são exemplificadas as várias situações:

(16) `[+definido]` $\sqcup$ `[+definido]` = `[+definido]`

`[+definido]` $\sqcup$ `[ ]` = `[+definido]`

`[+definido]` $\sqcup$ `[+plural]` = `[+definido, +plural]`

`[+definido]` $\sqcup$ `[@definido]` = `[+definido]`

`[+definido]` $\sqcup$ `[-definido]` = *falha*

A opção por AVMs, do ponto de vista linguístico, como forma de representar a informação semântica se deve à flexibilidade e potencial representacional do formalismo, que permite tanto representar a estrutura conceitual, quanto a estrutura sintática, como vemos nas diversas aplicações do mesmo (Gazdar et al., 1985, Bresnan, 2001, entre outros). Além disso, no que diz respeito à implementação, há uma excelente biblioteca para a linguagem Python, a NLTK[5], que entre as inúmeras funcionalidades que disponibiliza para processamento de línguas naturais inclui também um pacote para manipulação de AVMs, de modo que boa parte do trabalho de programação necessário para este fim pode ser poupado.

4.2.2 As bases da representação semântica

Como mencionado anteriormente, as bases da representação semântica concebida para esta simulação estão nas propostas de Filip (1999), Jackendoff (1983, 1990) e Pinker (1989). Parte da representação diz respeito aos atributos que constituem o que chamo de *predicado principal* e outra parte diz respeito à composição da estrutura (semântico-)conceitual. Começando por esta última, optei por adotar a ideia geral de decomposição lexical proposta em Jackendoff (1990), certamente não em todo seu potencial representativo. Essa escolha tem duas motivações principais. Primeiramente, ela é também a base da representação semântica assumida em Siskind (1996), cujo algoritmo para aquisição lexical serviu de modelo para a aquisição lexical no IASMIM.

Em segundo lugar, a proposta de Jackendoff (1990) vem ao encontro de alguns objetivos da pesquisa, em especial, o de não utilizar papéis temáticos explícitos e nem tampouco informações explícitas de subcategorização.[6] Na proposta de Jackendoff, papéis temáticos

[5] "Natural Language Toolkit", em http://nltk.org/ (último acesso, 22/06/2013).

[6] Embora este segundo efeito não se deva à representação conceitual em si, mas à interação entre a repre-

são vistos como (relações entre) posições da estrutura conceitual, de tal modo que um mesmo elemento pode ocupar mais de uma posição e, portanto, ter mais de um papel temático ou um mesmo papel temático pode ser determinado pelo conjunto de posições co-indexadas na estrutura conceitual. Para ilustrar melhor, tomemos a representação para o verbo *drink*, proposta pelo autor[7]:

(17) *drink*

$$[_{Event}\ \text{CAUSE}([_{Thing}\ ]_i,\ [_{Event}\ \text{GO}([_{Thing}\ \text{LIQUID}]_j,$$
$$[_{Path}\ \text{TO}([_{Place}\ \text{IN}([_{Thing}\ \text{MOUTH OF}([_{Thing}\ ]_i)])])])])]$$

Vamos chamar elementos CAUSE, GO, LIQUID, TO, IN e MOUTH OF, de *símbolos conceituais*. Note que o argumento indexado por i aparece em duas posições da estrutura, isto é, como argumento de CAUSE (em geral a posição relativa ao "agente") e como argumento de "MOUTH OF" (neste caso, com papel de "possuidor"). Jackendoff argumenta que não apenas um mesmo NP pode ter mais que um papel temático, como no caso acima, mas que também um mesmo papel temático pode ser compartilhado por dois ou mais NPs, como em "*The box has books in it*", em que o autor afirma que *the box* e *it* possuem o mesmo papel temático (digamos, o de "contêiner").

Ao propor que tais papéis são posições na estrutura conceitual, Jackendoff oferece uma forma menos rígida de capturar as correspondências entre a sintaxe e os papéis temáticos, ao mesmo tempo em que se abre à riqueza de papéis temáticos possíveis. No entanto, para os fins da presente simulação e tomando ainda o exemplo em (17), foi necessário desenvolver a representação, de modo a tornar mais explícitas as funções estruturais de cada elemento envolvido. Antes de prosseguir, vale ressaltar que para esta simulação os "tipos conceituais" ("Event", "Thing", etc.) indicados em (17) não são relevantes e podem ser desconsiderados. O elemento LIQUID aí funciona como um restritor semântico sobre o tipo de argumento

sentação semântica, a aquisição lexical e a aquisição sintática.

[7] Jackendoff (1990, p.53). Detalhes irrelevantes para esta discussão foram omitidos.

válido, que também não são utilizados nesta modelagem. Com isso, obtemos:

(18) *drink*

CAUSE($[]_i$, GO($[]_j$, TO(IN(MOUTH-OF($[]_i$)))))

De alguma forma, fica claro que o símbolo TO(...) compõe o símbolo GO, assim como GO(...) compõe CAUSE. Ademais, CAUSE, GO e MOUTH-OF, cada uma, possuem argumentos, identificados com os respectivos índices. Repare que decompor "drink" em CAUSE(GO(TO-IN-MOUTH-OF)) é uma opção relativamente arbitrária, especialmente o trecho TO-IN-MOUTH-OF, que poderia ser, por exemplo, DO-COPO + PARA-A-BOCA-DE (para usar o português) ou DESAPARECER-NA-BOCA-DE, entre outras coisas que poderíamos imaginar. A questão é se seria possível evitar este tipo de decisão visto que, se para alguns casos é mais fácil chegar a um consenso, para outros a possibilidade de decomposição parece ser ilimitada, como sugere o exemplo a seguir, retirado de Pinker (1989, p.215) (omitindo ainda alguns elementos):

(19) "*Bob told a story to the kids*"

tell

ACT(**Bob**, **story**–for/to(BE$_{epistemic}$(**story**, AT(**kids**))), GO$_{effect}$(**story**, TO(AT(**kids**))))

Portanto, optei por limitar a decomposição da seguinte forma: os predicados são decompostos em relação à causatividade, à estatividade (o que nos garante os argumentos interno e externo) e, quando necessário, à componente indireta do predicado, como aquela relativa ao objeto indireto do verbo *dar*. Para continuar identificando unicamente um predicado, utilizo o próprio como um símbolo conceitual. Assim, para os verbos "drink" e "tell" acima, teríamos:

(20) *drink*

CAUSE(x, BECOME(y, DRINK))

tell

CAUSE(x, BECOME(y, TELL(TO(z))))

Nas representações acima, os símbolos DRINK e TELL representam as propriedades particulares destes predicados, que permanecem não-decompostas, exceto no caso de TELL, em que o argumento indireto é indicado explicitamente por TO(z), sendo que este símbolo se compõe diretamente com TELL e deve ser visto como parte de sua semântica. Com isto, foi possível impor um limite bastante prático para a especificação dos predicados verbais. Incluí, ainda, ao repertório de predicados conceituais, elementos funcionais tais como definitude, tempo, entre outros, que nas propostas de Jackendoff (1983, 1990) e Pinker (1989) são deixados de lado. Assim, cheguei a representações conceituais como as seguintes:

(21) "O livro" : DEF(**livro**)

"O livro na mesa" : DEF(**livro**([EM(DEF(**mesa**))]))

"O livro é azul" : DECL(PRES(BE(**azul**, DEF(**livro**))))

"João comprou um livro" :

DECL(PAST(CAUSE(BE(**comprado**, INDEF(**livro**)), **João**))

"O livro é azul?" : INT(PRES(BE(**azul**, DEF(**livro**))))

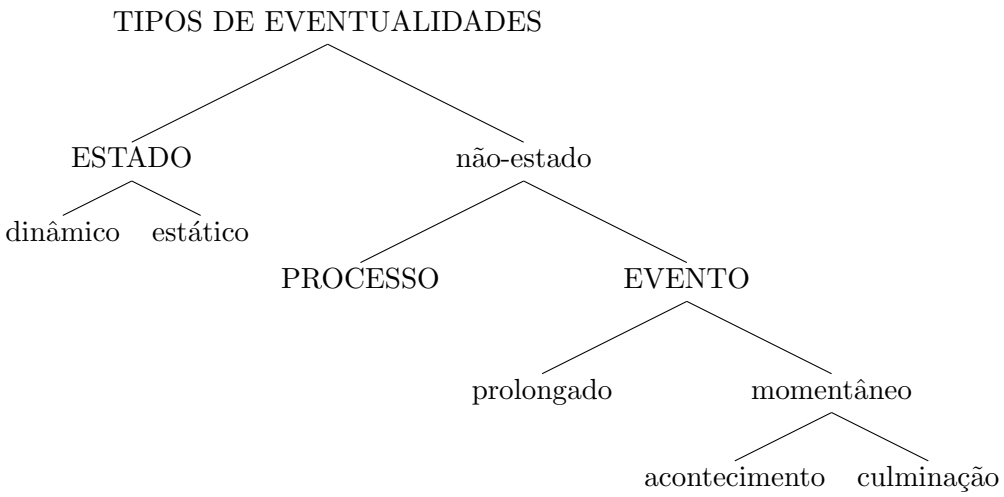
Porém, ainda faltava explicitar as funções estruturais dos símbolos. Para isso, defini uma *estrutura conceitual básica* da forma $F(CF, A, [M_1, M_2, ..., M_n])$, $n \geq 0$. A função estrutural F diz respeito ao que chamo de "predicado principal" que determina as propriedades principais do predicado. A função estrutural CF indica o "predicado componente" do predicado principal. A função estrutural A é utilizada para argumentos de relações, tais como o argumento interno, o externo e argumentos indiretos. Finalmente, a função estrutural M permite vincular modificadores opcionais que incidem sobre F (a lista de modificadores é não-ordenada, isto é, os índices indicados fazem referência à quantidade de modificadores, não à sua ordem relativa).

Enquanto F é obrigatória na especificação de qualquer predicado, as entradas CF e A são obrigatórias a depender do predicado e M é estritamente opcional. Aqui vale ressaltar: há um diferença de estatuto bastante importante entre CF e A, de um lado, e M, de outro. As duas primeiras compõem F, enquanto a última atua como uma espécie de "filtro" sobre propriedades inerentes ao predicado (local, maneira, etc.). As funções estruturais CF, A e M_i, quando presentes, expandem cada uma para a mesma estrutura conceitual básica. Abaixo, um exemplo de AVM conceitual para a sentença "*João comprou um livro*":

(22)
$$\begin{bmatrix} \text{UTT} & \texttt{João comprou um livro} \\ \text{CTT} & \begin{bmatrix} \text{F} & \text{DECL} \\ \text{CF} & \begin{bmatrix} \text{F} & \text{PAST} \\ \text{CF} & \begin{bmatrix} \text{F} & \text{CAUSE} \\ \text{CF} & \begin{bmatrix} \text{F} & \text{BECOME} \\ \text{CF} & \begin{bmatrix} \text{F} & \text{comprar} \end{bmatrix} \end{bmatrix} \\ \text{A} & \begin{bmatrix} \text{F} & \text{INDEF} \\ \text{CF} & \begin{bmatrix} \text{F} & \text{livro} \end{bmatrix} \end{bmatrix} \\ \text{A} & \begin{bmatrix} \text{F} & \text{DEF} \\ \text{CF} & \begin{bmatrix} \text{F} & \text{João} \end{bmatrix} \end{bmatrix} \end{bmatrix} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Definidas as propriedades estruturais da representação conceitual, era preciso ainda especificar os atributos definidores dos diferentes predicados, que chamarei "atributos lexicais". O conjunto destes atributos foi definido em parte de modo arbitrário – tomando atributos normalmente citados na literatura, tais como `[animado]` ou `[person]` – e em parte emprestando ideias de Filip (1999) e Pinker (1989), particularmente, para a especificação dos atributos dos predicados verbais. Tomemos Filip (1999), inicialmente, que propõe um conjunto de atributos para distinguir entre os tipos de eventualidades apresentados a seguir:

(23)



Esta classificação remonta à Bach (1986). Estados estáticos seriam, por exemplo, "possuir (uma casa)", "amar (alguém)", etc. Estados dinâmicos seriam "(estar) sentado", "(estar) doente", etc. Entre os processos, teríamos "andar" e "chover". Eventos prolongados seriam "construir (uma casa)", "comer um sanduíche", entre outros. Culminações incluem "decolar", "chegar", "partir", etc., enquanto acontecimentos incluem "piscar", "disparar", "chutar", etc. Vale ressaltar que a autora não deixa de observar que classificações deste tipo são apenas instrumentais, pois a classe de um dado verbo pode mudar, a depender dos elementos que o acompanham na sentença (p.e., "caminhar" é um processo, mas "caminhar para a universidade" é um evento prolongado). Para distinguir entre estas eventualidades, Filip (1999, p.110) propõe o seguinte esquema:

(24)

	change	quantization	temporal extent
Estado estático	–	–	+
Estado dinâmico	–	–	+
Processo	+	–	+
Evento prolongado	+	+	+
Culminações	+	+	+
Acontecimentos	+	+	–

Vê-se que neste esquema alguns pares de eventualidades apresentam os mesmos atributos. Como o objetivo era ter uma distinção total, optei por rever o esquema acima, incluindo

alguns atributos e renomeando outros, mas mantendo as distinções básicas do esquema anterior, como mostra (25):

(25)

	change	persistent	quantizable	punctual	culmination (*temporal extent*)
Estado estático	–	+	–	–	–
Estado dinâmico	–	–	–	–	–
Processo	+	–	–	–	–
Evento prolongado	+	–	+	–	–
Culminações	+	–	+	+	+
Acontecimentos	+	–	+	+	–

Finalmente, faltava distribuir estas propriedades através da estrutura conceitual, visto que o esquema não pressupõe decomposição lexical. O atributo `change` define o símbolo conceitual BE (`-change`) (quando não há mudança de estado, como em "estar sentado" e "amar"), e BECOME (`+change`) para aqueles em que há uma mudança (p.e., "construir" ou "chegar"). Os demais atributos compõem os próprios símbolos relativos aos predicados verbais. Para o símbolo CAUSE um novo atributo foi incluído, `control`, que permite distinguir verbos estritamente de ação (`+control`), como "bater" ou "andar", de verbos em que o papel do argumento externo é determinante, mas de uma forma não controlada por ele (`-control`)[8], como em "amar" ou "perceber". Com isso, temos a distribuição a seguir:

(26)
$$\begin{bmatrix} \text{F} & [\text{CONTROL} \quad \pm] & \\ \text{CF} & \begin{bmatrix} \text{F} & [\text{CHANGE} \quad \pm] & \\ \text{CF} & \begin{bmatrix} \text{F} & \begin{bmatrix} \text{PERSISTENT} & \pm \\ \text{QUANTIZABLE} & \pm \\ \text{PUNCTUAL} & \pm \\ \text{CULMINATION} & \pm \end{bmatrix} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

O intuito em estabelecer este conjunto de atributos para a especificação dos predicados

[8] Pinker (1989) propõe um atributo de mesmo nome, mas seu papel na proposta do autor é distinto do papel neste modelo. Ali, Pinker utiliza o atributo para distinguir entre dois grupos de funções semânticas: de um lado, ACT/HAVE (+control) e, de outro, GO/BE (–control).

verbais é o de permitir ao modelo capturar a relação de certas classes de verbos com certas classes de adjuntos, através da identificação de co-ocorrências entre os atributos definidores de cada classe. Dessa forma, o modelo dispõe de informação para induzir classes lexicais mais refinadas, capazes de capturar a distribuição de verbos e adjuntos.

4.2.3 O conjunto de atributos

Para os fins da presente simulação, o corpus de entrada foi construído com base no conjunto de atributos apresentados a seguir. Futuramente, nada impede que o conjunto incorpore novos atributos, não apenas para capturar distinções mais sutis entre tipos de adjetivo, nominais, etc., mas também para permitir que a representação conceitual cubra fenômenos que ficaram fora do escopo desta pesquisa, tais como aspectos informacionais (tópico, foco, etc.), fenômenos de ligação, orações relativas, etc.

Atributo	**Valores**	**Descrição**
		— Objetos e Eventualidades —
`concept`	*	Informa sobre a característica semântica última que distingue este conceito de qualquer outro. Este atributo representa, em geral, o conjunto de atributos semânticos que entram na determinação de objetos, eventualidades, adjetivos e advérbios, porém que parecem não ter relevância clara para os fenômenos sintáticos, ficando então implícitos nesse atributo.
`quantizable`	+/–	Para objetos, informa se é contável (+) ou não (–). Para eventualidades, distingue eventos (+) de processos (–).
		— Objetos —
`animacy`	+/–	Informa sobre a animacidade do objeto.
`definite`	+/–	Informa sobre a definitude do objeto.
`plural`	+/–	Informa se é um objeto singular (–) ou plural (+).
`person`	1/2/3	Informa se o objeto refere ao falante (1), ao interlocutor (2) ou a um terceiro (3).
`wh`	+/–	Informa se o objeto (ou propriedade, no caso de adjetivo) tem propriedades Qu (i.e., se é variável).
		— Eventualidades —

`control`	+/–	Distingue entre verbos transitivos com controle da ação (+), como "construir", e verbos como "amar" ou "ver" (–). Compõe o símbolo conceitual CAUSE.
`change`	+/–	Indica se há mudança de estado (+) e ou não (–). Compõe o símbolo conceitual BE/BECOME.
`persistent`	+/–	Distingue entre estados estáticos (+) e dinâmicos (–).
`punctual`	+/–	Culminações e acontecimentos, tais como "alcançar" ou "piscar", tem valor (+), enquanto processos e eventos com duração temporal e estrutura interna, como "construir", tem valor (–).
`culmination`	+/–	Distingue entre culminações (que pressupõem um processo ou evento que finalizam) (+) e acontecimentos (–).
`finite`	+/–	Distingue entre passado/futuro (+) e presente/infinitivo (–).
`realis`	+/–	Quando se tem `[+finite]`, distingue entre passado (+) e futuro (–), caso contrário, distingue entre presente (+) e infinitivo (–).[9]
		— Outros atributos —
`relation`	*	Indica relações envolvendo um argumento, tais como "ext-arg" (argumento externo ou "agente"), "to-poss" (transferência de posse), etc. Estão, nessa simplificação, no lugar de papéis temáticos. Numa representação conceitual com maior decomposição dos predicados, este atributo seria substituído por uma estrutura mais complexa.

Tabela 4.1: Conjunto de atributos lexicais utilizados na simulação

Note que não há atributos "formais", como ±N e ±V (cf. Chomsky, 1970), por exemplo. A ideia é que o efeito sintático de tais atributos seja obtido a partir da generalização das regras e categorias sintáticas, cuja base são os atributos acima. Na verdade, uma das hipóteses de base do modelo é exatamente a de que padrões sintáticos são determinados pela interação entre atributos do conjunto definido acima, propriedades das expressões conceituais e propriedades da sintaxe. Isso fica mais claro na próxima seção, quando falo das propriedades do conhecimento gramatical a ser adquirido. Em suma, a abrangência gramatical do modelo depende fundamentalmente da extensão do conjunto de atributos, bem como da

[9] O termo "realis" remete a aspectos envolvendo modalidade (realis/irrealis). Neste modelo, o termo é usado de um modo mais ingênuo (embora não de todo desvinculado dessa questão), especificando a "ancoragem" do evento no presente: quando '+', indica algo concluído ou habitual; quando '–', indica algo no futuro ou indeterminado (infinitivo) e, portanto, não "ancorado" no presente.

cobertura gramatical da representação conceitual.

4.2.4 Sumário

Nesta seção, foram apresentadas as propriedades estruturais dos dados de entrada, que consistem do par ⟨UTT,CTT⟩, sendo UTT a cadeia de palavras (pré-segmentada) e CTT a expressão (semântico-)conceitual que codifica o sentido próprio ao enunciado. O sistema de representação conceitual foi desenvolvido com base na articulação das propostas de Filip (1999), Jackendoff (1983, 1990) e Pinker (1989), e se caracteriza pela aplicação de decomposição lexical dos predicados da língua em conjunto com atributos lexicais para capturar nuances próprias às "partes" resultantes da decomposição.

Vale ressaltar que a representação semântica foi concebida para ser neutra em relação a línguas particulares. Assim, espera-se que as expressões conceituais possam ser mantidas relativamente fixas, bastando traduzir os enunciados correspondentes para a língua sendo investigada. Ou seja, assume-se aqui que as expressões conceituais sejam relativamente universais, com variações pontuais na estrutura de um ou outro predicado, a depender de como este é conceitualizado numa dada língua. Isso não impede, entretanto, que diferentes línguas investigadas impliquem o acréscimo de atributos lexicais (p.e., para lidar com a evidencialidade) e de novos predicados. Aliás, seria de fato o esperado.

O quanto essa hipótese se sustenta é uma questão empírica que, no contexto do presente modelo, depende da criação de corpora para diferentes línguas. Neste sentido, é importante salientar que em função do modelo não ser capaz de segmentar palavras e de acessar morfemas individualmente, a aplicação do modelo a línguas polissintéticas não poderia ser explorada de forma interessante, pelo menos não em relação às orações que consistem de uma única palavra. Ou seja, o modelo é capaz de explorar relações entre palavras e, portanto, alguma analiticidade é necessária para que o modelo retorne resultados interessantes.

4.3 A gramática

A gramática do aprendiz, em um estágio qualquer da aquisição, consiste de (i) um léxico LEX de palavras e seus respectivos sentidos (i.e., palavras podem ter mais de um sentido vinculado quando apresentam variantes homonímicas ou polissêmicas), (ii) um conjunto CAT de categorias sintáticas induzidas na aprendizagem sintática, e (iii) um conjunto R de regras do analisador para processamento das sentenças. O aprendiz inicia a aquisição com a gramática G_0 = {LEX=∅, CAT=∅, R=∅}, postulando novas palavras, categorias e regras à medida em que processa os dados de entrada.

4.3.1 O léxico

O léxico no IASMIM foi implementado segundo a proposta de Siskind (1996). Nesta, no intuito de tornar o algoritmo de aquisição lexical robusto à homonímia/polissemia e a eventuais ruídos (pares palavra/sentido inválidos), Siskind propõe um léxico organizado em duas camadas: na primeira camada estão as palavras (w_i) e na segunda camada estão os sentidos ($S(w_i) = \{s_1, s_2, ..., s_n\}$) conjecturados para cada palavra. Quando não-unitário, $S(w_i)$ indica que a palavra possui variantes homonímicas ou polissêmicas, ou que um ou mais sentidos foram incorretamente conjecturados.

A cada sentido s_j vinculado a uma palavra w_i está associado um "fator de confiança" (FC), que indica o número de vezes que s_j esteve envolvido no reconhecimento com sucesso das palavras de um enunciado. Este fator tem duas funções: a primeira, a de permitir que sentidos incorretamente conjecturados sejam distinguidos dos demais, visto que seu FC tenderá a ser baixo. E, ainda, permitem lidar com ambiguidade, caso o reconhecimento lexical deva proceder sem o auxílio de informação contextual: os sentidos com maior FC são favorecidos no reconhecimento. Veremos estes aspectos mais detalhadamente quando estivermos considerando o processador e a aprendizagem.

4.3.2 As categorias sintáticas

No IASMIM é implementada uma noção de "objeto sintático" – introduzida na teoria sintática pelo menos desde Chomsky (1993) – que aqui se caracterizam como os elementos de fato manipulados pela sintaxe para a composição da estrutura arbórea do enunciado, isto é, da raiz até os elementos terminais, a árvore consiste de objetos sintáticos combinados. Veremos outras propriedades de tais objetos mais adiante, sendo relevantes nesse momento as propriedades *sem* e *cat*. Para que possam ser processados pelo analisador, os itens lexicais reconhecidos na fase de processamento lexical são transformados em objetos sintáticos antes da análise iniciar (objetos, no caso, que irão se localizar em posições terminais da árvore obtida).

Como cada item lexical consiste do par $\langle w, s \rangle$, o elemento s (a expressão conceitual do sentido) é copiada para a propriedade *sem* do objeto correspondente. O procedimento gera, em seguida, um rótulo categorial para o objeto, tendo como base s. Portanto, as primeiras categorias sintáticas induzidas na aquisição são categorias exatamente equivalentes aos itens lexicais da língua. Na medida em que avança a aquisição sintática, entretanto, o aprendiz vai produzir objetos sintáticos constituintes (que incluem outros objetos), gerando novas categorias sintáticas (sintagmáticas). Além disso, à medida em que identifica contextos favoráveis, irá *generalizar*, subespecificando categorias sintáticas para produzir novas e mais abstratas.

Portanto, o caminho do conhecimento sintático do aprendiz consiste basicamente em postular e utilizar categorias sintáticas cada vez mais "distantes" das categorias inicialmente conjecturadas, as quais são fortemente ancoradas na semântica específica dos predicados e, por isso, bastante restritas em sua aplicação. Todas as categorias postuladas são "sintáticas", para falar estritamente. Porém, as categorias tardias é que podem ser mais propriamente consideradas sintáticas, visto que são estas que irão assumir o papel normalmente atribuído a categorias puramente formais, tais como as categorias obtidas pela combinação dos atributos

±N e ±V, já citados.

4.3.3 As regras sintáticas

As Figuras 4.2 e 4.3, abaixo, ilustram os tipos de regras sintáticas que o aprendiz pode conjecturar. O primeiro mostra uma regra do tipo COMBINE, que postula um objeto constituinte com base nas propriedades do elemento na primeira célula. O segundo exemplifica uma regra ANEXE responsável por retirar o elemento na primeira célula e anexar ao nó ativo (ACT). As regras em questão ilustram a construção de um DP (no caso, trata-se de um enunciado composto apenas pelo DP, já que CYC e a célula 3 são iguais a *nil*) e os detalhes das mesmas serão discutidos mais adiante, quando tratarmos do analisador sintático.

```
CYC: nil
ACT: nil
Células:
  (1) [ cf: [  ], f: [ +definite ], m: [  ] ]
  (2) [ cf: [ f: [ -animacy, concept: car, person: 3, +quantizable, -wh ],
        m: [  ] ], f: [ -plural ], m: [  ] ]
  (3) nil
Ação: COMBINE
```

Figura 4.2: Exemplo de regra para construção de um DP.

```
CYC: nil
ACT: Node225
Células:
  (1) [ cf: [  ], f: [ +definite ], m: [  ] ]
  (2) [ cf: [ f: [ -animacy, concept: car, person: 3, +quantizable, -wh ],
        m: [  ] ], f: [ -plural ], m: [  ] ]
  (3) nil
Ação: ANEXE
```

Figura 4.3: Exemplo de regra para anexação do determinante ao DP.

As regras em R encapsulam a gramática moderadamente sensível ao contexto adquirida pelo aprendiz. Uma importante propriedade desta gramática é que as análises produzidas por ela ignoram a ordem linear dos elementos terminais. Portanto, para um mesmo conjunto de elementos a mesma árvore será produzida, a despeito da ordem em que aparecem.

Em outras palavras, não há elementos deslocados nas representações arbóreas, pois não há regras de movimento. Pelos menos duas situações distintas podem ser identificadas, quando ocorre variação na ordem linear. A primeira diz respeito a variações de ordem envolvendo construções canônicas (i.e., não-marcadas) das línguas, como as que permitem caracterizar línguas como sendo SVO ou SOV, entre outras.

Para estas, espera-se que o modelo produza exatamente as mesmas estruturas subjacentes, visto que terá como entrada as mesmas expressões conceituais, com distinções apenas na ordem linear dos elementos do enunciado, o que – como já foi dito – não é considerado na estrutura sintática. Outro tipo de variação de ordem que incluo neste conjunto é a que diz respeito a variações como as que ocorrem por questões de "peso fonológico" de um sintagma, produzindo pares como "*O João deu [um carro magnífico e cheio de modernidades] ao filho*" e "*O João deu ao filho [um carro magnífico e cheio de modernidades]*". Nestes casos, também, assumo que não há mudança na expressão conceitual (doravante, EC) que justifique uma estrutura sintática distinta.

Por outro lado, é possível que variações na ordem tenham origem em distinções na EC. Seria o caso, no entanto, de identificar quais aspectos estão exatamente em jogo e de que forma estes aspectos poderiam ser representados no esquema conceitual proposto aqui. Uma vez feito isso, seria de se esperar que tais acréscimos surtissem efeitos sobre as estruturas sintáticas produzidas, mais especificamente, sobre a ordem hierárquica dos elementos. Um possível estudo dessa natureza poderia envolver fenômenos de topicalização, por exemplo.

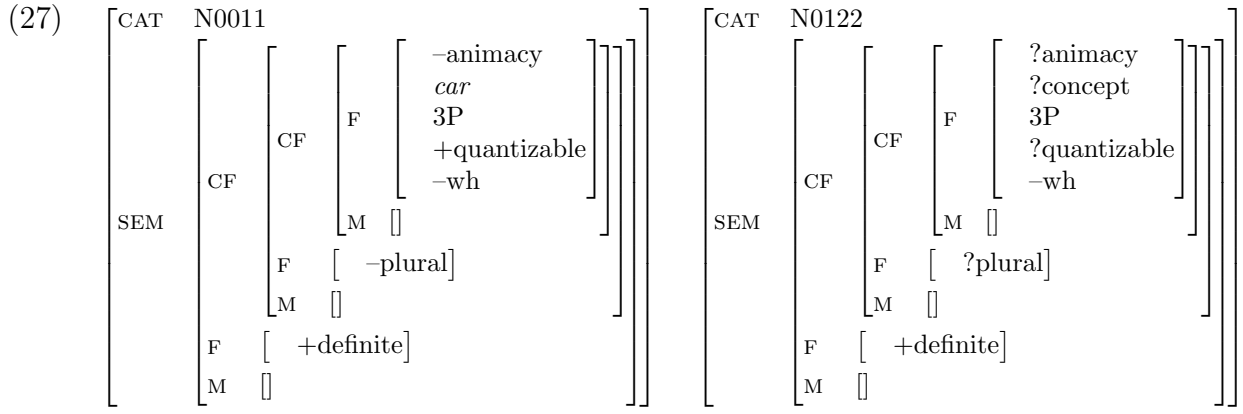
Finalmente, vale ressaltar que o modelo não lida com elementos omitidos, tais como sujeito nulo ou objeto nulo, elipses de VP, etc., ou seja, não há regras para analisar estas construções. Por outro lado, ele é capaz de lidar com elementos abstratos, isto é, elementos da expressão conceitual (sempre ou eventualmente) sem realização lexical. Um exemplo é a formação de relativas sem o complementizador *that*, no inglês. Nestes casos, o aprendiz é capaz de conjecturar categorias abstratas para "completar" a estrutura, visto que os itens

lexicais reconhecidos não contribuem com toda a informação existente na EC. Para isso, no entanto, nenhum tipo novo de regra é necessário: bastam as regras de combinação e anexação.

4.3.4 Sumário

Nesta seção foram apresentadas as características gerais do conhecimento gramatical assumido no modelo. A gramática G do aprendiz consiste na tripla {LEX, CAT, R}, sendo LEX o léxico (composto por palavras e seus respectivos sentidos), CAT o conjunto de categorias sintáticas e R o conjunto de regras sintáticas para análise dos enunciados. Na gramática inicial G_0 do aprendiz, todos estes conjuntos são vazios. Ao final da exposição aos dados de entrada, o aprendiz exibirá uma gramática G_i respectiva ao estágio de aquisição em que se encontra. Categorias sintáticas são inicialmente equivalentes aos itens lexicais, mas à medida em que a aquisição sintática tem curso, vão sendo criadas categorias novas e mais abstratas, tanto relativas a constituintes, como a conjuntos de itens com comportamento sintático equivalente. Em (27), temos exemplos de um estado inicial e de um estado tardio de um sintagma determinante ("?" indica subespecificação):

(27)



Regras sintáticas podem ser de dois tipos: regras de combinação (COMBINE) e de anexação de termos da combinação (ANEXE). Não há regras para movimento e nem regras para elementos omitidos (argumentos nulos e elipses). Regras do segundo tipo são uma

limitação momentânea do modelo. Quanto à operação de movimento, esta está ausente do modelo em função de certas assunções, tais como a de que a ordem linear não é parte da informação sintática e a de que elementos aparentemente deslocados devem ser introduzidos *in situ*. Portanto, *estruturas sintáticas* produzidas pelo analisador sintático ignoram a ordem linear, embora esta seja relevante durante o *processamento*. Quando não envolve distinções na EC respectiva, a variação na ordem linear é considerada "livre" e todas as variantes exibem a mesma estrutura sintática subjacente.

4.4 O processador

O processador no modelo se divide em duas componentes: a *componente lexical* responsável pelo *reconhecimento lexical* e a *componente sintática* responsável pela *análise sintática* do enunciado. Estes componentes são responsáveis também por disparar os procedimentos de aquisição lexical e sintática, em contextos em que fique evidente que o conhecimento gramatical momentâneo é insuficiente. Na presente implementação, a abordagem é estritamente serial: a análise sintática só tem início após o reconhecimento de todos os itens lexicais do enunciado.

4.4.1 A componente lexical

Na componente lexical é realizado o primeiro passo da análise, que consiste em identificar um conjunto de pares (palavra, sentido), doravante (w,s), consistente com o dado de entrada. Ser consistente implica, por um lado, conter a mesma lista de palavras do enunciado e, por outro, conter sentidos que contribuem expressões conceituais licenciadas pelo contexto (CTT). Seja s_i um sentido qualquer e ec_i a EC correspondente. O contexto CTT licencia ec_i se e somente se ec_i unifica com CTT ou com alguma sub-expressão de CTT. Este modelo de reconhecimento e aquisição lexical está fortemente baseado na proposta de Siskind (1996), com alguns ajustes e adaptações em função das características particulares desta modelagem.

Na proposta de Siskind, o reconhecimento e a aquisição lexical são parte de um mesmo processamento. Sobre os procedimentos de aquisição falarei mais adiante. Mas, estando o léxico adquirido, o procedimento de reconhecimento consiste basicamente em calcular todas as combinações possíveis envolvendo as palavras e seus sentidos de modo a obter um conjunto que Siskind designa de "atribuições de sentido possíveis" (ASPs), isto é, o conjunto de elementos do tipo $\{s_{w_1}, ..., s_{w_n}\}$, sendo n o número de palavras do enunciado, que exaure as combinações possíveis envolvendo os sentidos vinculados a cada palavra. De posse deste conjunto, cabe ao procedimento identificar a ASP que seja consistente com CTT e tenha o maior fator de confiança agregado (i.e., a soma dos fatores de confiança de todos os sentidos da ASP). Uma ilustração permite compreender melhor este procedimento. Tomemos o enunciado em (28a) e o léxico parcial em (28b):

(28) a. *Uma manga rosa*

CTT: INDEF(**manga-fruta**([**rosa-espécie**]))

b. Léxico parcial:

uma	INDEF(x), C=203
$manga_1$	**manga-fruta**, C=15
$manga_2$	**manga-parte-de-roupa**, C=33
$rosa_1$	**rosa-espécie**, C=12
$rosa_2$	**rosa-flor**, C=46

O primeiro passo seria calcular as combinações possíveis dos sentidos vinculados, o que produziria o conjunto

(INDEF(x), **manga-fruta**, **rosa-espécie**),
(INDEF(x), **manga-fruta**, **rosa-flor**),
(INDEF(x), **manga-parte-de-roupa**, **rosa-espécie**),
(INDEF(x), **manga-parte-de-roupa**, **rosa-flor**)

que exaure as combinações possíveis envolvendo os sentidos dos itens lexicais em questão.[10] De posse deste conjunto, a rotina identifica os que são consistentes com CTT. Neste exemplo, apenas um é consistente, a saber, a ASP (INDEF(x), **manga-fruta**, **rosa-espécie**). Se mais de uma ASP fosse consistente[11], o procedimento recorreria aos fatores de confiança para escolher a melhor candidata. Uma vez identificada uma ASP, os fatores de confiança dos sentidos envolvidos são incrementados e a ASP é enviada para a componente sintática.

4.4.2 A componente sintática

Primeiramente, é preciso enfatizar que o analisador opera com referência às expressões conceituais vinculadas às palavras do enunciado e não com as palavras propriamente. Por conveniência, continuarei usando os termos "palavra" ou "item lexical", tendo em mente que o que está sendo considerado no âmbito da análise sintática é a EC respectiva. O analisador sintático (*parser*) do modelo, apresentado na Figura 5.9, é semelhante ao proposto em Berwick (1985), que se caracteriza por processar a sentença da esquerda para a direita, num passo único. As estruturas de dados principais são a *área temporária* e a *pilha de nós constituintes* em construção.

A área temporária consiste de células sequenciais, inicialmente contendo as palavras na ordem em que vieram no enunciado. No decorrer da análise, entretanto, a área temporária vai eventualmente conter constituintes completados. Em alguns casos, é permitido ao analisador *antecipar* algumas células antes de decidir o que fazer (no momento, até duas, isto é, o parser opera com uma janela de três células, que inclui a célula em processamento e as duas consecutivas). A janela de três células é também utilizada para compor o contexto local das regras sintáticas, como forma de aumentar seu poder de discriminação.

O analisador funciona no esquema padrão-ação, isto é, a uma dada configuração mo-

[10] Formalmente, sendo L(w) o conjunto de sentidos de uma dada palavra, este cálculo não é nada mais que o produto cartesiano de L(w_1) × ... × L(w_n), sendo n o número de palavras do enunciado.
[11] Por exemplo, na ausência de informação contextual. Note ainda que o modelo ignora questões pragmáticas e de conhecimento de mundo, no caso, envolvendo as combinações de sentido mais prováveis.

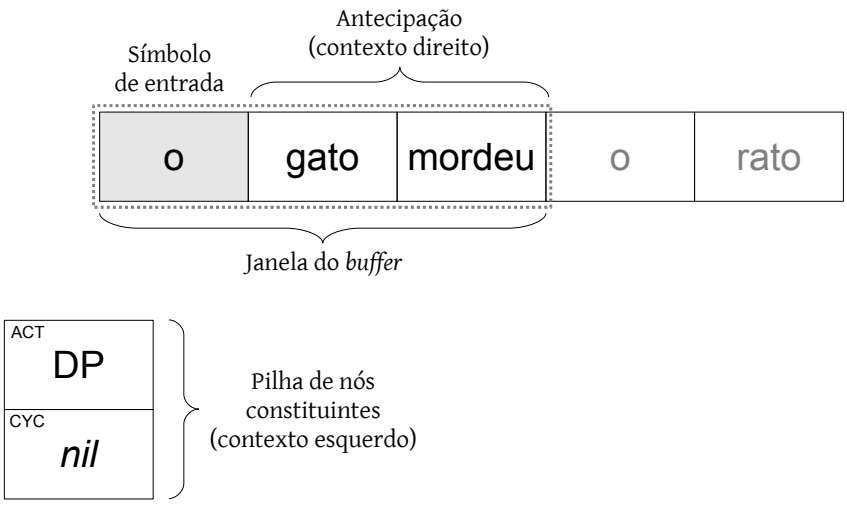


Figura 4.4: O analisador: primeiros estágios de uma análise.

mentânea do analisador (padrão) corresponde uma ação que altera a configuração. A aquisição sintática consiste, portanto, em criar regras para este fim. As regras se compõem de um *estado*, um *símbolo de entrada* e uma *ação* que produzirá uma mudança de estado. Portanto, o conhecimento sintático é modelado como um autômato de estados finitos determinístico. Um estado é composto por um *contexto esquerdo* e um *contexto direito*. O contexto direito corresponde à janela de antecipação comentada acima. O contexto esquerdo permite enriquecer o contexto local de discriminação disponível para as regras, ao mesmo tempo em que permitira testar a hipótese de aprendizagem grau-2 (cf. Culicover & Wexler, 1980).

Segundo esta hipótese, todas as distinções sintáticas das línguas naturais podem ser identificadas num contexto de até dois nós cíclicos, aí entendidos como nós sentenciais (tipo "S" ou "CP", numa notação mais atual).[12] Assim, o contexto esquerdo é composto por um

[12] Em termos estritos, o analisador do IASMIM – e também em Berwick (1985), como indica a implementação em Faria (2009) – não simula de fato a hipótese visto que os nós ACT e CYC são, na prática, os dois nós dominantes imediatos no contexto em análise (por exemplo, no caso de um determinante, o nó ativo seria um DP e o nó cíclico seria nulo ou, eventualmente, um VP ou PP). Esta diferença em relação à hipótese original de Culicover & Wexler (1980) não teve impacto (perceptível) sobre os resultados da modelagem, o que provavelmente se deve ao escopo limitado dos fenômenos tratados. Para fenômenos gramaticais mais complexos seria o caso, portanto, de adaptar o modelo para que além do nó ativo, considerasse como nós cíclicos nós sentenciais e não os nós imediatamente dominantes.

nó ativo (ACT) e um *nó cíclico* (CYC), ambos armazenados na pilha de nós constituintes em construção. Em certos contextos de análise, entretanto, ambos os nós podem ser nulos ou apenas o nó ativo pode estar presente. Os contextos esquerdo e direito compõem o contexto local das regras e são responsáveis pelo caráter moderadamente sensível ao contexto do conhecimento gramatical no modelo.

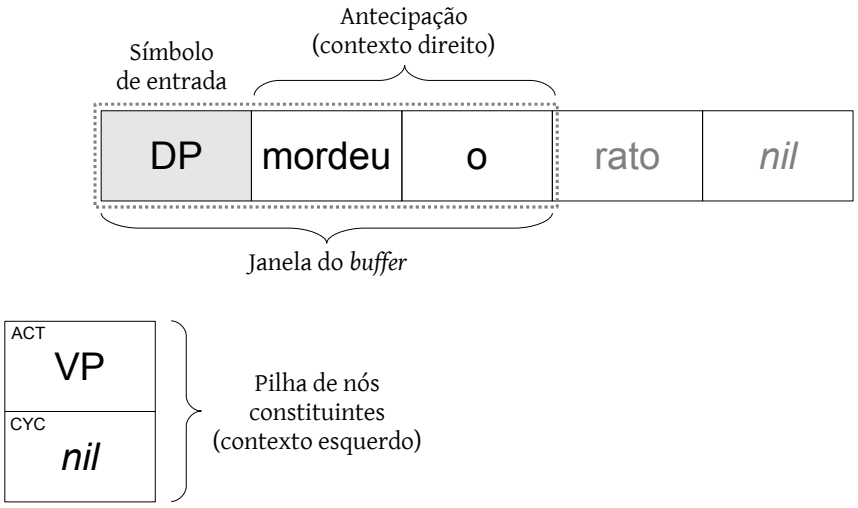


Figura 4.5: O analisador: construção de um sintagma verbal.

O símbolo de entrada corresponde ao elemento na primeira célula da janela momentânea da *área temporária*. Em função das características da componente sintática assumidas pelo modelo de aquisição, as ações de regras disponíveis são ANEXE e COMBINE. Regras transformacionais não fazem parte da gramática neste modelo, em oposição às regras deste tipo em Berwick (1985). Regras do tipo ANEXE tomam o símbolo de entrada momentâneo e o anexam como um dos termos de ACT, o constituinte em formação ativo. Regras do tipo COMBINE criam um novo constituinte – que terá o símbolo de entrada momentâneo como um de seus termos – e o inserem na pilha, como nó ACT.

No decorrer de uma análise, quando o constituinte em ACT é completado (seus termos anexados), este é devolvido para a área temporária, tornando-se o símbolo de entrada momentâneo. Neste caso, se houver um constituinte em formação em CYC este volta a ser o constituinte em formação ativo (ACT). À medida que a estrutura sintática vai sendo

construída, a EC respectiva também vai sendo derivada. Uma análise sintática termina com sucesso quando resta apenas um constituinte sintático na área temporária e sua EC respectiva, em função de ser um modelo de aquisição, unifica com CTT. Quando qualquer dessas condições não é obtida e o procedimento de análise não consegue evoluir, seja utilizando regras conhecidas ou conjecturando novas, a análise falha e o dado é descartado para fins de aquisição.

4.5 Os procedimentos de aprendizagem

Os procedimentos de aprendizagem, assim como o analisador, também se dividem em duas grandes classes: procedimentos para aquisição lexical e procedimentos para aquisição sintática. A aprendizagem lexical se baseia na noção de aprendizagem trans-situacional ("*cross-situational learning*"), isto é, na observação e comparação dos variados contextos em que um dado item lexical aparece. A partir dessa observação, o aprendiz pode identificar as expressões conceituais que necessariamente co-ocorrem com os itens lexicais e, por conseguinte, conjecturar que tais ECs são possíveis sentidos destes itens. O algoritmo de aprendizagem lexical do IASMIM é uma adaptação da proposta em Siskind (1996), que se mostrou bastante apropriada para a abordagem assumida nesta modelagem.

A aquisição sintática, por sua vez, se dá sobre a aquisição lexical e, portanto, será tanto mais bem-sucedida quanto maior for o sucesso daquela, visto que o reconhecimento completo dos itens lexicais dos enunciados é pré-condição para o estágio sintático da processamento. Os procedimentos de aprendizagem sintática recorrem essencialmente ao cruzamento entre o estado momentâneo do analisador e o contexto semântico relativo ao enunciado, para determinar quando se trata de criar uma regra de combinação ("*merge*") ou, alternativamente, de anexação. A indução de categorias sintáticas da gramática e a emergência de padrões recursivos nas regras se dá através do procedimento de generalização sintática.

4.5.1 A aquisição lexical

Assim como em Siskind (1996), para que a aquisição lexical seja possível, LEX prevê que cada sentido de uma palavra seja composto por três tabelas, a saber:

1. A tabela N, que mapeia um sentido para seus símbolos conceituais necessários;
2. A tabela P, que mapeia um sentido para seus símbolos conceituais possíveis;
3. A tabela D, que mapeia cada sentido para suas expressões conceituais possíveis.

Para ilustrar a operação do algoritmo, vou recorrer ao exemplo utilizado por Siskind (1996, p.57), acrescido das propriedades não previstas ali, mas relevantes para a presente modelagem (alguns nomes de símbolos também são diferentes, mas isto é irrelevante para a discussão). Vale ressaltar que o conjunto de heurísticas (Regras 1 a 5) apresentadas nesta seção é aplicado pelo algoritmo a cada uma das ASPs geradas conforme explicado anteriormente. Suponha que o aprendiz esteja a meio caminho em direção à aquisição lexical e que o léxico momentâneo contenha a seguinte informação:

	N	P
John	{**John**}	{**John**, **ball**}
took	{CAUSE}	{CAUSE, WANT, BECOME, **take**, PAST}
the	{}	{WANT, **arm**, DEF}
ball	{**ball**}	{**ball**, **take**}

Como mostra a tabela acima, o algoritmo ainda não convergiu para os conjuntos necessários dos itens lexicais envolvidos. Uma primeira adaptação feita sobre as heurísticas diz respeito à "Regra 1" (Siskind, 1996, p.57), cuja motivação ali era a de desfazer "incertezas referenciais". Por exemplo, diante de uma CTT como WANT(**John**, **ball**), o algoritmo po-

deria descartar esta CTT, visto que um dos símbolos necessários, CAUSE, não está contido nela (seria, portanto, uma CTT incorreta para o enunciado). No caso do IASMIM, entretanto, essa possibilidade não é modelada e apenas a CTT correta acompanha o enunciado em cada dado de entrada. A Regra 1 original é, portanto, irrelevante para este modelo.

Porém, um outro problema surge aqui. As representações semânticas consideradas por Siskind não continham os símbolos conceituais relativos a categorias funcionais adicionados por ocasião da presente simulação, tais como os atributos para lidar com tempo verbal, tipos oracionais, definitude, número, etc. Em função deste acréscimo, surge a seguinte questão: dado que um mesmo item lexical, como o verbo *construir*, por exemplo, pode ocorrer no contextos de orações declarativas e interrogativas (entre outras), o que é mais interessante? Que o aprendiz adquira duas entradas "polissêmicas" para o verbo, uma para o contexto declarativo e outra para o contexto interrogativo? Ou seria mais interessante que o aprendiz eventualmente convergisse para uma única entrada, subespecificada em relação ao tipo oracional, mantendo no léxico apenas os atributos de fato necessários?

A opção por incluir várias entradas parece menos desejável, se entendemos que esta situação se repetirá para a maior parte das palavras da língua, o que produziria uma significativa redundância no armazenamento lexical. Porém, a redundância em si não é um argumento conclusivo. Mais problemático é quando consideramos enunciados em que certos elementos estão omitidos, por estarem muito salientes no contexto. Nestes casos, definitivamente não seria possível fazer um pareamento estrito entre as ECs contribuídas pelos itens lexicais e CTT. Assim, embora o IASMIM não lide com tais enunciados em sua versão atual, considerei desejável adaptar o procedimento de aquisição lexical para que possa lidar com estas situações. Como desdobramento, isso resolveria o problema da redundância visto que verbos subespecificados em relação ao tipo oracional (ou outras propriedades, como tópico, por exemplo) não seriam impeditivo para o reconhecimento dos itens lexicais do enunciado. Assim, a Regra 1 original foi substituída aqui pela seguinte heurística:

Regra 1. *Ignore a ASP sendo verificada, caso nem todos os sentidos considerados contribuam símbolos necessários para a interpretação do enunciado e caso algum dos símbolos conceituais previstos em CTT não esteja presente nos símbolos possíveis dos itens do enunciado.*

A lógica dessa heurística é a seguinte: suponha que o aprendiz receba o dado de entrada "*John took the ball*", pareado com a CTT:

(29) DECL(PAST(CAUSE(**John**, BECOME(DEF(**ball**), **take**))))

Com base no conhecimento parcial do aprendiz, o algoritmo verifica que o símbolo DECL contido em CTT não está previsto como símbolo possível por nenhuma das entradas lexicais. Quando isso é detectado, o algoritmo verifica então se todos as entradas lexicais envolvidas contribuem com símbolos necessários. No caso, todos contribuem, exceto *the*, que ainda não tem nenhum símbolo em sua tabela N. Neste caso, não é possível saber se a falta de DECL se deve a uma aquisição ainda incompleta deste item (ou seja, dados futuros poderiam mudar esta entrada). Assim, a Regra 1 estipula que a ASP é inconsistente. Como o aprendiz não possui sentidos extras vinculados a estes itens lexicais, o algoritmo identifica o menor número de palavras que ele precisa atualizar, incluindo os símbolos do enunciado em análise em suas respectivas tabelas P e reprocessando o enunciado. Neste caso, basta incluir tais elementos em P(*the*). Com isso, o léxico parcial do aprendiz passa a exibir:

	N	P
John	{**John**}	{**John**, **ball**}
took	{CAUSE}	{CAUSE, WANT, BECOME, **take**, PAST}
the	{}	{WANT, **arm**, DEF, DECL, PAST, CAUSE, **John**, BECOME, **ball**, **take** }
ball	{**ball**}	{**ball**, **arm**}

Agora, a Regra 1 não aponta inconsistência, visto que P(*the*) garante que todos os símbolos de CTT estejam contemplados no conjunto de símbolos possíveis da ASP. Neste ponto, o algoritmo pode fazer a seguinte inferência: dado que a CTT não contém os símbolos WANT e **arm**, estes podem ser excluídos das tabelas P dos itens lexicais, atualizando o léxico para:

	N	P
John	{**John**}	{**John**, **ball**}
took	{CAUSE}	{CAUSE, BECOME, **take**, PAST}
the	{}	{DEF, DECL, PAST, CAUSE, **John**, BECOME, **ball**, **take** }
ball	{**ball**}	{**ball**}

Isso nos leva à segunda heurística proposta por Siskind (1996):

Regra 2. *Para cada palavra w do enunciado, remova de P(w) qualquer símbolo conceitual não previsto em CTT.*

A próxima inferência que o algoritmo pode fazer, diz respeito à relação entre os itens lexicais envolvidos: dado que todos os símbolos possíveis restantes estão contidos em CTT e dado que alguns símbolos possíveis são exclusivos a alguns itens lexicais, tais símbolos podem ser copiados para suas respectivas tabelas de símbolos necessários. Neste caso, o símbolo DEF é exclusivo à P(*the*) e pode ser copiado para N(*the*). Os demais símbolos ainda não podem ser atualizados aqui, em função da atualização feita por ocasião da Regra 1 sobre P(*the*).

	N	P
John	{**John**}	{**John**, **ball**}
took	{CAUSE}	{CAUSE, BECOME, **take**, PAST}
the	{DEF}	{DEF, DECL, PAST, CAUSE, **John**, BECOME, **ball**, **take** }

ball	{**ball**}	{**ball**}

Esta terceira heurística é descrita como segue:

Regra 3. *Para cada palavra w do enunciado, adicione à N(w) qualquer símbolo conceitual que apareça em CTT mas que esteja ausente de P(w′) para qualquer outra palavara w′ do enunciado.*

Agora, o algoritmo pode fazer outra inferência: dado que **ball** e **John** aparecem apenas uma vez em (29) e que ambas fazem parte, respectivamente, de N(*ball*) e N(*John*), estes símbolos não podem ser contribuídos (potencialmente) por outros itens lexicais. Com isso, estes símbolos podem ser removidos de P(w) para as demais palavras:

	N	P
John	{**John**}	{**John**}
took	{CAUSE}	{CAUSE, BECOME, **take**, PAST}
the	{DEF}	{DEF, DECL, PAST, CAUSE, BECOME, **take** }
ball	{**ball**}	{**ball**}

Este processo de inferência é descrito na quarta heurística:

Regra 4. *Para cada palavra w do enunciado, remova de P(w) qualquer símbolo conceitual que apareça apenas uma vez em CTT e que conste de N(w′) para alguma outra palavara w′ do enunciado.*

Após este passo, o algoritmo se aproximou um pouco mais dos conjuntos de símbolos conceituais próprios a cada item lexical. Porém, alguns outros enunciados seriam necessários para a convergência completa. Suponha, por exemplo, que o aprendiz receba o enunciado

"The kids", pareado com a CTT = DEF(**kids**). Neste caso, aplicando as Regras 1 a 4, o aprendiz convergiria para o seguinte léxico parcial:

	N	P
John	{**John**}	{**John**}
took	{CAUSE}	{CAUSE, BECOME, **take**, PAST}
the	{DEF}	{DEF}
ball	{**ball**}	{**ball**}

Note que *the* convergiu totalmente, de modo que N(*the*) = P(*the*). Com isso, sendo exposto novamente ao enunciado (29) e executando as Regras 1 a 4, o aprendiz finalmente convergiria para:

	N	P
John	{**John**}	{**John**}
took	{PAST, CAUSE, BECOME, take}	{PAST, CAUSE, BECOME, **take**}
the	{DEF}	{DEF}
ball	{**ball**}	{**ball**}

Agora, o aprendiz poderia passar para o que Siskind (1996) chama de "estágio dois" da aquisição lexical, em que após descobrir os conjuntos de símbolos conceituais dos itens lexicais, o algoritmo passa à tarefa de descobrir a EC respectiva a cada palavra. Este estágio consiste de outras duas regras, 5 e 6. A **Regra 5** tem a função de, para cada item lexical, calcular todas as combinações possíveis dos símbolos em N(w) que sejam consistentes com CTT. As combinações obtidas vão fazer parte de D(w). A **Regra 6** tem a função de pegar todas as expressões conceituais em D, para todas as palavras do enunciado, e testar suas combinações para ver se alguma delas unifica com CTT (segundo o autor, é uma forma generalizada de consistência de arco).

Aqui entra em cena mais uma adaptação feita de modo a adequar a proposta às particularidades da presente modelagem. Aparentemente, Siskind (1996) não impõe qualquer restrição a priori sobre as combinações possíveis envolvendo os símbolos conceituais. Portanto, o algoritmo poderia produzir expressões como:

(30) BECOME(**John**, CAUSE(DECL(**ball**), PAST(**take**)))

Expressões como essa, embora logicamente possíveis, são completamente implausíveis do ponto de vista linguístico. Portanto, não faz sentido considerá-las nesta tarefa de aquisição lexical, razão pela qual o algoritmo foi adaptado da seguinte maneira: quando um novo item lexical (ou um novo sentido) é postulado, o algoritmo calcula as sub-expressões possíveis a partir de CTT, segundo certas condições. Uma primeira condição é a de que a função estrutural F é sempre a base de uma sub-expressão. Uma segunda condição é que no cálculo de uma sub-expressão, o algoritmo não pode "saltar" um nível. Por exemplo, tomando a CTT em (29), alguns exemplos de sub-expressões possíveis são:

(31) DECL(PAST(CAUSE(x, BECOME(x, **take**))))
BECOME(x, **take**)
CAUSE(x, BECOME(x, **take**))
PAST(CAUSE(x, BECOME(x, **take**)))
John
DEF(**ball**)
DEF(x)
etc...

Algumas possibilidades lógicas, porém, seriam restringidas pelas condições acima mencionadas:

(32) DECL(CAUSE(x, BECOME(x, **take**)))
CAUSE(x, **take**)

etc...

Com isso, foi possível reduzir significativamente o número de cálculos combinatórios a serem efetuados, ao mesmo tempo em que o conjunto de sub-expressões obtido contém expressões que são, por definição, consistentes com CTT. Assim, a quinta heurística pode ser definida como segue:

> **Regra 5**. *Para cada palavra w do enunciado, se w convergiu para seu conjunto de símbolos conceituais, N(w) = P(w), remova de D(w) qualquer EC que não envolva exatamente os símbolos conceituais em N(w); caso não tenha convergido, remova de D(w) qualquer EC que inclua um símbolo ausente de P(w).*

Seguindo com o processamento do enunciado (29), após executar a Regra 5, o algoritmo converge para as expressões conceituais próprias a cada palavra. A Regra 6, que em Siskind (1996, p.61) tem a função de verificar se a composição das expressões conceituais das palavras do enunciado é consistente com CTT, deixa de ser relevante aqui, quando o algoritmo passa a permitir que os itens contribuam com menos informação do que aquela contida em CTT. Nestes casos, não se pode garantir que alguma combinação das expressões seja consistente com CTT. Portanto, para os fins desta modelagem, apenas as heurísticas de 1 a 5 são utilizadas.

Após executar as regras de 1 a 5 sobre todas as ASPs do enunciado, três situações podem ocorrer: (i) o algoritmo converge para uma única ASP consistente com CTT; (ii) o algoritmo converge para um conjunto de ASPs consistentes com CTT; e (iii) nenhuma ASP converge e nem é consistente com CTT. No primeiro caso, o algoritmo incrementa os fatores de confiança dos sentidos envolvidos na ASP e salva as atualizações conjecturadas no léxico. No segundo caso, o algoritmo identifica a ASP com maior fator de confiança agregado, incrementa os fatores de cada sentido e salva estas atualizações no léxico.

Por fim, no terceiro caso, o algoritmo identifica o menor número de itens lexicais para os quais, ao incluir uma nova entrada, o algoritmo consiga convergir para alguma ASP. Sendo possível, as entradas são incluídas e o enunciado é reprocessado. Caso contrário, o dado é descartado. Nos dois primeiros casos, porém, tendo sido identificada a ASP mais adequada, as entradas lexicais respectivas são enviadas para o analisador sintático, para que a análise e aquisição sintática possa ter andamento.

4.5.2 Blocos de construção da sintaxe: o resultado da aquisição lexical

A aquisição lexical, como concebida, vai produzir o conjunto de elementos com os quais a sintaxe vai trabalhar. Aqui vale ressaltar as principais características destes "blocos de construção", para que se possa entender o processamento e aquisição sintática no modelo. Tomemos novamente o conjunto de itens lexicais exemplificados acima, após convergirem para seus sentidos corretos (é omitida a coluna P):

	N	D
John	{**John**}	**John**
took	{PAST, CAUSE, BECOME, **take**}	PAST(CAUSE(x, BECOME(y, **take**)))
the	{DEF}	DEF(x)
ball	{**ball**}	**ball**

O bloco de construção mais simples consiste dos itens lexicais *saturados*, isto é, dos itens que não apresentam funções estruturais em aberto. Do conjunto acima, **John** e **ball** são deste tipo, pois apresentam respectivamente as ECs em (33a) e (33b), a seguir:

(33) a.
$$\begin{bmatrix} \text{ITEM} & \textbf{john} \\ \text{EC} & \begin{bmatrix} \text{F} & \begin{bmatrix} \text{DEFINITE} & + \end{bmatrix} \\ \text{CF} & \begin{bmatrix} \text{F} & \begin{bmatrix} \text{ANIMACY} & - \\ \text{CONCEPT} & \textit{john} \\ \text{PERSON} & 3 \\ \text{QUANTIZABLE} & + \\ \text{WH} & - \\ \text{PLURAL} & - \end{bmatrix} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

b.
$$\begin{bmatrix} \text{ITEM} & \textbf{ball} \\ \text{EC} & \begin{bmatrix} \text{F} & \begin{bmatrix} \text{ANIMACY} & - \\ \text{CONCEPT} & \textit{ball} \\ \text{PERSON} & 3 \\ \text{QUANTIZABLE} & + \\ \text{WH} & - \\ \text{PLURAL} & - \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Alguns itens lexicais, entretanto, serão *insaturados*, quando convergirem para uma EC contendo alguma função estrutural em aberto. Durante a aquisição lexical, apenas as funções 'cf' e 'a' são identificadas (como estando em aberto) como parte da entrada lexical de um item. A função 'm' (para modificadores), por ser estritamente opcional, aparece de modo contingente na CTT de um dado enunciado e, portanto, não há a possibilidade de que algum sentido (de um item lexical qualquer) apresente esta função em aberto em sua EC respectiva armazenada no léxico. Só quando chega à sintaxe, é que a relação do adjunto com o elemento modificado será capturada. Quando um item é insaturado, a sintaxe o considera "dependente" e as dependências são aqui consideradas como a realização da noção de "(atributo) seletor". Dos itens acima, **the** e **took** são dependentes. Vejamos suas ECs respectivas:

(34) a.
$$\begin{bmatrix} \text{ITEM} & \textbf{the} \\ \text{EC} & \begin{bmatrix} \text{F} & \begin{bmatrix} \text{DEFINITE} & + \end{bmatrix} \\ \text{CF} & [] \end{bmatrix} \end{bmatrix}$$

b. $$\begin{bmatrix} \text{ITEM} & \textbf{took} & & & & & & \\ \text{EC} & \begin{bmatrix} \text{F} & \begin{bmatrix} \text{FINITE} & + \\ \text{REALIS} & + \end{bmatrix} & & & & & \\ \text{CF} & \begin{bmatrix} \text{F} & [\text{CONTROL} \quad +] & & & & \\ \text{CF} & \begin{bmatrix} \text{F} & [\text{CHANGE} \quad +] & & \\ \text{CF} & \begin{bmatrix} \text{F} & \begin{bmatrix} \text{CONCEPT} & \textbf{take} \\ \text{QUANTIZABLE} & + \\ \text{CULMINATION} & - \\ \text{PUNCTUAL} & + \end{bmatrix} \end{bmatrix} \\ \text{A} & [] \end{bmatrix} \\ \text{A} & [] \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Em (34a), temos a função 'cf' em aberto no nível máximo da EC. Portanto, o item lexical *the* é dependente de um argumento que sature esta função estrutural. O item *took*, por sua vez, apresenta duas funções 'a' em aberto aninhadas em sub-níveis da EC. À sequência de atributos percorridos para acessar um dado seletor, vamos chamar de "caminho". No processamento sintático, estes seletores vão determinar o estatuto de núcleo para tais itens lexicais e vão determinar também a ordem de composição da estrutura sintática, como é explicado mais abaixo. Assim, os itens em questão apresentam os seguintes seletores e caminhos para eles:

(35) **the**, seletor: 'cf'

took, seletores: 'cf:a' e 'cf:cf:a'

Uma categoria sintática é determinada pelo conjunto de atributos lexicais que contém e pelo conjunto de seletores (que será vazio para itens saturados). Veremos que apenas um item saturado (seja primitivamente ou derivacionalmente) pode satisfazer o seletor de outro item. Adjuntos também podem saturar seletores, porém tais seletores só podem ser identificadas com recurso à CTT, durante o processamento sintático do enunciado. Neste ponto, podemos retomar os conceitos de *set-Merge* e *pair-Merge* (cf. Chomsky, 2004): set-Merge se aplica às combinações de termos para satisfação de seletores 'cf' e 'a' (inerentes aos predicados), enquanto pair-Merge se aplica na satisfação de seletores 'm' (contingentes). O conhecimento adquirido pelo aprendiz reflete esta distinção.

4.5.3 A aquisição sintática

Finalizado o processamento lexical do enunciado, a lista de palavras (quando reconhecidas) é enviada para o analisador, para que este faça a análise sintática do enunciado e, eventualmente, dispare procedimentos de aquisição. Neste ponto, o analisador recebe itens lexicais caracterizados como pares (w_i, s_i) e os converte em objetos sintáticos, momento no qual são designados rótulos categoriais para eles. Itens inéditos irão receber rótulos novos, enquanto itens já processados na sintaxe irão receber os anteriormente designados. Feito isto, tem início o fluxo de processamento sintático, em que os itens são processados, de modo a construir um objeto final a partir de sua combinação (a árvore sintática). Assim como em outros modelos de aquisição (Berwick, 1985, Gaylard, 1995, entre outros), a aprendizagem ocorre a partir de erros de análise, considerados sempre como indício de insuficiência da gramática momentânea do aprendiz.

No reconhecimento e aquisição lexical, a série de heurísticas (Regras 1 a 5) é executada para cada ASP envolvendo os itens lexicais em questão. Aqui, de modo semelhante, sobre cada estado do analisador para o qual a gramática momentânea é insuficiente são executadas heurísticas (Passos 1 a 4, abaixo) cujo objetivo é criar uma nova regra gramatical que permita ao analisador prosseguir com a análise e que, eventualmente, leve a alguma generalização na gramática. Vale ressaltar que a análise de um enunciado é composicional, embora não estritamente ascendente ("*bottom-up*"). Em outras palavras, os itens são combinados de modo a atender os seletores mais locais dos itens, de modo que a árvore final resultante reflita a localidade das relações. Usando termos da teoria gramatical, seria como construir a estrutura profunda[13] do enunciado (lembrando que a estrutura superficial, neste modelo, não é parte do conhecimento sintático a ser adquirido).

[13] Com o advento do minimalismo e a mudança de um paradigma representacional para um derivacional, a noção de estrutura profunda perdeu estatuto teórico. Porém, as relações capturadas pela estrutura profunda (especialmente as argumentais) continuam presentes na derivação minimalista, mais exatamente sendo estabelecidas nos primeiros passos da derivação. O análise no modelo reconstituiria, portanto, estes primeiros passos.

Para obter esta composicionalidade, faz parte da pré-disposição do aprendiz o conhecimento sobre a ordem intrínseca entre as funções estruturais da expressão conceitual, a saber, 'cf' > 'a' > 'm' (do mais interno para o mais externo), ordem esta que guiará a composição sintática da estrutura. Para compreendermos estes aspectos e também o fluxo de aquisição sintática, vamos tomar um enunciado simples, mas que permite ilustrar as heurísticas envolvidas, além do caráter da análise. Suponha que o aprendiz receba o enunciado "*the red car*", com parte de sua gramática momentânea exibindo o seguinte estado (leva-se um tempo para se acostumar com a forma de apresentação das informações):

```
N0009 → the, [cf=[], f=[+definite]]
N0019 → car, [f=[animacy=@animacy, concept=@concept, person='3', +quantizable,
               -wh, -plural]]
N0011 → [cf=[f=[-animacy, concept='car', person='3', +quantizable, -wh,
               -plural]], f=[+definite]]

Regra A0002
CYC: None, ACT: N0011, (head)
  (1) N0009
  (2) [f=[-animacy, concept=car, person=3, +quantizable, -wh, -plural]]
  (3) None

Regra A0003
CYC: None, ACT: N0011, (comp)
  (1) N0019
  (2) None
  (3) None

Regra M0002
CYC: None, ACT: None, (seletor: ('cf',))
Sense: [cf=[f=[animacy=@animacy, concept=@concept, person='3',
        quantizable=@quantizable, -wh, plural=@plural]],
        f=[+definite]]
  (1) N0009
  (2) [f=[animacy=@animacy, concept=@concept, person=3,
       quantizable=@quantizable, -wh, plural=@plural]]
  (3) None
```

Acima, vemos que a gramática já dispõe de rótulos sintáticos para *the* (N0009) e *car* (N0019, já bastante subespecificado), além de um rótulo para o DP (N0011), ainda específico

para *the cat*. Possui também três regras relevantes para este exemplo, as regras: M0002, para criação do constituinte e já bastante subespecificada; A0002, para anexar o determinante como núcleo do sintagma; e A0003, para anexar o item nominal como complemento. A palavra *red* foi adquirida na aquisição lexical e chega pela primeira vez para análise sintática. O analisador então converte os itens para objetos sintáticos, criando um novo rótulo para a palavra *red* e apresentando o contexto de análise inicial a seguir:

```
Criou objeto sintático N0009/<the.1> : [cf=[], f=[+definite]]
  Dependência(s): [('cf',)]

Criou objeto sintático N0078/<red.1> : [f=[concept='red', -wh]]

Criou objeto sintático N0019/<car.1> : [f=[-animacy, concept='car',
  person='3', +quantizable, -wh, -plural]]

Contexto do parser:
  CYC: None, ACT: None
  Buffer: the | red | car
```

O analisador busca por uma regra conhecida compatível mas não encontra. No caso de M0002, ela espera um item nominal na segunda célula e *None* na terceira. Já as regras de anexação demandam um constituinte ativo (ACT) em formação, o que não é o caso. Portanto, o procedimento dispara os procedimentos de aquisição, que visam identificar, *nessa ordem*, se é possível *anexar* o símbolo de entrada em algum constituinte em formação ou se é possível conjecturar um novo constituinte.

> **Passo 1**. *Verifique se há um nó ativo (ACT) na pilha de constituintes. Se não houver um nó ativo, a rotina segue para o próximo passo. Caso contrário, verifique se CTT licencia o item na primeira célula da área temporária como um de seus termos (núcleo, complemento ou adjunto). Se for compatível, crie uma regra do tipo ANEXE, execute e siga para o Passo 4, em que o algoritmo tenta generalizar a regra. Se não houver compatibilidade, passe para o Passo 2.*

A compatibilidade existe quando a EC do termo esperado unifica com a EC de um dos termos do objeto sintático em construção e, quando em fase de aquisição, se CTT licencia o resultado da unificação. Como não é o caso neste exemplo, visto que ACT é vazio, o procedimento segue para o Passo 2, em que tenta criar uma regra do tipo COMBINE:

> **Passo 2**. *Verifique se é o caso de criar uma regra do tipo COMBINE, para a formação de um constituinte sintático. Duas situações são possíveis:*
>
> > **Passo 2A**. *Verifique (com recurso à CTT) se há adjuntos a serem processados para este objeto sintático (aninhados em níveis mais baixos que eventuais seletores). Se houver, crie uma regra do tipo COMBINE relativa à adjunção, execute e vá para o Passo 4. Caso contrário, siga para 2B.*
> >
> > **Passo 2B**. *Se o objeto contiver algum seletor, significa que é um núcleo sintático. Com recurso à CTT, crie uma regra do tipo COMBINE relativa ao complemento esperado. A regra é executada e o procedimento segue para o Passo 4. Caso o objeto não seja um núcleo, siga para o Passo 3.*

No caso das funções estruturais 'cf' e 'a', os respectivos seletores são identificáveis diretamente no símbolo de entrada, visto que são determinadas pelo processo de aquisição lexical. Já para a função estrutural 'm', a necessidade de adjuntos só pode ser determinada com o auxílio de CTT, visto que modificadores são assumidos como estritamente opcionais. No caso do determinante, o passo 2B identifica que 'cf' é um seletor e portanto cria a regra para criação de um constituinte. Antes de prosseguir com o exemplo, entretanto, vale considerar o papel da terceira heurística.

Suponha que não fosse possível anexar o símbolo de entrada e nem conjecturar um novo constituinte do qual fosse núcleo. Resta ao procedimento de aquisição tentar uma última

heurística, a saber, a que permite identificar possíveis núcleos "abstratos" na estrutura, isto é, núcleos que não correspondem a nenhuma palavra do enunciado (e, potencialmente, da língua). Não é o caso do presente enunciado, mas seria o caso, por exemplo, da opcionalidade envolvendo o complementizador *that*, em construções como "*the cat that John saw*" e "*the cat John saw*". Nos dois casos, esta modelagem assume que uma CTT idêntica está em jogo e, portanto, o símbolo conceitual relativo à "*that*" estaria presente em ambas as CTTs.

Porém, se a palavra está omitida, o analisador não tem condições de adjungir a oração relativa ao núcleo nominal, visto que este espera um constituinte cujo núcleo seja a categoria correspondente ao complementizador. Assim, para que a análise possa proseguir, o procedimento de aquisição conjectura um núcleo abstrato, criando uma regra do tipo COMBINE. Como esta é um heurística com grandes chances de sobregerar, o procedimento faz uso da antecipação e também do contexto esquerdo, de modo a ter alguma segurança neste passo. Ou seja, de certo modo o aprendiz "resistiria" a conjecturar tais categorias. Esta heurística pode ser descrita como a seguir:

> **Passo 3**. *Se houver um constituinte em formação em ACT, recorra à CTT para identificar a categoria do termo esperado e se o símbolo de entrada, por sua vez, é termo da categoria identificada. Se for, verifique antes se o símbolo de entrada é termo de CYC ou se a janela de antecipação contém objetos que satisfazem ACT. Se qualquer um dos casos ocorrer, não prossiga com a criação da categoria abstrata. Caso contrário, crie a regra correspondente, marque a categoria como sendo abstrata, execute a regra e siga para o Passo 4.*

Se nenhum dos Passos 1 a 3 conseguir criar uma regra, significa que não há nada a fazer, pelo menos não para esta posição do buffer. No caso do exemplo acima, o Passo 2B cria uma regra para formação de um sintagma determinante, como mostra o quadro a seguir:

```
Criou regra M0051
CYC: None, ACT: None, (seletor: 'cf')
Sense: [cf=[f=[-animacy, concept='car', person='3', +quantizable,
        -wh, -plural]], f=[+definite]]
  (1) N0009
  (2) [f=[concept='red', -wh]]
  (3) [f=[-animacy, concept='car', person='3', +quantizable,
       -wh, -plural]]

Execução da regra M0051

Criou objeto sintático N0011 : [cf=[f=[-animacy, concept='car', person='3',
  +quantizable, -wh, -plural]], f=[+definite]]
```

Quando uma nova regra dos tipos ANEXE ou COMBINE é criada e executada, ela é em seguida enviada para um procedimento de generalização que tem como objetivo identificar regras já conhecidas que tenham contextos semelhantes o suficiente para permitir a generalização. O procedimento pode ser descrito assim:

> **Passo 4**. *Para as regras conhecidas de mesma ação e posição no buffer, se houver uma regra conhecida com mesmo símbolo de entrada, verifique se o contexto esquerdo ou (exclusivamente) o direito das regras são idênticos. Se for, crie uma nova regra mais geral a partir das duas, descartando-as. Se as regras não tiverem o mesmo símbolo de entrada, mas ambos os contextos esquerdo e direito forem idênticos, crie uma classe de equivalência para os símbolos de entrada, descartando a nova regra. Em qualquer outro caso, não generalize e mantenha as regras como estão.*

Na generalização, os nós CYC, ACT e as células de antecipação tem suas ECs atualizadas para conter apenas valores em comum, com os conflitantes se tornando subespecificados. Caso os pares comparados envolvam um elemento *None* e outro cujo valor é uma matriz, a generalização leva ao símbolo '*' (coringa), indicando que qualquer elemento é válido, inclusive nenhum. As condições para generalização são:

(i) As regras devem ser de mesmo tipo, fazer referência à mesma posição do buffer e, se forem do tipo anexe, fazerem referência ao mesmo termo (núcleo, complemento ou adjunto).

(ii) Os nós ACTs das regras sendo avaliadas devem ser ambos nulos ou ambos devem apresentar a mesma categoria de símbolo conceitual como predicado principal (entrada 'f', na representação semântica), isto é, os atributos do predicado devem ser idênticos, embora seus valores possam diferir. Quando atendem a esta condição, os nós ACT são considerados "equivalentes". Essa condição confere um caráter lexical à aquisição sintática no modelo, visto que o aprendiz pressupõe que a equivalência (lexical) dos nós é indício de uma mesma categoria sintática em jogo.

(iii) Se os símbolos de entrada forem idênticos, há três situações: (i) ambos os contextos são iguais; (ii) apenas um dos contextos é igual (direito ou esquerdo); (iii) nenhum contexto é igual. No caso de (i), significa que as regras são iguais e não há nada a fazer. No caso de (iii), significa que há muitas diferenças e não é o caso de generalizar. No caso de (ii), significa que as regras podem ser generalizadas. Se o lado esquerdo for idêntico, generaliza o lado direito para os atributos em comum ou vice-versa.

(iv) Se os símbolos de entrada forem distintos, mas os contextos direito e esquerdo das regras forem ambos idênticos, significa que os símbolos são equivalentes. Este procedimento induz uma classe de equivalência envolvendo os símbolos em questão, que poderá ser ampliada para abarcar outros mais, na medida em que a regra entrar em outras generalizações.

No Passo 4, o procedimento identifica que a regra M0052 criada acima pode ser generalizada com a regra M0002, visto que CYC, ACT e o símbolo de entrada (célula 1) são idênticos. Portanto, uma regra mais geral é criada, agora apresentando uma especificação mais abstrata que a anterior (que abarca tanto sequências artigo + nome, quanto sequências artigo + adjetivo + nome):

```
Regra generalizada M0002
CYC: None, ACT: None, (seletor: 'cf')
Sense: [cf=[f=[animacy=@animacy, concept=@concept, person='3',
        quantizable=@quantizable, -wh, plural=@plural]],
        f=[+definite]]
Células:
  (1) N0009
  (2) [f=@f]
  (3) *
```

Neste ponto, o analisador executou a regra e criou um constituinte em ACT, produzindo um novo contexto sintático:

```
Contexto do parser (célula 0)
  CYC: None, ACT: N0011
  Células:
    N0009
    N0078
    N0019
```

O fluxo completo é então repetido para o novo contexto. Neste caso, ainda não haverá regra conhecida para anexar N0009 à N0011. Assim, os procedimentos de aquisição são disparados novamente e o Passo 1 cria uma regra ANEXE. A regra é executada, o objeto N0009 é retirado do *buffer* e anexado à N0011. A nova regra, A0085, pode ser generalizada com A0002, como mostra o fluxo a seguir:

```
Regra A0002
CYC: None, ACT: N0011, (head)
  (1) N0009
  (2) [f=[-animacy, concept=car, person=3, +quantizable, -wh, -plural]]
  (3) None

Regra A0085
CYC: None, ACT: N0011, (head)
  (1) N0009
```

```
  (2) [f=[concept='red', -wh]]
  (3) [f=[-animacy, concept='car', person='3', +quantizable,
       -wh, -plural]]

Generalizou: A0085 → A0002

Regra generalizada A0002
CYC: None, ACT: N0011, (head)
  (1) N0009
  (2) [f=@f]
  (3) *

Contexto do parser (célula 0)
  CYC: None, ACT: N0011
  Células:
    N0078
    N0019
    None
```

O novo contexto do analisador agora tem o objeto sintático relativo à *red* como símbolo de entrada. O fluxo é repetido, mas dessa vez nada pode ser feito, pois *red* não é núcleo, não possui adjuntos licenciados por CTT e nem pode disparar um constituinte abstrato. Portanto, o analisador segue para a próxima célula, e o contexto passa a ser:

```
Contexto do parser (célula 1)
  CYC: None, ACT: N0011
  Células:
    N0019
    None
    None
```

Agora temos N0019 (*car*) como símbolo de entrada. O fluxo é repetido. O Passo 1 não consegue criar uma regra ANEXE para anexar N0019 ao sintagma determinante, pois CTT não licencia a combinação, visto que falta o adjunto. Assim, o Passo 2A vai disparar a criação de uma regra COMBINE, para formação do constituinte de adjunção envolvendo *red car*. Uma nova regra é criada, executada e o analisador volta à primeira célula, produzindo

um novo contexto (neste caso, como é o primeiro adjetivo visto pelo aprendiz, não há regras conhecidas para generalizar):

```
Nova regra M0052
CYC: None, ACT: N0011, (seletor: 'cf:m')
Sense: [f=[-animacy, concept='car', person='3', +quantizable,
        -wh, -plural], m=[[f=[concept='red', -wh]]]]
  (1) N0019
  (2) None
  (3) None

Execução da regra M0052

Criou objeto sintático N0079 : [f=[-animacy, concept='car', person='3',
  +quantizable, -wh, -plural], m=[[f=[concept='red', -wh]]]]

Contexto do parser (célula 0)
  CYC: N0011, ACT: N0079
  Células:
    N0078
    N0019
    None
```

Note que agora o constituinte relativo ao determinante se tornou CYC, enquanto o constituinte relativo à adjunção está ativo (ACT). O símbolo de entrada voltou a ser *red* e agora o Passo 1 pode conjecturar uma regra do tipo ANEXE. A regra é executada e N0078 é retirado do *buffer* e anexado à N0079, como adjunto. O contexto passa a apresentar N0019 como símbolo de entrada. Novamente, o Passo 1 cria uma regra ANEXE, retirando N0019 e anexando-o à N0079, como núcleo da adjunção. Neste ponto, o analisador percebe que N0079 está completo, devolvendo-o ao *buffer*:

```
Nova regra A0086
CYC: N0011, ACT: N0080, (adjuncts)
  (1) N0078
  (2) [f=[-animacy, concept=car, person=3, +quantizable,
       -wh, -plural]]
  (3) None
```

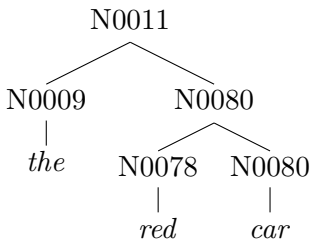
```
Execução da regra A0086
Anexou N0078/<red.1> como 'adjuncts' de N0079

Contexto do parser (célula 0)
  CYC: N0011, ACT: N0089
  Células:
    N0019
    None
    None

Criou regra A0087
CYC: N0011, ACT: N0079, (adjuncts)
  (1) N0019
  (2) None
  (3) None

Execução da regra A0087
Anexou N0019/<car.1> como 'head' de N0079
Completou objeto sintático N0079

Contexto do parser (célula 0)
  CYC: None, ACT: N0011
  Células:
    N0079
    None
    None
```

Como é uma estrutura de adjunção inédita, ainda não há regra conhecida para anexá-la à N0011. Assim, o Passo 1 conjectura uma nova regra ANEXE, que é executada, retirando N0079 do *buffer* e anexando à N0011. Porém, a rotina de generalização verifica que há uma regra existente (A0003) muito semelhante à nova regra e generaliza, como vemos abaixo. O interessante é que esta generalização captura o fato de que o sintagma nominal – pelo menos em relação à formação de um DP – se comporta sintaticamente do mesmo modo, à despeito de conter uma estrutura de adjunção. Daí, que o mesmo rótulo é atribuído (N0080) tanto à *car*, quanto à *red car*, indicando que eventualmente um adjunto pode estar presente (propriedade 'm=@m' na categoria N0080). Vemos, assim, que o procedimento é capaz de induzir este conhecimento gramatical, sem a necessidade de uma codificação explícita dessa propriedade.

```
Criou regra A0088
CYC: None, ACT: N0011, (comp)
  (1) N0079
  (2) None
  (3) None

Execução da regra A0088
Anexou N0079 como 'comp' de N0011
Completou objeto sintático N0011

Generalizou (equivalência): A0088 → A0003

Regra generalizada A0003
CYC: None, ACT: N0011, (comp)
  (1) N0080
  (2) None
  (3) None

N0079 → [f=[-animacy, concept='car', person='3', +quantizable,
          -wh, -plural], m=[[f=[concept='red', -wh]]]]
N0080 → [f=[animacy=@animacy, concept=@concept, person='3',
          +quantizable, -wh, -plural], m=@m]
         [Equivalência: 'N0079', 'N0019']

Contexto do parser (célula 0)
  CYC: None, ACT: None
  Células:
    N0011
    None
    None
```

O analisador identifica então que N0011 está completo, devolvendo-o ao *buffer*. Agora, há apenas este objeto em análise e sua EC respectiva unifica com CTT. Portanto, a análise sintática teve sucesso e as regras conjecturadas são incluídas definitivamente na gramática. Caso a análise terminasse num estado de insucesso (sem completar objetos ou com mais de um objeto no *buffer* sem possibilidade de combinação), as regras conjecturadas no processo seriam descartadas para fins de aprendizagem. Em outras palavras, o aprendiz só aproveita um enunciado, quando obtém uma análise completa do mesmo. A árvore sintática produzida pelo analisador para o exemplo é apresentada abaixo:

(36)



4.5.4 Outras questões envolvendo aprendizagem no modelo

Ordenamento das heurísticas

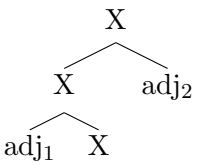
Assim como ocorre em Berwick (1985) e Gaylard (1995), as heurísticas de aquisição sintática também são ordenadas no modelo, no caso, segundo a sequência *anexação, criação de constituinte* e *criação de constituinte abstrato*. Esta ordenação advém do princípio do subconjunto, garantindo que o aprendiz, a cada passo, amplie a gramática o mínimo possível: a anexação não postula novos sintagmas; se for preciso postular um novo sintagma, a combinação pelo contexto postula apenas sintagmas cujos termos correspondem a itens presentes no enunciado; por fim, a criação de um sintagma abstrato permite que certos elementos (núcleos) sem realização lexical possam ser conjecturados. A ordenação das heurísticas visa evitar que o procedimento de aquisição sobregere na criação de regras.

Adjuntos

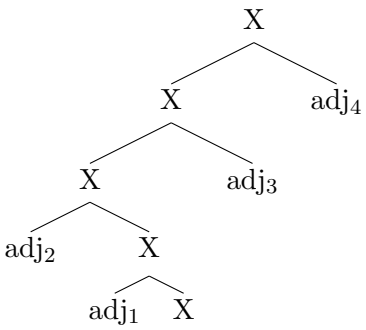
Como mencionado no início deste capítulo, há apenas um aspecto da aquisição sintática em que a ordem linear tem influência: este é o da composição das estruturas de adjunção. Como explicado na Seção 4.2.2, um dado símbolo conceitual pode conter zero, um ou mais modificadores. É assumido aqui que a lista de modificadores é não-ordenada na expressão conceitual, portanto, não há como a sintaxe ordená-los na composição, de modo intrínseco e independente da ordem linear. Assim, esta é que vai determinar a ordem hierárquica dos adjuntos (se mais de um) de um elemento. No modelo, o núcleo da adjunção funciona como um "pivô", de tal maneira que os adjuntos que o precedem no enunciado são processados antes

daqueles que o sucedem e, para cada conjunto, os mais próximos ao núcleo são processados primeiro. Alguns exemplos:

(37) a. ... adj$_1$ X adj$_2$ ...



b. ... adj$_2$ adj$_1$ X adj$_3$ adj$_4$ ...



4.6 Sumário

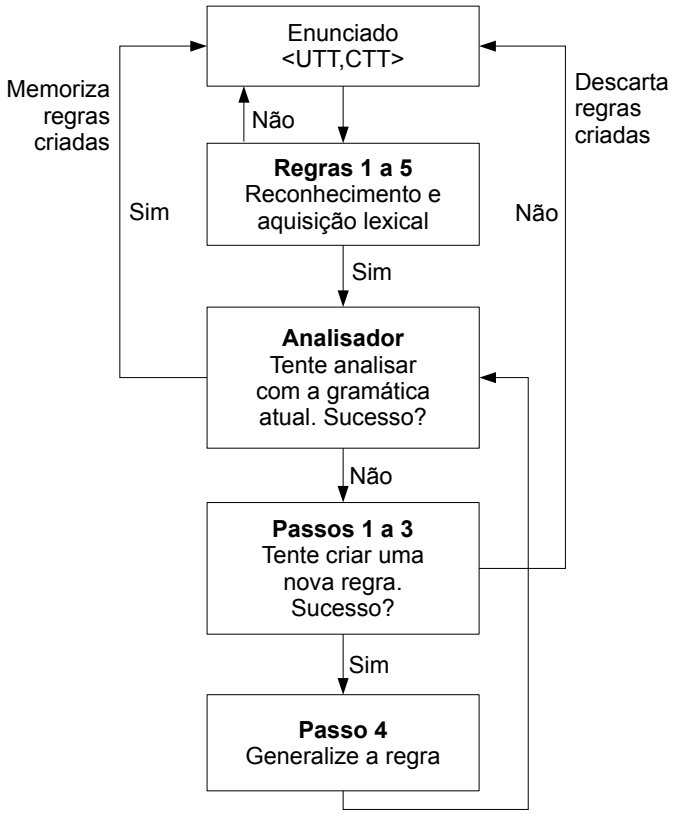


Figura 4.6: O fluxo de análise e aquisição sintática no IASMIM

A Figura 4.6, acima, resume o fluxo geral de processamento e aquisição no modelo.

Neste capítulo foram apresentadas as principais características do IASMIM, tanto no que diz respeito aos pressupostos teóricos que guiaram sua concepção, quanto também às estruturas de dados envolvidas, às propriedades dos processadores lexical e sintático e, finalmente, dos procedimentos de aquisição lexical e sintática. O modelo integra os processos de aquisição lexical e de aquisição sintática, de modo que sua interação em conjunto com as propriedades dos dados de entrada determinam a natureza do conhecimento adquirido. No próximo capítulo, são apresentados e discutidos os resultados obtidos pelos algoritmos, quando submetidos a corpora de entrada do inglês e do português brasileiro.

5

Resultados e discussão

5.1 Propriedades do corpus de entrada

Para permitir uma avaliação concreta do IASMIM, um corpus submetido a ele deveria apresentar duas características principais: abrangência gramatical, dentro dos limites da representação assumida e daquilo que se considera plausível para o período de aquisição simulado, e quantidade de dados suficiente para exaurir (ou quase) as possibilidades combinatórias envolvendo os itens lexicais presentes no corpus. Além disso, seria interessante que as frequências das construções no corpus criado refletissem na medida do possível as frequências verificadas na fala dirigida à criança, nos corpora de aquisição, pois esta similitude conferiria maior plausibilidade à simulação.

5.1.1 Tipos e frequência dos enunciados

Dois trabalhos serviram como fonte para as frequências relativas e também para a definição dos tipos de construções a serem incluídas no corpus. Hoff-Ginsberg (1986) apresenta um estudo sobre os efeitos de propriedades funcionais (objetivos comunicativos) e estruturais (tipos de construções) da fala das mães sobre o desenvolvimento da sintaxe na fala dos filhos. Os dados considerados consistem de transcrições de conversação, coletadas em intervalos de

2 meses por um período de 6 meses, para 22 crianças de dois anos e suas mães. Deixando de lado detalhes do estudo (que indicam alguma correlação, diga-se de passagem), trago aqui o levantamento feito pela autora sobre as propriedades dos enunciados produzidos pelas mães, do qual destaco alguns dos achados, apresentados na Tabela 5.1.

Propriedades estruturais da fala das mães

Medida	*M*
Medidas de complexidade sintática	
MLU	4.47
VP/enunciado	.95
NP/enunciado	1.60
Auxiliares/VP	.29
Palavras/NP	1.33
Frequências de formas sentenciais (% sobre todos os enunciados)	
Declarativas	25
Questões sim/não	15
Questões-Qu	17
Imperativas	8
Interjeições	17

Tabela 5.1: Levantamento sobre propriedades da fala dirigida à criança em Hoff-Ginsberg (1986).

Vê-se, pela tabela, que os enunciados não são muito longos (MLU igual à 4.47 – lembrando que se trata de número médio de *morfemas*, não de palavras). Além disso, praticamente todo enunciado dirigido à criança contém um verbo, um ou mais NPs, e um auxiliar aparece em média a cada três enunciados. Quanto à frequência por tipos de construções, vemos que questões (sim/não e Qu) compõem a maior parte dos dados de entrada, com as declarativas vindo a seguir e finalmente imperativas (desconsidero interjeições, pois não são, em geral, sintaticamente relevantes). Estes dados já seriam suficientes para obter uma ideia razoável das frequências e tipos de construções que deveriam compor o corpus.

No entanto, seria interessante identificar pelo menos mais uma fonte de dados, a fim de diminuir o efeito de qualquer viés que poderia estar incidindo sobre as frequências em Hoff-Ginsberg (1986). Cameron-Faulkner et al. (2003), num estudo sobre a fala dirigida à criança

a partir de amostras de doze mães falantes de inglês, apresentam um levantamento mais detalhado quanto ao tipo das construções envolvidas. O interessante é que o levantamento apresenta as médias, a variação mínima/máxima entre as mães e um contra-ponto com medidas apresentadas em Wells (1981, *apud* Cameron-Faulkner et al., 2003), como podemos ver na Tabela 5.2, abaixo:

Tipo de construção	O presente estudo		Wells (1981)	
	Proporção média	Tokens	Proporção média	Tokens
Fragmentos (variação)	.20 (.13–.32)	3351	.27 (.21–.35)	92
Uma palavra	.07		.08	
Multi-palavra	.14		.19	
Questões (variação)	.32 (.20–.42)	5455	.21	74
Qu-	.16		.08	
Sim/não	.15		.13	
Imperativas (variação)	.09 (.05–.14)	1597	.14 (.06–.24)	48
Cópulas (variação)	.15 (.08–.20)	2502	.15 (.10–.19)	51
Sujeito–predicado (variação)	.18 (.14–.26)	2970	.18	64
Transitivas	.10		.09	
Intransitivas	.03		.02	
Outras	.05		.07	
Complexas (variação)	.06 (.03–.09)	1028	.05 (.03–.07)	18

Tabela 5.2: Levantamento sobre propriedades da fala dirigida à criança em Cameron-Faulkner et al. (2003).

Segundo os autores, os fragmentos consistem de enunciados de uma ou mais palavras, sendo que metade dos enunciados de uma palavra são substantivos, enquanto os fragmentos com mais de uma palavra compõem-se de NPs (43%), VPs (23%), PPs (10%) e outros (24%). Construções complexas correspondem a orações com complemento sentencial ("*I think it's going to rain*") e orações subordinadas adverbiais, introduzidas por *because*, *if* e *when*. Se considerarmos as construções de cópula, sujeito–predicado e complexas como o conjunto das declarativas, obtemos o contraste entre os dois levantamentos apresentados na Tabela 5.3.

Na especificação do corpus do modelo foram excluídas as interjeições e acrescidas sentenças passivas. Apesar de não contempladas pelos levantamentos citados, passivas são construções às quais a criança está exposta e que se desenvolvem em sua fala no período

Tipo de construção	Hoff-Ginsberg (1986)	Cameron-Faulkner et al. (2003)	IASMIM
Fragmentos	–	.20	.20
Questões	.32	.31	.32
Qu	.17	.16	
Sim/não	.15	.15	
Imperativas	.08	.09	.09
Declarativas	.25	.39	.39
Total			1.00

Tabela 5.3: Tipos de construções e frequências aplicadas ao corpus submetido ao IASMIM.

de aquisição simulado pelo modelo (ver Israel et al., 2000, por exemplo). Vale ressaltar que as frequências não foram controladas no interior de cada classe de construções, ou seja, os sub-tipos apresentam frequência aleatória. Por incluir os tipos de construções acima, o corpus submetido ao aprendiz pode ser considerado relativamente abrangente, especialmente quando comparado às modelagens analisadas no Capítulo 3. Ainda assim, ficaram de fora várias construções e elementos comuns desse período, tais como negações, sentenças com omissão de elementos (sujeito ou objeto nulo e elipses), quantificadores, etc. Dos auxiliares, apenas os verbos *do* e *will* foram considerados.

5.1.2 Classes de palavras

Sendo um dos objetivos deste estudo o desenvolvimento de um modelo com potencial translinguístico, foram preparados corpora tanto para o inglês, quanto para o português brasileiro. Além disso, um terceiro corpus para avaliar questões envolvendo ordem foi criado artificialmente, utilizando o léxico do inglês, mas em que a ordem das palavras era estritamente núcleo-final (exceto em interrogativas, em que os verbos auxiliares iniciam as sentenças, como no inglês). Em geral, os itens lexicais do PB são uma tradução dos itens lexicais do inglês, visto que a ideia era testar construções equivalentes das duas línguas. A Tabela 5.4 sumariza as classes de palavras contidas no léxico:

Determinantes definidos e indefinidos
Pronomes de primeira, segunda e terceira pessoas (formas nominativas e acusativas)
Palavras-Qu (“quem”, “o que” e “qual/quais”)
Preposições (locativa, dativa, genitiva, instrumental e agentiva)
Complementizador “que” (para complementos sentenciais)
Nomes próprios e comuns (animados, inanimados, referentes a locais, instrumentos, etc.)
Adjetivos intersectivos (p.e., “vermelho”) e operadores (p.e., “falso”)
Advérbios de frequência, maneira, local, etc.
Auxiliares (“will” e “do”)
Cópulas (“be” e “get”)
Verbos intransitivos (inergativos, inacusativos e incoativos), transitivos (incluindo verbos com complemento sentencial) e ditransitivos

Tabela 5.4: Classes de palavras contempladas no léxico dos corpora submetidos ao IASMIM.

5.1.3 Corpora criados automaticamente

Exceto por um dos corpora usados no experimento, todos os demais foram criados automática e aleatoriamente, a partir de um pequeno aplicativo desenvolvido para este fim. O aplicativo consiste de gramáticas sensíveis ao contexto especificadas manualmente para cada língua (inglês, português e uma língua artificial simulando ordem núcleo-final) e de um procedimento que utiliza tais gramáticas para gerar a quantidade desejada de dados de entrada, lembrando que cada dado consiste do par ⟨UTT,CTT⟩. Como dito acima, para os tipos básicos de construção, há frequências vinculadas, de modo que o procedimento calcula a quantidade proporcional de enunciados para cada tipo. Em função do caráter aleatório do procedimento e do foco apenas na estrutura sintática, eventualmente são produzidas sentenças semanticamente ruins, seja por combinações improváveis entre nomes, verbos e adjuntos, seja pela repetição de palavras numa mesma sentença. Em (38), são exibidos alguns exemplos de enunciados gerados para o corpus em português, alguns semanticamente adequados (a-d), outros nem tanto (e-h).

(38) a. Uns ratos pequenos crescerão

b. Eu enviei eles para o César

c. Eu dançarei

d. Nos quartos

e. Um primeiro milagre derreterá

f. Nós daremos quais pés para uns tios?

g. Dêem os falsos pés do primeiro rapaz para uns falsos rapazes

h. Uns gatos nos quartos grandes sempre darão uns primeiros caixotes para um professor

i. etc...

5.2 A aprendizagem no modelo

5.2.1 Visão geral

A Tabela 5.5 resume as características dos corpora preparados para as simulações. O valor da MLUw indicado faz menção ao *número médio de palavras*, não de morfemas, que é a medida tradicional. Isto se deve ao fato de que o modelo considera a palavra como unidade mínima. Porém, há estudos, como o de Parker & Brorson (2005), que indicam que as duas medidas estão quase perfeitamente correlacionadas e que, portanto, a medida MLUw pode ser usada tão efetivamente quanto a medida em morfemas. Note também que para o inglês foi gerado um conjunto de pouco mais de 40 mil enunciados, enquanto para o português foram 100 mil. O fato é que há um número significativamente maior de itens lexicais no corpus do português (especialmente o corpus ampliado), em função da morfologia da língua. Por esta razão, foi submetido ao aprendiz o dobro de enunciados, no intuito de evitar que a extensão do corpus fosse um impedimento para a convergência, no caso do português. Como veremos mais adiante, este fator acabou não sendo relevante.

Quando exposto ao *corpus mínimo*, preparado especialmente para o desenvolvimento

Corpora	Enunciados	Palavras	MLUw	Léxico
Corpus mínimo (dados do inglês)	985	3065	3.11	52
Corpus "núcleo-final"	2071	10347	5.00	56
Corpus inglês	40863	245111	6.00	91
Corpus português	100000	575449	5.75	133
Corpus português (ampliado)	100000	577349	5.77	464

Tabela 5.5: Os corpora submetidos ao IASMIM.

do modelo e pensado de modo a acelerar a aprendizagem, o aprendiz se mostrou capaz de convergir, tanto na aquisição lexical, quanto na sintática. No caso da primeira, convergir significa – assumindo o mesmo critério de Siskind (1996) – adquirir pelo menos um sentido por palavra, para 95% das palavras do léxico. Convergir na aquisição sintática significa formar categorias mais abstratas, de modo a capturar duas propriedades tidas aqui como reveladoras de conhecimento sintático: a repetição de rótulos em *estruturas de adjunção* e a emergência de *padrões recursivos* na árvore.

Para os corpora mais extensos, cujo objetivo era o de impor condições mais difíceis e realistas ao modelo, o aprendiz adquiriu com relativa estabilidade (entre as simulações) itens funcionais (determinantes, pronomes, palavras-Qu, preposições e o complementizador), nomes, adjetivos, advérbios, cópulas e verbos no imperativo. Porém, quando se trata dos demais verbos, em suas diversas flexões e ocorrências em declarativas e interrogativas, a performance do aprendiz foi bastante limitada em geral, sistematicamente convergindo para sentidos incorretos (por um ou outro símbolo conceitual incorretamente conjecturado como necessário) ou simplesmente não convergindo. Mesmo para o corpus do inglês, em que atingiu um alto grau de convergência, uma parte dos sentidos adquiridos não corresponderam aos almejados, inclusive com mais ocorrências de falsos positivos (sentidos convergentes incorretos).

No que diz respeito à aquisição sintática, a dificuldade do aprendiz com os corpora mais extensos inviabilizou uma avaliação mais abrangente, visto que sem aquisição lexical, não é

possível à sintaxe processar os enunciados. Outro fator inviabilizador foi o tempo de execução necessário para que o modelo processasse os dados. Para se ter uma ideia, apenas para os 2071 enunciados do corpus núcleo-final, várias horas de processamento foram necessárias. Este nível de desempenho inviabiliza a avaliação da aquisição sintática no modelo para corpora mais extensos, como os de 50 ou 100 mil enunciados submetidos para aquisição lexical.

Em parte, este problema se deve a ajustes necessários sobre a implementação, de modo a otimizar seu desempenho. Mas outra parte decorre do fato de que o modelo nunca deixa de checar os procedimentos de aprendizagem, que fazem parte do fluxo de processamento. Apesar desta limitação, as simulações com o corpus mínimo e o corpus núcleo-final mostraram que o aprendiz é capaz de adquirir a sintaxe, tendo sido capaz de analisar fragmentos, declarativas e interrogativas simples (envolvendo cópulas). Também para o corpus núcleo-final, os padrões de adjunção e recursividade emergiram para estas construções, o que indica que o sucesso do aprendiz não resulta de um viés na preparação do corpus mínimo.

5.2.2 Aquisição lexical

Exceto para os corpora mínimo e núcleo-final, em que o aprendiz conseguiu convergir completamente, para os demais corpora seu comportamento foi sempre o de adquirir rapidamente o conjunto formado pelos itens funcionais, nomes próprios e comuns, adjetivos, advérbios, cópulas, auxiliares e verbos no imperativo, estabilizando aí e praticamente deixando de adquirir novos itens (especialmente para as várias formas flexionais de verbos intransitivos, transitivos e ditransitivos). Com um pouco mais de sucesso, o aprendiz conseguiu convergir para verbos na voz passiva. A Tabela 5.6 apresenta um quadro resumido geral do desempenho do aprendiz na aquisição lexical e os gráficos apresentados nas Figuras 5.1 à 5.5, mostram as curvas de aprendizagem do aprendiz para cada corpus.

O corpus mínimo foi o mais fácil de ser processado, como era de se esperar, visto que serviu de base para o desenvolvimento do modelo. Por ser um corpus pequeno, composto por 197

		Previsto	Adquirido		
Corpus	Enunciados	Léxico	Léxico	Falsos positivos	Convergência
Corpus mínimo	985	52	52	5	100%
Corpus núcleo-final	2071	56	54	0	96,4%
Corpus EN	40083	91	87	11	95,6%
Corpus PB	100 mil	133	70	2	52,63%
Corpus PB (ampliado)	100 mil	464	183	3	39,43%

Tabela 5.6: Quadro geral dos resultados da aquisição lexical no modelo.

enunciados, o aprendiz fez cinco passagens seguidas sobre ele, totalizando uma experiência de 985 enunciados.[1] Na primeira passagem, os enunciados vinham em uma ordem que facilitava a aquisição, pois iniciava por fragmentos de sintagmas nominais simples, seguidos de fragmentos de sintagmas nominais com adjuntos (adjetivos e sintagmas preposicionados), e finalmente orações declarativas e questões sim/não simples, envolvendo cópulas. Nas demais passagens, o corpus era ordenado aleatoriamente. Como se vê no gráfico, o corpus mínimo favoreceu bastante a aquisição, de modo que quase todas as palavras foram adquiridas logo na primeira passagem, com as demais sendo adquiridas no decorrer da segunda.

O corpus núcleo-final é uma versão do corpus do inglês adaptado de modo a apresentar ordem quase estritamente núcleo-final e também para conter, de modo um pouco mais ampliado que o corpus mínimo, apenas fragmentos, sentenças declarativas, questões sim/não e questões-Qu. Os resultados mostraram que a ordem artificial de exposição imposta ao corpus mínimo não é uma condição para a convergência da aquisição lexical, pelo menos no que diz respeito aos tipos de construções mais simples. Diferentemente do que ocorreu para o próximo corpus, aqui os sentidos para os quais o aprendiz convergiu foram os almejados, inclusive sem apresentar nenhum falso positivo.

[1] Neste caso, foi uma forma de simular as várias exposições que a criança tem aos enunciados, porém sem a interferência de enunciados mais complexos, que atrasariam a aprendizagem. Obviamente não é o caso da criança, mas aqui o intuito era o de verificar se os procedimentos concebidos convergiam, dada uma situação ideal de exposição. A partir daí, confirmada a convergência, os demais corpora tinham o objetivo de impor condições mais realistas.

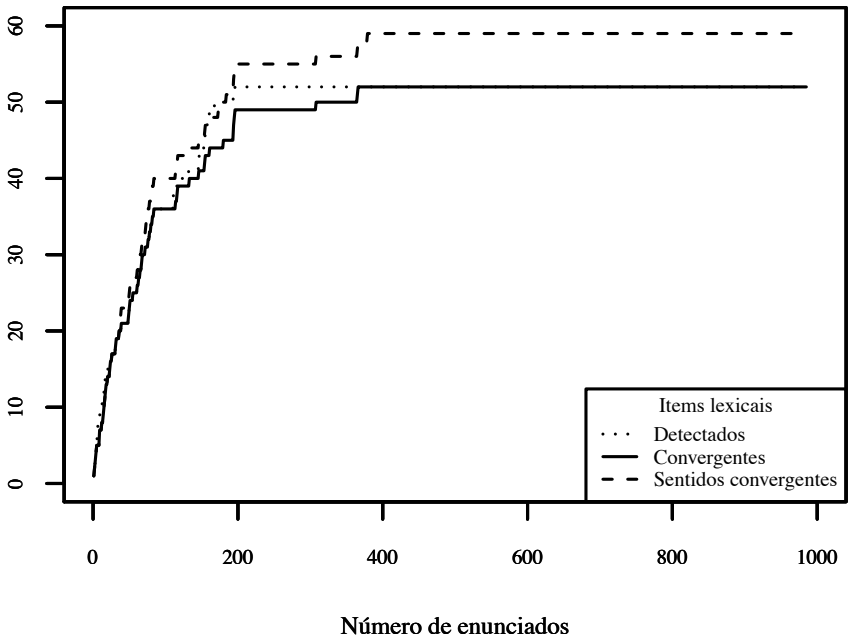


Figura 5.1: Aquisição lexical do aprendiz para o corpus mínimo.

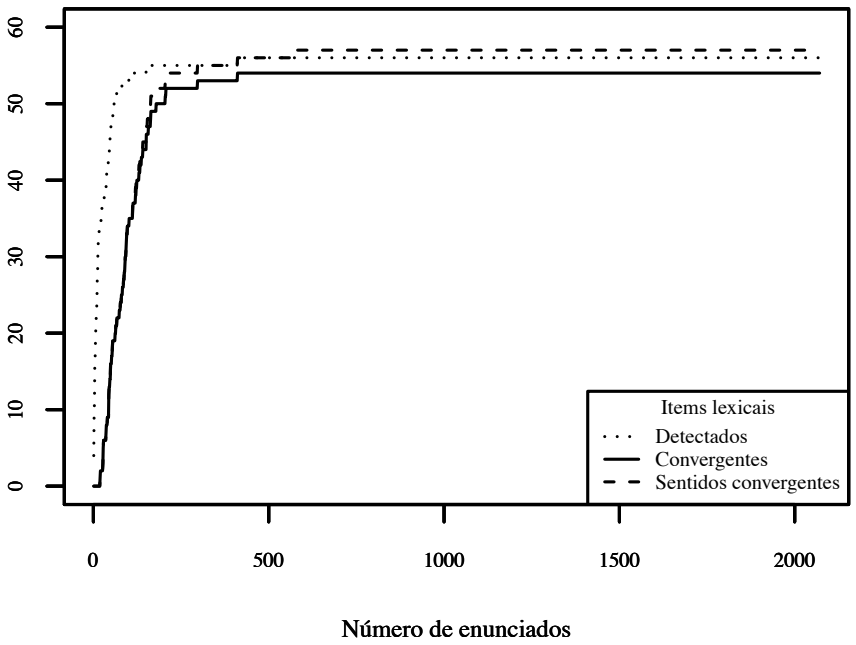


Figura 5.2: Aquisição lexical do aprendiz para o corpus núcleo-final.

O corpus do inglês preparado para a simulação foi um corpus mais complexo, incluindo os demais tipos de construção deixados de fora nos dois primeiros. A quantidade de itens lexicais, no entanto, foi controlada, de modo a manter um baixo número de verbos, com

cada classe (inacusativos, transitivos, etc.) contendo dois exemplares, visto que o acréscimo no número de verbos tem impacto bastante prejudicial sobre o sucesso do aprendiz, como veremos para o corpus ampliado do português. Aqui, também, o aprendiz apresentou uma convergência bastante alta, acima de 95%. Porém, os sentidos conjecturados se afastaram dos almejados, particularmente em relação aos verbos, artigos, nomes próprios e pronomes.

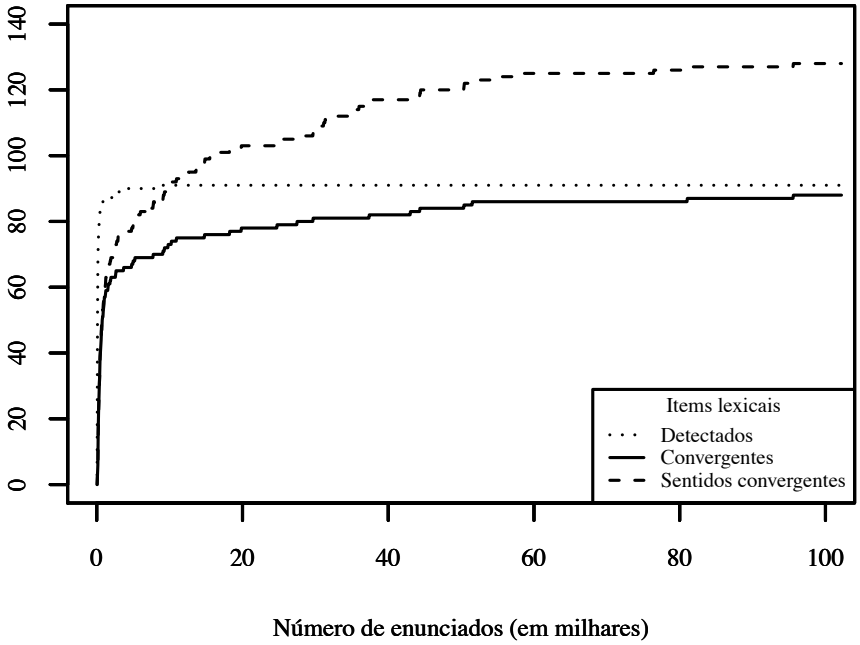


Figura 5.3: Aquisição lexical do aprendiz para o corpus do inglês.

Ocorreu que em muitos casos o aprendiz conjecturou que artigos e pronomes representavam propriedades tais como força da sentença (declarativa ou interrogativa) e finitude. Em função disto, os verbos acabaram convergindo para sentidos que incluíam a definitude dos argumentos como parte de seu sentido e os nomes próprios, ao contrário, convergiram para sentidos que excluíam a informação sobre definitude. Este comportamento desviante resultou em um grau maior de polissemia, visto que um mesmo verbo teria que exibir várias entradas para dar conta da variação de definitude de seus argumentos. Isto, em conjunto com os casos de falsos positivos, explica a maior diferença observada nessa simulação entre as curvas de palavras e de sentidos convergentes.

O desempenho cai drasticamente, porém, quando foram submetidos ao aprendiz os corpora do português. A principal diferença destes corpora em relação ao do inglês é o grande número de verbos, em função de suas formas flexionais específicas. Com isto, o corpus do português, que também foi controlado (como o do inglês) para apresentar duas instâncias de cada classe de verbo, acabou por apresentar um léxico de 133 palavras, contra 91 do anterior. Esta diferença causou grande impacto na performance do aprendiz, que passou dos 90% para os 50%. Como já foi dito, para itens funcionais, cópulas, nomes próprios e comuns, verbos no imperativo e na voz passiva, a aquisição seguiu convergindo. Porém, para as demais formas verbais, o desempenho foi bastante problemático.

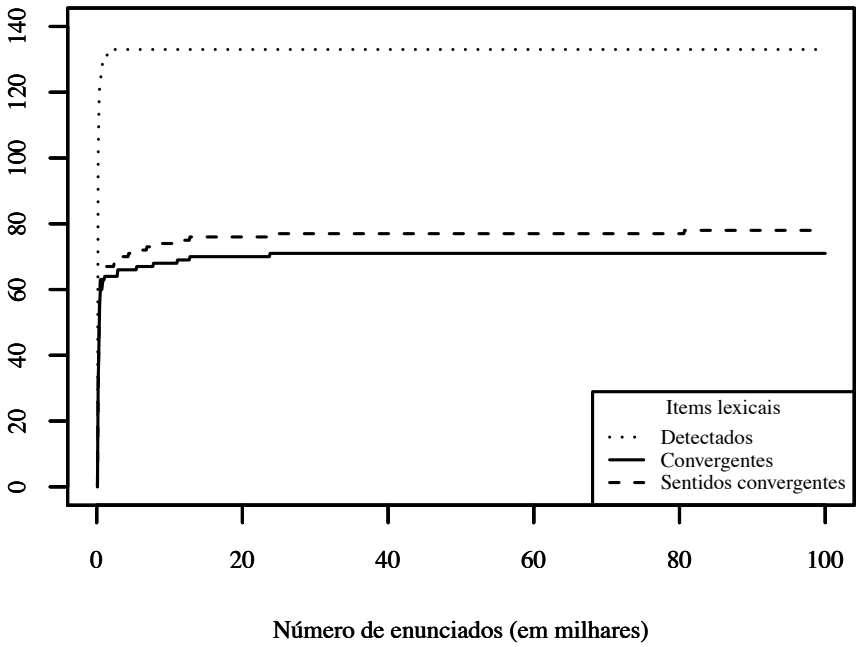


Figura 5.4: Aquisição lexical do aprendiz para o corpus do português.

Finalmente, a simulação com o corpus ampliado do português apenas replicou o comportamento verificado na anterior, com a diferença que agora, havendo ainda mais verbos, a aquisição se estabilizou por volta dos 40% de aquisição do léxico, correspondente aos itens também adquiridos no anterior.

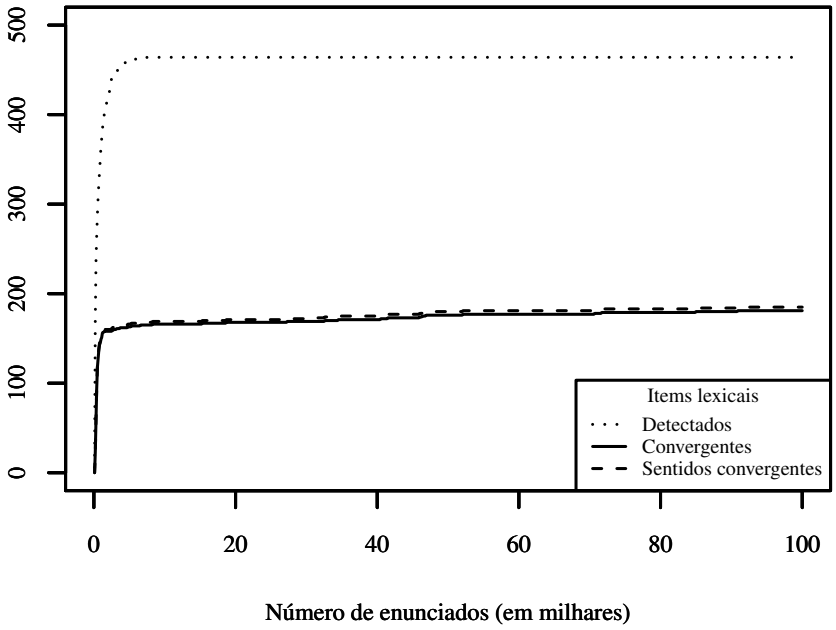


Figura 5.5: Aquisição lexical do aprendiz para o corpus ampliado do português.

Polissemia e esparsidade nos corpora do português

É possível que as dificuldades do aprendiz em convergir para os corpora do português se devam à quantidade insuficiente de dados. Ocorre que o algoritmo de Siskind é bastante sensível ao grau de homonímia/polissemia do corpus. Na presente simulação, os corpora criados são significativamente mais complexos (contém mais atributos e tipos de sentenças) do que os corpora nos experimentos em Siskind (1996), inclusive em termos de polissemia, entendida aqui como palavras que possuem duas ou mais entradas lexicais que compartilham parte dos símbolos conceituais. Nos experimentos de Siskind, cada palavra correspondia, em geral, a apenas um símbolo conceitual, enquanto aqui nomes próprios, verbos e preposições (no caso do português) correspondem a vários. Portanto, uma primeira razão para o desempenho abaixo do esperado nas simulações apresentadas pode ser a polissemia dos dados.

Mas mesmo considerando que tais dados se enquadrem nos casos de homonímia, Siskind (1996) mostra que o grau de homonímia (palavras com dois ou mais sentidos completamente

distintos) do corpus tem impacto direto sobre a convergência, de tal forma que um grau médio de 2 (i.e., 2 sentidos por palavra), demandaria quase um milhão de palavras, ou seja, o dobro do que foi disponibilizado nas simulações do português e quase quatro vezes mais dados no caso do inglês. Portanto, para averiguar corretamente a capacidade do IASMIM, seria necessário submeter pelo menos essa mesma quantidade de dados. Não foi possível averiguar isto, entretanto, visto que o algoritmo precisa ser otimizado de modo a tornar viável lidar com tal quantidade de dados.

Outro fator possível, especialmente para explicar a desempenho claramente pior para o português, é a morfologia. Dado que o algoritmo de Siskind (op.cit.) foi testado apenas para o inglês, é possível que a morfologia do português esteja interferindo decisivamente, causando um grau de esparsidade nos dados maior do que o algoritmo consegue lidar. Em outras palavras, cada forma flexionada de um verbo tenderá a ter baixa frequência e grande distância (em número de enunciados) entre cada ocorrência. Nos gráficos abaixo, note como os corpora do português apresentam frequências relativas em geral bem mais baixas do que os corpora núcleo-final e do inglês.

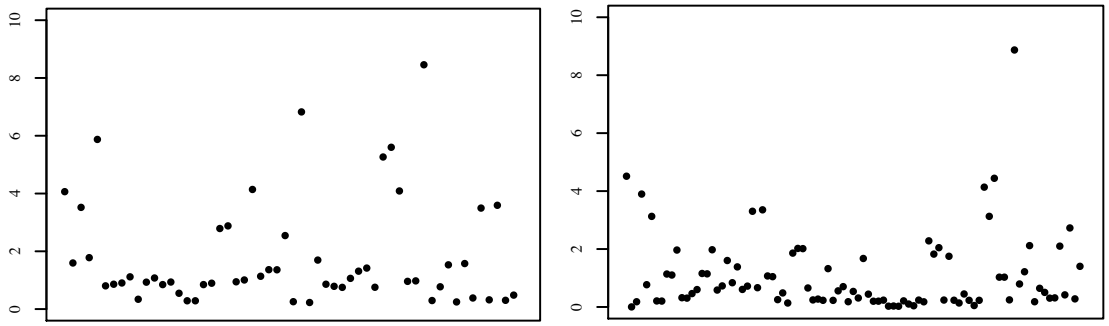


Figura 5.6: Frequência relativa das palavras para o corpus núcleo-final e do inglês.

Em geral, as frequências ficam abaixo de 10% para todos os corpora. Note como, no português, as frequências tendem a formar uma linha baixa, próxima do zero. Por serem em geral tão baixas, as palavras correspondentes serão relativamente raras no corpus. Com isso, é muito provável que não haja tempo hábil para que a palavra venha a convergir, antes de ser

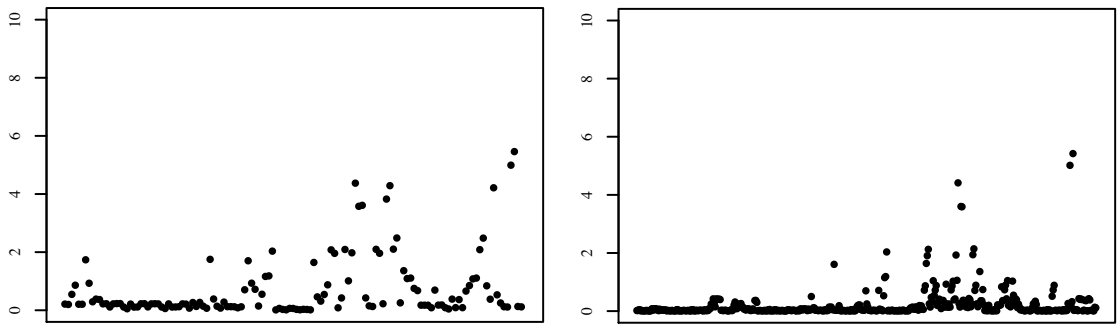


Figura 5.7: Frequência relativa das palavras para os corpora (reduzido e ampliado) do português.

recolhida pelo procedimento de "garbage collection" do algoritmo, que a cada 500 enunciados remove os sentidos que não "congelaram", isto é, que não convergiram e foram utilizados algumas vezes com sucesso. Tal procedimento é necessário, entre outras coisas, para permitir que o algoritmo manipule uma quantidade viável de hipóteses durante a aquisição.

Por esta razão, foi testada uma outra medida para o "garbage collection", no caso uma que só descartava a palavra após 50 ocorrências sem convergir. Os resultados, porém, permaneceram baixos. Isso indica, portanto, que seria necessário tornar o algoritmo mais sensível a distribuições como as do português, de modo a superar o problema de esparsidade causado pela morfologia da língua. Obviamente, se isto ocorre para o português, os resultados seriam ainda piores para línguas com marcação de caso, visto que nem mesmo os artigos, substantivos e adjetivos, escapariam ao efeito da esparsidade.

Reconhecimento de enunciados

A aquisição lexical cria condições para que o aprendiz possa, de modo relativamente gradual, reconhecer enunciados e enviá-los para o analisador. Nas simulações, o aprendiz demostrou uma preferência clara por enunciados mais curtos, mesmo que o tipo de enunciado tenha uma frequência mais baixa no corpus, como é o caso das imperativas. Isto mostra que o nível de complexidade (para a aquisição lexical) do enunciado se sobrepõe à frequência, em termos de impacto sobre a ordem de aquisição. Assim, a maior parte do conhecimento

sintático desenvolvido nos primeiros estágios corresponderá a regras para processamento de fragmentos (DPs e PPs), sentenças imperativas e declarativas. O gráfico na Figura 5.8 exibe o número de enunciados reconhecidos por tipo, em função da exposição aos dados.

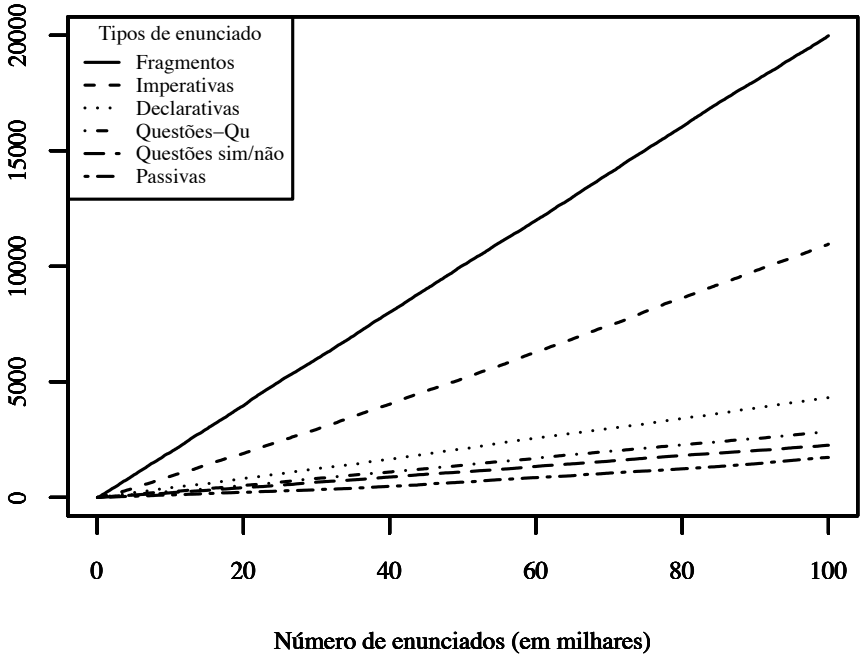


Figura 5.8: Reconhecimento de enunciados por tipo para o corpus ampliado do português.

5.2.3 Aquisição sintática

A aquisição sintática no modelo tem a característica de ser bem mais simples do que a lexical. Porém, o desenvolvimento do conhecimento sintático depende de modo crucial do sucesso da mesma. O fato é que, uma vez que a aquisição lexical chega ao ponto de identificar corretamente as expressões conceituais respectivas a cada item e estas seguem para o analisador, este *quase sempre* vai – em condições normais e com o auxílio dos procedimentos de aquisição – produzir análises para os enunciados. Por condições normais, me refiro a dados de entrada que não extrapolam os limites impostos ao analisador, tal como o limite de antecipação, e a sentidos consistentes adquiridos na etapa lexical, de modo que cheguem elementos "combináveis" para o analisador.

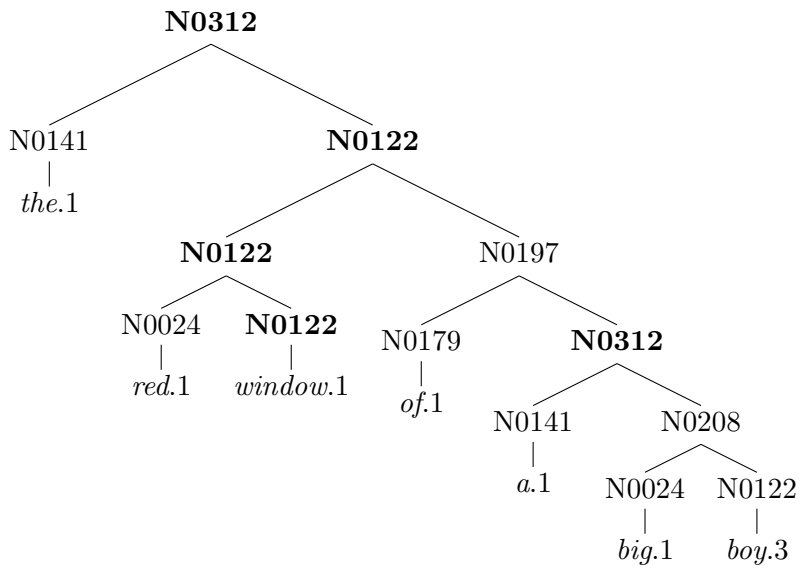
A aquisição sintática vai, então, proceder no sentido de criar categorias e regras grama-

ticais para analisar os enunciados, generalizando-as à medida em que acumula experiência. Desse modo, o conjunto de categorias e regras tende, inicialmente, a crescer mais rapidamente até que, num certo ponto da experiência, começa a desacelerar e a retroceder, visto que a generalização vai eliminando regras específicas em favor de regras mais gerais. Numa gramática madura, a tendência é que modificações sobre a gramática passem a ser esporádicas e marginais, com a mesma apresentando uma relativa estabilidade (em termos do número de regras e categorias envolvidas). Uma medida relevante, neste sentido, é o grau de "compressão" que a generalização confere à gramática, isto é, o quanto a gramática obtida é reduzida (em número de regras e categorias) em relação à gramática que seria obtida sem generalizações.

Recursividade e adjunção

O principal fato a destacar na aquisição sintática, foi a capacidade do aprendiz de generalizar as análises ao ponto em que padrões de adjunção, representados como segmentos de uma mesma categoria, e padrões de recursividade pudessem emergir nas estruturas sintáticas obtidas. Como mostram os exemplos (39), abaixo, o padrão de adjunção emerge na estrutura evolvendo os segmentos da categoria N0122 (núcleo da adjunção) e os adjuntos N0024 e N0197. Já o padrão recursivo se reflete na ocorrência da categoria N0312 aninhada no interior da estrutura, cuja raiz é também da mesma categoria.

(39)



Como os procedimentos de aquisição e generalização se aplicam independentemente da categoria específica em jogo, não há dúvidas de que os mesmos padrões surgiriam em contextos de adjuntos adverbiais e orações subordinadas, bastando para isso que o aprendiz fosse exposto à experiência respectiva. Berwick (1985) e Gaylard (1995) parecem identificar o aspecto recursivo da gramática com a presença de construções contendo orações subordinadas, razão pela qual seus modelos necessitam ou da estrutura X-barra como parte do aparato inato (Berwick, 1985) ou de sentenças com dois níveis de subordinação (Gaylard, 1995), para que a recursividade emerja em suas modelagens.

No modelo, a recursividade resulta da interação entre propriedades das expressões conceituais (que também são estruturas recursivas), propriedades da componente sintática (a operação *merge* e rótulos) e propriedades dos procedimentos de aquisição e generalização. Ela não tem que ser aprendida, portanto, mas vai emergir ou não nas estruturas sintáticas a depender das características da língua. No caso do inglês e do português, padrões de recursividade podem emergir com base em dados muito mais simples, tais como fragmentos de sintagmas nominais contendo adjuntos, um resultado à primeira vista mais plausível do que o de Gaylard (op.cit.), que depende de orações com dois níveis de subordinação, e o de Berwick (op.cit.), na medida em que o presente modelo obtém este resultado sem o aparato

da teoria X-barra.

Compressão e adequação descritiva

A compressão da gramática não indica, por si só, se os procedimentos de aquisição estão trabalhando corretamente. É preciso que a gramática seja compacta *mas que também* descreva apenas as construções gramaticais da língua, isto é, que não sobregeneralize. Na simulação com o corpus mínimo, sem generalização, o aprendiz induz uma gramática constituída por 260 categorias sintáticas e 625 regras gramaticais. Se o procedimento é habilitado, a gramática é reduzida drasticamente, passando a apresentar 28 categorias e 82 regras, que em termos percentuais equivalem, respectivamente, a 89,2% e 86,88% de redução. Os demais MCAs considerados aqui não apresentam dados em relação a este aspecto, o que impede uma comparação direta.

O modelo obtém esta redução sem, no entanto, gerar regras demasiadamente gerais, conforme pode ser conferido no Apêndice A (note a presença dos atributos definidores das categorias, mesmo quando subespecificados; isto é o que garante que uma regra não se aplica a qualquer elemento). Esta redução faz todo o sentido, se pensarmos que sem a generalização, a gramática tem que criar uma regra específica para cada combinação particular de palavras, à despeito de serem da mesma categoria. Desse modo, o número de regras da gramática será diretamente proporcional ao número de itens lexicais da língua. Não há dúvidas, portanto, sobre a importância dos procedimentos de generalização. Além da simulação com o corpus mínimo, a aquisição sintática foi também simulada com o corpus núcleo-final. Os gráficos apresentados a seguir sumarizam o desempenho do aprendiz para ambos.

No gráfico da Figura 5.10, além do número bem maior de regras e categorias (em função de um corpus maior e mais diversificado), note que a quantidade de dados a que o aprendiz foi exposto parece ter sido insuficiente para que a curva de aquisição exibisse o padrão de queda mais acentuada no número de categorias antes de estabilizar na gramática-alvo, observado

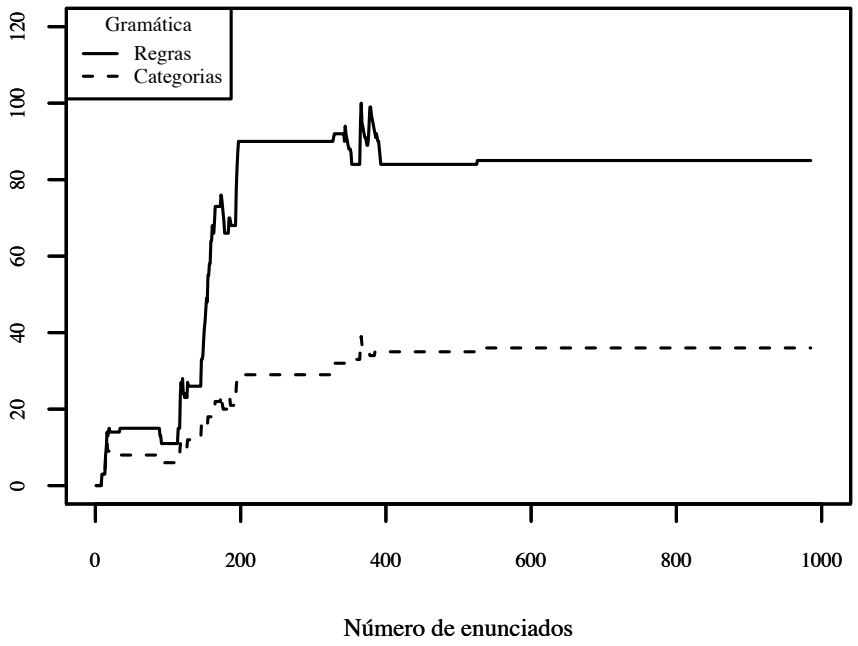


Figura 5.9: Aquisição sintática do aprendiz para o corpus mínimo.

na simulação com o corpus mínimo. Ainda assim, é possível observar a tendência inicial de crescimento acentuado da gramática, seguido de uma gradual estabilização da curva, inclusive com alguns pequenos trechos de queda no número de regras e categorias, indicando que a gramática está convergindo para a gramática-alvo.

Finalmente, era preciso averiguar o desempenho da heurística voltada para a identificação de núcleos abstratos. Os resultados mostraram que ela funcionou como esperado, à despeito do núcleo abstrato ocorrer no topo ou no interior da árvore. Para simular elementos no interior da estrutura, foram fornecidos exemplos como "*the cats of boys*", simulando a hipotética ausência de artigos na língua (caso em que "boys" poderia ser definido ou indefinido). Neste caso, para que "boys" pudesse ser anexado ao sintagma preposicional, era necessário completar a estrutura com uma categoria abstrata relativa à definitude.[2] Para categorias no topo da árvore, a própria aquisição lexical fez com que verbos e auxiliares con-

[2] O que poderia implicar que no modelo o comportamento de sintagmas nominais (NPs) em línguas sem artigo seria semelhante ao de NPs em línguas em que há artigos, o que não parece correto. Isso indica que o conceito de "categorias abstrata" no modelo talvez não seja adequado e precisaria ser repensado.

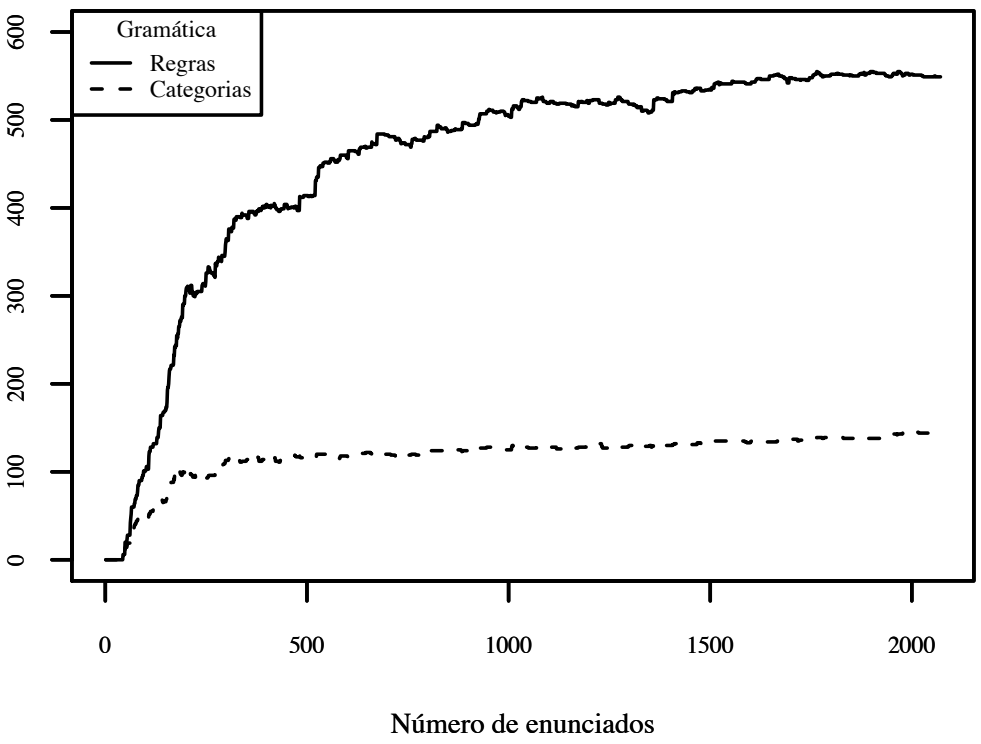
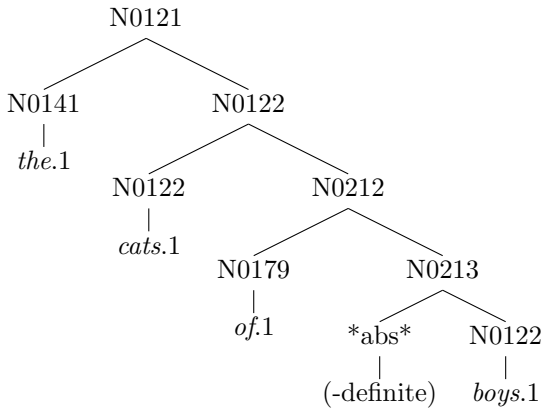


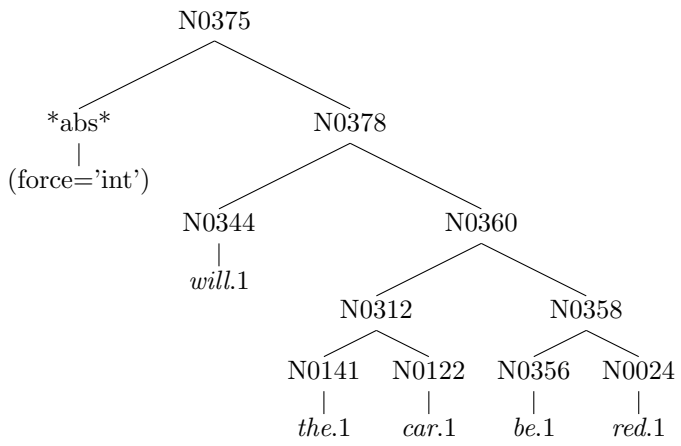
Figura 5.10: Aquisição sintática do aprendiz para o corpus núcleo-final.

vergissem para sentidos mais restritos, que não continham o atributo `[force]`, por exemplo. Com isso, o analisador foi capaz de processar enunciados como os exemplificados em (40) e (41).

(40)



(41)



Processamento de ordem núcleo-final

Em termos gerais, o processamento de enunciados cuja ordem é quase estritamente núcleo-final[3] foi natural para o modelo. Alguns pequenos ajustes tiveram que ser feitos na forma como ele percorria o *buffer*, especialmente quando antecipava algumas células. No mais, o modelo como estava concebido já exibia a capacidade para processar este tipo de ordem. Assim, esta simulação permitiu observar o impacto da ordem nas análises sintáticas e o que se observou, como esperado, é que para tais tipos de línguas, dada a limitação de antecipação imposta ao aprendiz, ocorrem mais falhas de análise, visto que para muitos enunciados, a janela de antecipação de até duas células consecutivas se mostra insuficiente. Aproximadamente 18,83% dos dados de entrada do corpus núcleo-final foram descartados em função disso.[4]

Por exemplo, suponha que o aprendiz receba o enunciado "*small who seldom is?*", cujo sentido é "*who is seldom small?*". Cada elemento do enunciado vai ocupar uma posição no *buffer*, de modo que as três primeiras corresponderão à `[ small, who, seldom ]`. Ao iniciar a análise, o símbolo de entrada será *small*. Como este símbolo não possui seletores e ainda não há nenhum sintagma em formação, o analisador não pode aplicar nenhuma regra

[3] A exceção foram as interrogativas, em que o auxiliar iniciava as sentenças.

[4] É claro que a limitação da antecipação não seria necessariamente absoluta na criança, tal como é no modelo. É possível que o valor '3' para a janela seja uma "média" da antecipação psicolinguística, com o falante podendo trabalhar com janelas um pouco maiores em algumas situações.

e nem pode criar alguma, com base nas heurísticas. Assim, ele utiliza a antecipação para checar se alguma ação pode ser tomada com base numa das duas células seguintes. Porém, nem *who*, nem tampouco *seldom*, permitem disparar alguma ação, visto que também não possuem seletores e não estão relacionados entre si por adjunção.

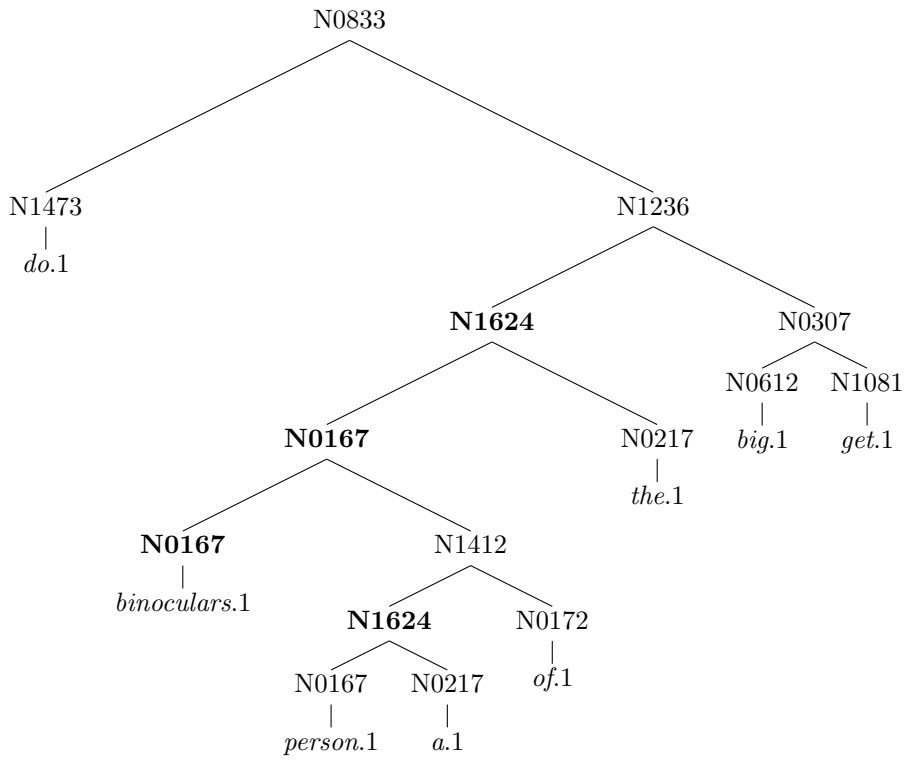
Desta forma, o analisador chega à terceira célula sem que nenhuma ação possa ser tomada. Neste caso, o analisador falha e abandona a análise. Isso significa que tais enunciados estão fora do escopo da aprendizagem do modelo, o que pode ser interpretado como evidência para alguma dificuldade que a criança enfrentaria com tais construções nestas línguas. Em outras palavras, a previsão do modelo é que a criança vai tirar proveito, pelo menos nos primeiros estágios multi-palavras, dos enunciados para os quais a antecipação de até duas células é suficiente para processá-los. No exemplo em questão, vale ressaltar, há a presença de um advérbio, exatamente o elemento que neste caso inviabiliza a análise. Bastaria, portanto, que a criança fosse exposta também a quantidades relevantes de enunciados sem advérbios.

Fica claro, também, que fenômenos como a ordem verbo-final em subordinadas do alemão, poderiam ser mais difíceis para o modelo adquirir, pelos motivos expostos. Novamente, é uma questão de haver *também* dados que caibam na janela de antecipação. É importante salientar que o fato de processar línguas que exibem ordem núcleo-final de modo mais acentuado é um passo adiante que o IASMIM dá, quando comparado a outras modelagens em aquisição. Mas seria interessante contar com corpora realistas para estes tipos de línguas, ao invés de um corpus artificial como o utilizado neste estudo, para observar mais adequadamente o desempenho do modelo.

A seguir, apresento um exemplo de análise de um enunciado deste corpus, a saber, "*do binoculars person a of the big get*" ("*do the binoculars of a person get big?*"). O exemplo em (45) demonstra a aprendizagem de enunciados menos triviais, já que estão em jogo um verbo contendo dois seletores, um auxiliar e um adjunto preposicionado. Além disso, o exemplo mostra novamente a emergência de padrões de adjunção e recursividade na gramática

adquirida, como se pode ver nas subestruturas envolvendo as categorias N0167 e N1624.

(42)



5.3 Outros aspectos

5.3.1 Ambiguidade dos dados

Apesar de ter como meta realizar análises determinísticas para os enunciados (uma vez que cheguem a ele), o analisador vai eventualmente errar em função de ambiguidade. Por exemplo, tomemos o enunciado "do cats fake some station a at get" ("*Do some fake cats get at a station?*"). O analisador, após processar "do" (criando o sintagma respectivo e anexando o auxiliar como núcleo) e "cats fake some", criando o DP respectivo, chega ao ponto de processar "station a", para também formar um DP. Se ele não tiver uma regra já adquirida e compatível no momento desta análise, o procedimento de aquisição vai tentar criar uma regra. Com base em "station", não há nada a ser feito, pois este não contém seletores.

Para o modelo, ademais, "a" e "some" são elementos iguais: ambos possuem a expressão conceitual `[f=[-definite], cf=[], m=[]]`, que é o que o analisador de fato considera.[5] Assim, ao olhar para o contexto, em busca de uma combinação envolvendo "a", o modelo pode se equivocar e criar uma regra para processar o DP que já foi criado (informação que não é considerada nas heurísticas). Neste caso, a análise vai fracassar, pois "station" não poderá ser anexado ao DP, visto que este esperaria um NP com núcleo "fake". No entanto, o aprendiz poderia superar este problema: bastaria que ele fosse exposto a enunciados em que haja não dois artigos indefinidos, mas sim um artigo definido e outro indefinido. Assim, não haveria ambiguidade e a análise teria sucesso. Com o tempo, as regras seriam generalizadas e então enunciados como este poderiam ser analisados, pois não dependeriam do procedimento de aquisição.

5.3.2 Constituintes descontínuos

O modelo se mostrou capaz de construir os sintagmas, mesmo quando seus termos estavam separados por outros elementos. Por exemplo, o modelo é capaz de processar "*here Mary is*" (corpus núcleo-final), em que o sujeito se interpõe ao locativo e à cópula (que, devido a ordem composicional estrita da análise, precisam ser combinados antes de o sujeito ser anexado à estrutura). Isto é possível exatamente pelo fato de que a ordem linear não é parte da estrutura sintática e portanto elementos linearmente intrusivos não afetam a composição da estrutura.[6]

Assim, o modelo atinge maior abrangência do que o modelo de Gaylard (1995), que assume que a ordem é parte do conhecimento sintático a adquirir. No caso de Berwick (1985), o modelo também é mais abrangente, visto que ali, Berwick trata a ordem de modo absoluto em relação aos componentes X-barra ("spec", "head" e "comp"), de modo que uma vez configurada, o aprendiz não é capaz de processar variações, exceto no caso da inversão

[5] A informação sobre número no modelo é adquirida como parte da semântica dos substantivos.
[6] Exceto nos casos envolvendo o limite de antecipação, como comentado anteriormente para o corpus núcleo-final.

sujeito-auxiliar para a qual há uma regra transformacional prevista no modelo.

5.3.3 Desempenho em relação à Villavicencio (2002)

Villavicencio (2002) também implementou um algoritmo semelhante ao de Siskind (1996) para aquisição lexical. Assim, seus resultados permitem ter um referência para avaliação do IASMIM, visto que ambos se assentam em bases mais ou menos comuns para a aquisição lexical. Villavicencio não fornece a proporção em termos do número de palavras adquiridas, mas relata um desempenho de 63,6% de enunciados processados (i.e, reconhecidos) de um total de 1517 que compõem o corpus. Em relação a esta medida e considerando apenas os corpora com palavras do inglês, o IASMIM teve um desempenho de 96,15% de enunciados processados para o corpus mínimo, 92,32% para o corpus núcleo-final e 82,28% para o corpus do inglês. Considerando que o modelo de Villavicencio não conta com uma representação conceitual tão granular quanto a adotada aqui, o que dificulta a aquisição lexical, os resultados parecem bastante positivos e promissores.

5.4 Sumário

Neste capítulo foram apresentados e discutidos os resultados das simulações envolvendo o modelo. Corpora diferentes foram preparados e submetidos ao aprendiz, de modo que sua performance pôde ser melhor avaliada. Vários tipos de construções são contempladas nos corpora, a saber, fragmentos (DPs e PPs), sentenças declarativas com verbos de diferentes valências, orações subordinadas, passivas e imperativas, além de questões sim/não e questões-Qu. Frequências relativas a cada tipo de construção – obtidas a partir de levantamentos estatísticos disponíveis na literatura – foram aplicadas na geração dos corpora, de modo a mimificar minimamente a distribuição que estas apresentam na fala dirigida à criança.

Para os corpora com palavras do inglês, o modelo em geral alcança seus objetivos, isto é, consegue convergir de modo gradual na aquisição lexical e sintática, adquirindo sentidos corretos (inclusive casos de polissemia) e generalizando as regras ao ponto em que emergem

padrões de adjunção (i.e, segmentos de uma mesma categoria) e padrões recursivos. Além disso, os procedimentos conseguem compactar a gramática em quase 90%, mantendo, ainda assim, uma descrição precisa dos dados (i.e., não sobregera). O aprendiz é capaz de reconhecer acima de 80% dos enunciados e enviá-los ao analisador, para posterior processamento sintático.

Por outro lado, o aprendiz mostra dificuldades significativas para aquisição lexical em relação aos corpora (como os do português) em que há maior esparsidade dos dados, em particular, corpora em que grande parte do léxico a ser adquirido apresenta frequências de ocorrências muito baixas. Este é certamente um resultado indesejado diante da meta de obter um maior alcance translinguístico. Porém, os resultados da aquisição sintática para o corpus núcleo-final mostram que com relação à capacidade de lidar com línguas com diferentes ordens lineares canônicas, o modelo deu passos concretos nessa direção, restando para pesquisa posterior o desenvolvimento do algoritmo de aquisição lexical de modo a torná-lo mais sensível a línguas com morfologia mais rica e, portanto, maior esparsidade na distribuição das palavras pelo corpus.

Outro gargalo do modelo diz respeito à quantidade de dados com que é capaz de lidar em um tempo viável de processamento. Para a versão atual, é inviável testar a aquisição lexical em mais que 100 mil enunciados e, quando a aquisição sintática também é habilitada, 2 mil enunciados já impõem uma carga relativamente alta para o modelo. Parte desse alto custo computacional poderá ser reduzido, através de melhorias nos algoritmos de modo a torná-los mais leves e eficientes. Enfim, há otimizações a fazer, necessárias para que o modelo possa ser melhor avaliado, com corpora contendo na casa do milhão de palavras.

Outro aspecto importante a mencionar, é que a representação semântico-conceitual assumida é também relativamente independente de particularidades das línguas, o que permite que outros corpora sejam preparados com relativa facilidade. Certamente a representação é ainda bastante limitada, quando se consideram fenômenos gramaticais mais complexos que

foram deixados de fora deste estudo. Ainda assim, a representação semântica definida aqui parece propiciar uma base sólida para posteriores desenvolvimentos. Portanto, é possível dizer que o modelo alcança os objetivos de maior abrangência gramatical e universalidade, quando comparado aos MCAs anteriores a ele.

6

Conclusão

6.1 Uma nota sobre a aquisição da ordem linear

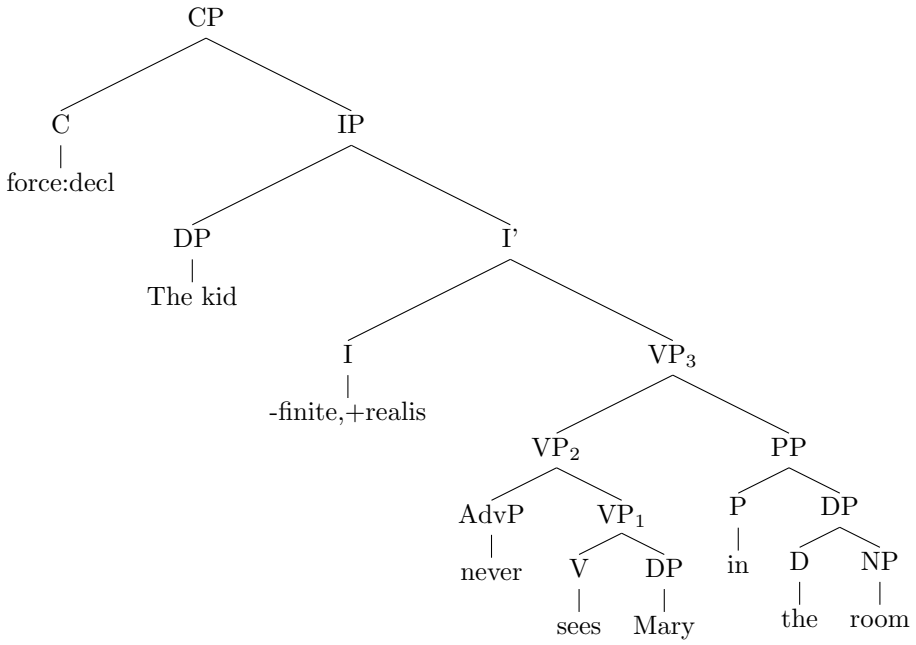
Embora tenha sido cogitado durante o desenvolvimento do modelo, a aquisição da ordem linear acabou ficando fora do escopo da aprendizagem prevista, principalmente por não se tratar do que considerei aqui como conhecimento sintático. Entretanto, seria interessante comentar brevemente de que modo esta aquisição se daria, dadas as características do modelo e dos objetos produzidos pela análise sintática. De acordo com as assunções do modelo, sendo a ordem linear um aspecto pós-sintático, caberia à interface PF dar conta dessa informação. Vale ressaltar que a interpretação dada à "interface PF" aqui difere da tradicional em pelo menos um ponto crucial: aqui, propõe-se que a interface pode ter acesso às relações estruturais (dominância e c-comando) da árvore sintática, embora não possa mais modificá-la.

Tomemos, por exemplo, o enunciado *The kid never sees Mary in the room* e a árvore[1] sintática correspondente em (43). Vale lembrar que a saída do processamento sintático, quando tem sucesso, é um objeto sintático correspondente à árvore sintática do enunciado. Assumindo que a sintaxe envia este objeto para a interface PF ("*spell-out*"), poderíamos

[1] Por razões expositivas, utilizo rótulos comumente encontrados nas análises propostas pela teoria sintática.

assumir algo semelhante à proposta de Kayne (1994) e estabelecer um procedimento em PF capaz de extrair os pares de c-comando envolvendo terminais da árvore. Para evitar os problemas com c-comando simétrico, assumo uma versão derivacional de c-comando, determinada por ocasião do *merge*, de modo que o núcleo sempre c-comanda o complemento (ou adjunto).[2]

(43)



Com isto, é possível aplicar o Axioma de Correspondência Linear (LCA) de Kayne, derivando pares de c-comando assimétricos entre todos os terminais da árvore. Lembrando que o LCA faz uso da relação de dominância para ordenar terminais não diretamente relacionados por c-comando, de modo que, se um não-terminal A c-comanda um não-terminal B, então todos os terminais dominados por A c-comandam os terminais dominados por B. Assim, a partir da árvore acima, a interface PF poderia extrair os seguintes pares $\langle \alpha, \beta \rangle$, tal que α c-comanda β:

(44) ⟨the, room⟩, ⟨in, the⟩, ⟨in, room⟩, ⟨sees, Mary⟩, ⟨sees, {in, the, room}⟩, ⟨Mary, {in,

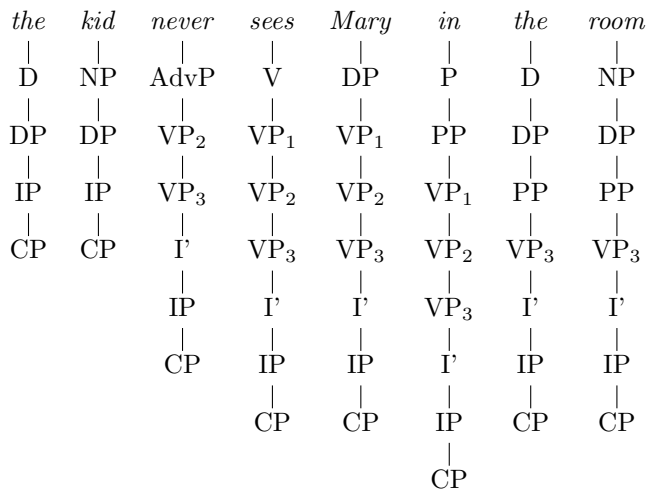
[2] Deixemos de lado – para manter o argumento – os problemas descritivos desta definição, como por exemplo o fato de que o I' (núcleo) c-comandaria o DP sujeito, visto que nessa representação não há especificador.

the, room}⟩, ⟨sees, never⟩, ⟨Mary, never⟩, ⟨never, {in, the, room}⟩, ⟨{never, sees, Mary, in, the, room}, {the, kid}⟩ e ⟨the, kid⟩

Porém, temos um problema com estes pares. Tomemos ⟨the, room⟩ e ⟨the, kid⟩, por exemplo. A qual instância de *the* no enunciado estes pares fazem referência? Sabemos que devem ser distintas, mas como PF poderia determinar, se as ocorrências de *the* apresentam todas o mesmo rótulo categorial? Mais dramático é o caso do penúltimo par acima, em que há uma instância de *the* (em *the room*) c-comandando outra (em *the kid*). PF não poderia fazer nada com isso, pois seria como dizer que A precede A (i.e., ele próprio). Fica claro, portanto, que os pares extraídos em PF devem fazer referência a algo mais do que apenas a informação presente no item terminal.

Talvez haja outras opções, mas um passo que me parece bastante natural é utilizar as relações de dominância, já previstas em estruturas sintáticas. Assim, ao invés de fazer referência apenas aos terminais, cada par poderia fazer referência ao terminal acrescido de cadeia de nós sintáticos que o dominam. Ou seja, *the* (em *the kid*) não seria referido apenas como D, mas sim como (CP-IP-DP-D), isto é, exatamente a lista de não-terminais que o domina. Já *the*, em *the room*, seria referido como (CP-IP-I'-VP_3-PP-DP-D), ou seja, um elemento bastante distinto do anterior. Aplicando isto a todos os pares, extrairíamos uma série de elementos corrrespondentes aos itens do enunciado, porém absolutamente distintos uns dos outros (as cadeias de dominância estão exibidas na ordem inversa):

(45)



Se PF puder acessar esta informação – que é apenas uma informação disponível na árvore sintática, ressalto – poderia estabelecer a ordem linear entre os elementos de cada par. Utilizando 0 para sucessão e 1 para precedência, poderia ir marcando os pares, parametricamente, utilizando as informações sobre ordem linear do enunciado. Tais parâmetros seriam utilizadas, posteriormente, para linearização de estruturas geradas na sintaxe.[3] Para cada enunciado o número de pares de c-comando é finito e pequeno. Além disso, mesmo quando se considera a infinitude de enunciados possíveis numa língua, a tendência é que os pares de c-comando também sejam finitos, visto que fazem referência às categorias sintáticas envolvidas, não aos itens lexicais específicos. Portanto, a quantidade de informação a ser armazenada é de um modo geral finita.

Eventualmente, PF teria que lidar com ordens conflitantes, isto é, pares configurados de um modo anteriormente, mas que aparecem numa ordem distinta noutro momento. Neste caso, duas situações são possíveis. Primeiramente, é possível que a ordem conflitante seja motivada por alguma categoria presente na estrutura, que provavelmente apareceria na cadeia de dominância de um dos (ou de ambos os) itens, o que tornaria o par único, anulando o conflito. Caso sejam cadeias exatamente iguais, neste caso o procedimento de aquisição da ordem teria que contar com algum mecanismo probabilístico capaz de vincular probabilidades

[3] Inclusive, poderiam ser utilizadas pelo analisador, se quiséssemos robustecê-lo com capacidade de análise preditiva (i.e., *top-down*).

para os valores 0 e 1 vinculados a um mesmo par. É também uma tarefa relativamente simples.

Por fim, nada impede, ainda, que PF faça generalizações sobre as configurações paramétricas, de modo a capturar regularidades que permitam ter conjuntos menores de parâmetros capazes de descrever a ordem linear da língua. Novamente, é uma questão de conceber um procedimento para este fim, baseado nas propriedades disponíveis na árvore sintática (categorias, dominância, etc.). Com isso, vê-se que é plenamente possível desenvolver o modelo de modo a lidar também com a ordem linear e de um modo linguisticamente relevante, pois a estratégia proposta aqui faz uso de relações próprias à representação sintática.

Mais que isso, este aparato tornaria possível lidar com variação de ordem sem a operação de movimento. O primeiro passo foi dado na concepção do modelo, ao assumir que o conhecimento sintático independe da ordem linear, isto é, que as representações sintáticas de enunciados equivalentes em diferentes línguas são as mesmas, a despeito de diferenças superficiais. Isto tem uma dupla consequência. Primeiramente, confere um caráter mais universal ao modelo, na medida em que a natureza do conhecimento sintático será ainda mais semelhante, seja qual for a língua em estudo.

Em segundo lugar, ao assumir que não há movimento para derivar variações de ordem linear, o modelo se coaduna com estudos sobre o processamento, que indicam não haver diferenças significativas de processamento para construções canônicas entre línguas que exibem diferentes ordens lineares (ver Erdocia et al., 2009, entre outros). Embora não seja absolutamente necessário que a operação de movimento tenha correlação com o custo de processamento, seria mais interessante se pudéssemos ter uma teoria sintática que faça previsões em relação ao processamento. Neste sentido, o modelo caminha nesta direção, visto que as estruturas sintáticas obtidas com as análises independem[4] da ordem linear.

[4] Exceto para adjuntos, como explicado no Capítulo 4.

6.2 Estágios intermediários da aquisição

Dadas as propriedades dos dados de entrada e dos procedimentos de aquisição, garante-se também que as gramáticas intermediárias do aprendiz serão sempre "naturais", isto é, estarão em conformidade com a GU (conforme assumida na modelagem), mesmo que sejam momentaneamente desviantes em relação à língua-alvo. Por exemplo, tomemos o caso dos desvios ocorridos na simulação com o corpus do inglês, em que alguns verbos incluíram a definitude de um de seus argumentos em sua entrada lexical. Tal característica aparece, por exemplo, no húngaro, em que a definitude do objeto é marcada nos verbos, de modo que verbos intransitivos recebem marca indefinida, enquanto verbos transitivos recebem a marca correspondente à definitude do objeto.

Vê-se, assim, que o modelo – mesmo quando se desvia da língua-alvo (no decorrer do percurso) – ainda exibe gramáticas intermediárias que se caracterizam como línguas naturais possíveis. O importante, para o aprendiz modelado, é que a combinação entre os itens lexicais adquiridos continue possível, o que depende apenas da satisfação dos seletores envolvidos. Desta forma, o IASMIM está de acordo – pelo menos para o que se propõe a aprender – com a hipótese de que os estágios intermediários exibidos pela criança no decorrer da aquisição são restringidos pela GU, sempre exibindo propriedades de línguas naturais possíveis (ver Clahsen, 1990, Poeppel & Wexler, 1993, Clahsen et al., 1993, entre outros).

6.3 O modelo e a teoria linguística

Este trabalho apresentou uma simulação computacional que envolveu a modelagem de um aprendiz da língua, cujas tarefas de aprendizagem consistiram em aprender o léxico (mapeamento entre palavras e sentidos) e, a partir daí, aprender regras gramaticais. O modelo incorporou várias noções e assunções da teoria sintática vigente e ainda uma versão da proposta de representação semântico-conceitual em Jackendoff (1990). Ao mesmo tempo, o modelo se destaca na medida em que não assume a operação de movimento e trata a ordem

linear como algo extrínseco ao conhecimento sintático.

Na verdade, ao integrar os processos de aquisição lexical e sintática e articular as propriedades da representação semântica com as propriedades da componente sintática, o modelo é capaz de capturar uma parte significativa do conhecimento gramatical, para a qual a operação de movimento é uma alternativa descritiva. O que o modelo faz, grosso modo, é "devolver" à semântica[5] a parte que lhe cabe, isto é, a que normalmente seria atribuída à estrutura profunda. Com isso, não apenas a sintaxe deixa de ter que recuperar níveis de representação distintos, mas obtém-se o benefício de descartar o uso de papéis temáticos explícitos nas representações, embora o efeito dos papéis esteja presente, bem como as correspondências mais ou menos estáveis entre posições na árvore sintática e posições na expressão conceitual.

Outra virtude do modelo foi não necessitar explicitar informações de subcategorização para itens lexicais. Tais informações são um desdobramento da aquisição lexical, na medida em que esta identifica itens lexicais insaturados, isto é, itens que contém seletores, no sentido conferido ao termo nesta modelagem. Os seletores não são atributos lexicais, mas sim dependências semânticas estruturais dos itens. Tais seletores são identificados por inspeção, quando os itens chegam para o processamento sintático e os objetos sintáticos correspondentes são criados. Uma vez identificados, os seletores irão determinar a ordem de composição da estrutura sintática no decorrer da análise.

Neste sentido, é possível sugerir que a aquisição lexical no modelo seja uma interpretação "aquisicional" para o conceito de *numeração* em Chomsky (1993, e posteriores), na medida em que a sintaxe é "cega" em relação ao modo como os itens são identificados e adquiridos, aguardando apenas que uma lista de elementos lhe seja enviada. Se a lista contiver elementos consistentes, uma estrutura sintática poderá ser formada, caso contrário, a análise irá falhar. Vê-se uma grande semelhança entre esta descrição e a descrição de

[5] No sentido dado em Jackendoff (op.cit.).

uma derivação sintática, que iniciaria com o processo pré-sintático de "montagem" dos itens lexicais que irão compor a numeração.

Não é sem razão, portanto, que afirmei anteriormente que o IASMIM é um modelo de aquisição voltado para aspectos da competência linguística. O modelo está em íntima relação com a teoria e permite avaliar empiricamente como alguns dos componentes da FL podem interagir de modo a permitir a aquisição da linguagem. Vale lembrar que uma das principais críticas aos modelos teóricos da gramática gerativa, por parte de pessoas interessadas em modelagens computacionais, tem sido a de que estes são, em grande medida, impossíveis de implementar e, neste sentido, não são computacionais.

Creio eu que os resultados da presente modelagem são uma evidência na direção contrária, isto é, mostram claramente que é possível aplicar a teoria em modelagens computacionais. O problema é que esta tarefa irá exigir de quem modela uma estudo aprofundado dos modelos teóricos, tanto para compreendê-los apropriadamente, quanto para lidar criticamente com eles, de modo a propor alternativas para lacunas e propriedades indesejáveis. É neste contexto que modelagens computacionais podem ser de grande valia para a teoria linguística.

Por fim, vale ressaltar que este modelo aponta para uma direção teórica distinta da que a teoria tem em geral seguido: aponta-se para uma divisão mais estrita entre semântica e sintaxe, de um lado, e entre sintaxe e morfologia/fonologia de outro. Esta separação deixa a sintaxe bastante "enxuta", visto que as relações temáticas são expressas no nível semântico (embora com reflexo na sintaxe) e as informações de ordem linear são expressas apenas na interface-PF (ou após ela), sem reflexos para a estrutura sintática produzida pelo analisador. Estas duas opções determinam o caráter não-transformacional do modelo. O quanto esta direção teórica pode produzir explicações mais adequadas que as atuais é uma questão a ser investigada.

Sua virtude mais saliente nesta modelagem foi a de permitir articular vários conceitos

da teoria sintática vigente com uma representação semântica independentemente motivada e uma teoria de aprendizagem explícita, que integrou aquisição lexical e sintática. Infelizmente, o escopo reduzido da aprendizagem no modelo impediu que aspectos mais complexos do conhecimento gramatical pudessem ser simulados e investigados, o que certamente conferiria um peso muito maior ao modelo face à teoria linguística vigente. De todo modo, uma visão restrita da componente sintática como a que foi assumida neste modelo pode ser um caminho promissor para demonstrar o papel central desta componente na cognição humana, isto é, talvez a importância desta componente não esteja na extensão de seu papel (junto aos demais módulos), mas sim no fato de ter um papel bastante restrito, mas *capital*, de modo que sem ela não seria possível à FL ter o grau de poder expressivo que apresenta.

6.4 Desenvolvimentos futuros

Há inúmeras possibilidades para desenvolvimentos posteriores sobre o modelo. Primeiramente, como discutido no capítulo anterior, seria o caso de melhorar sua performance, buscando meios de otimizar o processamento, para que o modelo fosse capaz de lidar com mais dados, especialmente em relação à aquisição da sintaxe. Afora isto, o modelo precisaria ser desenvolvido para lidar com omissão de elementos (sujeito e objeto nulo e elipses). Tanto o algoritmo de aquisição lexical, quanto o de aquisição sintática são atualmente limitados neste aspecto, embora um primeiro passo tenha sido dado, ao flexibilizar o reconhecimento lexical, de modo que os itens não precisam exaurir o contexto.

Ampliar a abrangência gramatical é, também, fundamental. O primeiro grande desafio para tornar o modelo capaz de lidar com quantificadores, relações anafóricas, topicalizações, relativas, foco, negação, etc., é desenvolver a representação semântico-conceitual assumida, visto que hoje ela não dispõe de recursos para isso. Aliás, mesmo o aparato atual pode ser revisto, pois há assunções bastante simplificadas, como por exemplo a representação assumida para aspectos de tempo (e aspecto) verbal. Um sistema mais robusto, inclusive envolvendo algum tipo de álgebra própria, de modo a explorar ainda mais os efeitos da

composicionalidade[6], seria bastante bem-vindo. Neste caso, adaptações teriam que ser feitas também no analisador, para que pudesse lidar com uma estrutura semântica mais complexa. É provável, ainda, que novos tipos de regras sintáticas ou de restrições precisassem ser criadas.

Outro caminho de desenvolvimento, diz respeito a repensar a capacidade de processamento do modelo, de modo que este possa fazer análises na ausência de uma representação do contexto. Para isso, o analisador teria que contar com recursos de análise preditiva (*top-down*), *backtracking* para recorrer de escolhas incorretas (como em sentenças *garden-path*) e preferências de processamento, como a de *minimal-attachment*, entre outras. Vale ressaltar que, a priori, as mesmas regras gramaticais adquiridas seriam utilizadas. A diferença estaria nos recursos de que disporia o analisador para fazer um uso mais robusto destas regras.

Outro objetivo interessante seria tornar o aprendiz capaz de aprender relações de concordância. Um módulo "morfológico", pós-sintático e anterior ao processamento da ordem, poderia ser acoplado ao modelo e seu papel – entre outras coisas – seria o de registrar, para cada item lexical, possíveis relações com outros elementos na árvore sintática. À medida que a experiência avançasse, um mecanismo de aprendizagem trans-situacional poderia identificar se há e quais seriam tais elementos e quais atributos estariam envolvidos. Lembrando que a sintaxe, no modelo, opera com a categoria sintática, não com o item lexical em si. A sintaxe é, portanto, cega à forma específica da palavra.

Por exemplo, suponha que a categoria respectiva a um verbo transitivo (no português) chegasse a este módulo. Ali, as informações da categoria seriam acrescidas da lista de elementos (da árvore) que co-ocorrem com o verbo. A partir de outras experiências, os elementos ou atributos não recorrentes iriam sendo podados, até que apenas os itens e atributos relevantes permanecessem. É possível que este procedimento identificasse os atributos de concordância com o sujeito, no caso do PB. Provavelmente seria necessário assumir um conceito mais

[6] Por exemplo, para explorar o modo como a composição de certos verbos, argumentos e adjuntos afeta a telicidade da sentença.

explícito de c-comando no modelo, mas à primeira vista parece algo viável.

Por fim, é certamente desejável que outros corpora sejam submetidos ao modelo, se possível para línguas com características bastante distintas das analisadas aqui. Isso permitiria identificar limitações outras que poderiam comprometer o potencial translinguístico do modelo e fazer as correções, adaptações ou melhorias necessárias.

Referências

Allen, S. E. M. & Crago, M. B. (1996). Early passive acquisition in inuktitut. *Journal of Child Language*, 23(1):129–155. 31

Altmann, G. & Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition*, 30(3):191 – 238. 47

Anderson, J. R. (1978). Computer simulation of language acquisition: A second report. In LaBerge, D. & Samuels, S. J., editors, *Perception and Comprehension*. Erlbaum, Hillsdale, NJ. 24, 45, 75

Angluin, D. (1982). Inference of reversible languages. *Journal of the Association for Computing Machinery*, 29(3):741–765. 88

Bach, E. (1986). The algebra of events. *Linguistics and philosophy*, 9(1):5–16. 109

Bertolo, S. (2001). A brief overview of learnability. In Bertolo, S., editor, *Language acquisition and learnability*, pages 1–14. Cambridge University Press. 43

Berwick, R. C. (1985). *The Acquisition of Syntactic Knowledge*. The MIT Press, Massachusetts. xxi, xxiii, 2, 3, 48, 52, 53, 59, 61, 63, 65, 66, 67, 68, 69, 70, 73, 75, 78, 79, 80, 91, 95, 121, 122, 123, 136, 148, 168, 175

Berwick, R. C., Abney, S. P., & Tenny, C., editors (1992). *Principle-Based Parsing: Computation and Psycholinguistics*, volume 44. Springer Netherlands, Boston/Dordrecht/London. 43

Berwick, R. C., Pietroski, P., Yankama, B., & Chomsky, N. (2011). Poverty of the stimulus revisited. *Cognitive Science*, 35:1207–1242. 17, 18, 20, 21

Berwick, R. C. & Pilato, S. (1987). Learning syntax by automata induction. *Machine Learning*, 2(1):9–38. 88, 90, 92

Berwick, R. C. & Weinberg, A. S. (1984). *The Grammatical Basis of Linguistic Performance: Language Use and Acquisition*. The MIT Press, Cambridge, Massachusetts. 41, 42

Braine, M. D. (1992). What sort of innate structure is needed to "bootstrap" into syntax? *Cognirion*, 45:77–100. 24

Bresnan, J. (2001). *Lexical-functional syntax*. Blackwell Publishers Ltd, Oxford, UK. 41, 73, 104

Referências

Broeder, P. & Murre, J., editors (2000). *Models of Language Acquisition: Inductive and Deductive Approaches*. Oxford University Press, New York. 92

Brown, R. (1973). *A First Language: The Early Stages*. Allen & Unwin, London. 29, 30, 77

Cameron-Faulkner, T., Lieven, E., & Tomasello, M. (2003). A construction based analysis of child directed speech. *Cognitive Science*, 27:843—873. xxiii, 152, 153, 154

Chomsky, N. (1956). Three models for the description of language. *IEEE Transactions on Information Theory*, 2(3):113–124. 39

Chomsky, N. (1957). *Syntactic Structures*. de Gruyter Mouton. 11, 39, 40

Chomsky, N. (1959). On certain formal properties of grammars. *Information and Control*, 2(2):137–167. 3, 23, 33, 34, 35, 36, 39

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT Press. 18, 22, 23, 24, 59, 67

Chomsky, N. (1968). Language and mind. 18, 19, 21, 96

Chomsky, N. (1970). Remarks on nominalizations. In Jacobs, R. & Rosembaum, P., editors, *Readings in English Transformational Grammar*, chapter 12, pages 184–221. Ginn and Company, Waltham, Mass. 112

Chomsky, N. (1981). Lectures on government and binding. 97

Chomsky, N. (1986). *Knowledge of Language: Its Nature, Origin and Use*. Praeger, New York. vii, ix, 1, 10, 11, 18, 19, 25, 30, 40, 42, 67

Chomsky, N. (1993). A minimalist program for linguistic theory. In Hale, K. & Keyser, S. J., editors, *The View from Building 20 – Essays in Linguistics in Honor of Sylvain Bromberger*, pages 1–52. The MIT Press. 4, 67, 91, 115, 185

Chomsky, N. (1995a). Bare phrase structure. In Webelhuth, G., editor, *Government and Binding Theory and the Minimalist Program*, pages 383–439. Blackwell, Oxford. 4, 68, 96

Chomsky, N. (1995b). *The Minimalist Program*. The Mit Press. vii, ix, 40, 97

Chomsky, N. (1998). *Minimalist Inquires: The Framework*. Number 15 in MIT Occasional Papers in Linguistics. The MIT Press. vii, ix

Chomsky, N. (2004). Beyond explanatory adequacy. In Belletti, A., editor, *The Cartography of Syntactic Structures*, chapter 3, pages 104–131. Oxford University Press, Oxford. 135

Chomsky, N. (2005). Three factors in language design. *Linguistic Inquiry*, 36(1):1–22. 18

Christiansen, M. H. & Chater, N. (1999). Connectionist natural language processing: The state of the art. *Cognitive Science*, 23(4):417–437. 92

Cinque, G. (1999). *Adverbs and functional heads: a cross-linguistic perspective*. Oxford, New York. 67, 91

REFERÊNCIAS

Clahsen, H. (1990). Constraints on parameter setting: A grammatical analysis of some acquisition stages in german child language. *Language Acquisition*, 1(4):pp. 361–391. 184

Clahsen, H., Penke, M., & Parodi, T. (1993). Functional categories in early child german. *Language Acquisition*, 3(4):pp. 395–429. 184

Clifton, C., Speer, S., & Abney, S. P. (1991). Parsing arguments: Phrase structure and argument structure as determinants of initial parsing decisions. *Journal of Memory and Language*, 30:251–271. 25, 47

Corrêa, L. M. S. (2008). Relação processador lingüístico-gramática em perspectiva: Problemas de unificação em contexto minimalista. *D.E.L.T.A.*, 24(2):231–282. 25, 26

Crain, S. & Nakayama, M. (1987). Structure dependence in grammar formation. *Language*, 63(3):522–543. 20

Culicover, P. W. & Wexler, K. (1980). *Formal Principles of Language Acquisition*. The MIT Press, Cambridge, Massachusetts. 46, 61, 90, 122

Dowty, D. (1991). Thematic proto-roles and argument selection. *Language*, 67(3):547–619. 68, 99

Erdocia, K., Laka, I., Mestres-Missée, A., & Rodriguez-Fornells, A. (2009). Syntactic complexity and ambiguity resolution in a free word order language: Behavioral and electrophysiological evidences from basque. *Brain and Language*, 109(1):1–17. 183

Faria, P. (2009). *Propriedades das línguas naturais e o processo de aquisição: reflexões a partir da implementação do modelo em Berwick (1985)*. Mestrado, Universidade de Campinas (UNICAMP), Campinas, SP, Brasil. 3, 95, 122

Filip, H. (1999). *Aspect, Eventuality Types and Nominal Reference*. Routledge, Taylor & Francis Group (Garland), New York. 80, 99, 104, 108, 109, 113

Fisher, C., Hall, D. G., Rakowitz, S., & Gleitman, L. (1994). When it is better to receive than to give: Syntactic and conceptual constraints on vocabulary growth. *Lingua*, 92:333–375. 84

Frank, M. C. (2011). Computational models of early language acquisition. Online. URL: http://langcog.stanford.edu/papers/F-underreview-b.pdf. Acesso em 30/06/2013. 16, 17, 51, 54, 92

Frazier, L. & Fodor, J. D. (1978). The sausage machine: a new two-stage parsing. *Cognition*, 6:291–325. 47

Gaylard, H. L. (1995). *Phrase Structure in a Computational Model of Child Language Acquisition*. PhD thesis, University of Birmingham. xxi, 2, 24, 52, 53, 65, 70, 74, 77, 78, 81, 91, 92, 96, 136, 148, 168, 175

Referências

Gazdar, G. (1983). Phrase structure grammars and natural languages. In *Proceedings of the Eighth international joint conference on Artificial intelligence - Volume 1*, IJCAI'83, pages 556–565, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. 40

Gazdar, G. (1988). Applicability of indexed grammars to natural languages. In Reyle, U. & Rohrer, C., editors, *Natural Language Parsing and Linguistic Theories*, volume 35, pages 69–94. Springer Netherlands. 35

Gazdar, G., Klein, E., Pullum, G. K., & Sag, I. A. (1985). *Generalized phrase structure grammar*. Blackwell, Oxford. 67, 104

Gibson, E. & Wexler, K. (1994). Triggers. *Linguistic Inquiry*, 25:407–454. 2

Gold, E. M. (1967). Language identification in the limit. *Information and Control*, 10(5):447–474. 3, 43, 45, 50, 88

Gropen, J., Pinker, S., Hollander, M., & Goldberg, R. (1991). Affectedness and direct objects: The role of lexical semantics in the acquisition of verb argumentstructure. *Cognition*, 41:153–195. 24

Guasti, M. T. (2002). *Language acquisition: a linguistic perspective*. The MIT Press, Cambridge, Mass. xxiii, 2, 14, 15, 16, 17, 23, 24, 27, 28, 30, 31

Hammond, M. (2010). Introduction to the mathematics of language. U. of Arizona. Course Website http://dingo.sbs.arizona.edu/ hammond/ling501-f12/. xxi, 32, 37, 38, 39

Heinz, J. (2010). Computational theories of learning and developmental psycholinguistics. In Lidz, J. & Pater, J., editors, *The Cambridge Handbook of Developmental Linguistics*. Cambridge University Press. 43, 45

Hoff-Ginsberg, E. (1986). Function and structure in maternal speech: Their relation to the child's development of syntax. *Developmental Psychology*, 22(2):155–163. xxiii, 151, 152, 154

Hopcroft, J., Motwani, R., & Ullman, J. D. (2001). *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley, USA, second edition edition. 32, 37, 39

Ingram, D. (1989). *First language acquisition: method, description and explanation*. Cambridge University Press, New York. xxiii, 2, 12, 13, 14, 29

Israel, M., Johnson, C., & Brooks, P. J. (2000). From states to events: The acquisition of english passive participles. *Cognitive Linguistics*, 11(1/2):103–129. 154

Jackendoff, R. (1983). *Semantics and Congnition*, volume 8. MIT press, Cambridge, Massachusetts. 3, 82, 99, 104, 107, 113

Jackendoff, R. (1990). *Semantic structures*, volume 18. The MIT Press, Cambridge, Massachusetts. vii, ix, 3, 99, 104, 105, 107, 113, 184

REFERÊNCIAS

Jain, S., Osherson, D. N., Royer, J. S., & Sharma, A. (1999). *Systems That Learn: An Introduction to Learning Theory*. Learning, Development, and Conceptual Change. Bradford Book, 2nd edition edition. 51

Joshi, A. (2003). Mildly context-sensitive grammars. In Frawley, W. & Bright, W., editors, *International encyclopedia of linguistics*. Oxford University Press, 2nd edition. 45

Joshi, A. K., Shanker, K. V., & Weir, D. (1990). The convergence of mildly context-sensitive grammar formalisms. Technical reports (cis), University of Pennsylvania. 40, 41, 42

Jurafsky, D. & Martin, J. H. (2008). *Speech and Language Processing*. Pearson Prentice Hall, 2 edition. 70

Jusczyk, P. W. (1997). *The discovery of spoken language*. The MIT Press, Cambridge, MA. 14, 27, 45

Kaiser, E. & Trueswell, J. C. (1994). The role of discourse context in the processing of a flexible word-order language. *Cognition*, 94:113–147. 25, 47

Kaplan, F., Oudeyer, P.-Y., & Bergen, B. (2008). Computational models in the debate over language learnabiliy. *Infant and Child Development*, 17(1):55–80. 16, 43, 51, 92

Kayne, R. S. (1981). Unambiguos paths. In May, R. & Koster, J., editors, *Levels of Syntactic Representations*. Reidel. 91

Kayne, R. S. (1994). *The Antisymmetry of Syntax*. MIT Press, Cambridge, MA. 180

Kornai, A. & Pullum, G. K. (1990). The x-bar theory of phrase structure. *Language*, 66(1):pp. 24–50. 36

Langley, P. (1982). A model of early syntactic development. pages 146–151. 5, 52, 55

Langley, P. (1987). Machine learning and grammar induction. *Machine Learning*, 2(1):5–8. 46

Levelt, W. J. (2008). *An introduction to the theory of formal languages and automata*. John Benjamins Publishing Company, Amsterdam / Philadelphia. 32, 33, 39, 43

Lidz, J., Waxman, S., & Freedman, J. (2003). What infants know about syntax but couldn't have learned: experimental evidence for syntactic structure at 18 months. *Cognition*, 89:B65–B73. 21

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). Lexical nature of syntactic ambiguity resolution. *Psychological Review*, 101(4):676–703. 25

Marcus, G. F. (1993). Negative evidence in language acquisition. *Cognition*, 46:53–85. 23

Marcus, G. F. (1995). The acquisition of the english past tense in children and multilayered connectionist networks. *Cognition*, 56(3):271 – 279. 17

Referências

Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., Xu, F., & Clahsen, H. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development*, 57(4). 78

Marcus, M. P. (1980). *A theory of syntactic recognition for natural language*. MIT Press, Cambridge, Mass. 48, 59

Mazuka, R. (1998). *The Development of Language Processing Strategies: a cross-linguistic study between Japanese and English*. Lawrence Erlbaum Associates, Publishers, Mahwah, New Jersey. 45, 67

Mohri, M. & Sproat, R. (2006). On a common fallacy in computational linguistics. *SKY Journal of Linguistics*, 19:432–439. 40

Nowak, M. A., Komarova, N. L., & Niyogi, P. (2002). Computational and evolutionary aspects of language. *Nature*, 417(6889):611–617. 2, 43

Parker, M. D. & Brorson, K. (2005). A comparative study between mean length of utterance in morphemes (mlum) and mean length of utterance in words (mluw). *First Language*, 25(3):365–376. 156

Partee, B. H., ter Meulen, A., & Wall, R. E. (1993). *Mathematical Methods in Linguistics*. Kluwer-Dordrecht, Boston/London. 32, 39, 45

Pearl, L. (2010). Using computational modeling in language acquisition research. In Blom, E. & Unsworth, S., editors, *Experimental Methods in Language Acquisition Research*. John Benjamins. 9, 43, 49, 50, 53

Pearl, L. & Lidz, J. (2009). When domain-general learning fails and when it succeeds: Identifying the contribution of domain specificity. *Language Learning and Development*, 5(4):235–265. 21

Pearl, L. & Sprouse, J. (2011). Syntactic islands without universal grammar: A computational model of the acquisition of constraints on long-distance dependencies. 2, 89, 92

Peters, P. S. & Ritchie, R. W. (1971). On restricting the basecomponent of transformational grammars. *Information and Control*, 18:483–501. 40

Peters, P. S. & Ritchie, R. W. (1973). On the generative power of transformational grammars. *Information Sciences*, 6:49–83. 40

Pinker, S. (1979). Formal models of language learning. *Cognition*, 7:217–283. 43, 45, 53

Pinker, S. (1984). *Language Learnability and Language Development*. Reprinted version, 1996. Harvard University Press, Cambridge, MA. 16, 24, 48, 68, 92

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. MIT press, Cambridge, Massachusetts. 3, 84, 99, 104, 106, 107, 108, 110, 113

Pinker, S. (2004). Clarifying the logical problem of language acquisition. *Journal of Child Language*, 31(4):949–953. 21, 23

Poeppel, D. & Wexler, K. (1993). The full competence hypothesis of clause structure in early german. *Language*, 69(1):pp. 1–33. 184

Pollard, C. & Sag, I. A. (1994). *Head-Driven Phrase Structure Grammar*. University of Chicago Press, Chicago. 41

Pollock, J.-Y. (1989). Verb movement, universal grammar, and the structure of IP. *Linguistic Inquiry*, 20:365–424. 67

Pullum, G. K. (1996). Learnability, hyperlearning, and the poverty of the stimulus. In Johnson, J., Juge, M. L., & Moxley, J. L., editors, *Proceedings of the Twenty-Second Annual Meeting of the Berkeley Linguistics Society: General Session and Parasession on The Role of Learnability in Grammatical Theory*, pages 498–513. Berkeley, California: Berkeley Linguistics Society. 20, 21

Pullum, G. K. & Gazdar, G. (1982). Natural languages and context-free languages. *Linguistics and Philosophy*, 4(4):pp. 471–504. 40

Pye, C. & Poz, P. Q. (1988). Precocious passives (and antipassives) in quiche mayan. *Papers and Reports on Child Language Development*, 27:71–80. 31

Rubin, M. C. d. B. P. (2004). *A Passiva na Síndrome de Down*. Tese (doutorado em linguistica), Universidade Federal do Paraná. 31

Saxton, M. (2000). Negative evidence and negative feedback: immediate effects on the grammaticality of child speech. *First Language*, 20:221–252. 23

Saxton, M., Backley, P., & Gallaway, C. (2005). Negative input for grammatical errors: effects after a lag of 12 weeks. *Journal of Child Language*, 32(3):643–672. 23

Seidenberg, M. S. (1997). Language acquisition and use: Learning and applying probabilistic constraints. *Science*, 275(14):1599–1603. 16, 17, 51, 92

Selfridge, M. (1986). A computer model of child language learning. *Artificial Intelligence*, 29:171–216. 52, 56

Shieber, S. M. (1985). Evidence against the context-freeness of natural language. *Linguistics and Philosophy*, 8:333–343. 40

Shieber, S. M. (1986). *An Introduction to Unification-Based Approaches to Grammar*. Microtome Publishing, Brookline, Massachusetts, 3rd, 2003 edition. 3, 99, 102

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, 61(1):39–91. 53, 81, 84, 86, 92, 104, 114, 119, 124, 125, 128, 130, 131, 132, 157, 163, 176

Referências

Skinner, B. F. (1957). *Verbal Behavior*. The Century Psychology Series. Appleton-Century-Crofst, Inc., New York. 10, 14, 16

Sprouse, J., Wagers, M., & Phillips, C. (2012). A test of the relation between working-memory capacity and syntactic island effects. *Language*, 88(1):82–123. 90

Steedman, M. (2000). *The Syntactic Process*. MIT Press, Cambridge. 67

Steedman, M. & Baldridge, J. (2011). Combinatory categorial grammar. In Borsley, R. D. & Börjars, K., editors, *Non-Transformational Syntax: Formal and Explicit Models of Grammar*, pages 181–224. Blackwell Publishing Ltd. 39, 41

Uriagereka, J. (1999). Multiple spell-out. In Epstein, S. D. & Hornstein, N., editors, *Working Minimalism*, pages 251–282. The MIT Press, Cambridge. vii, ix, 4, 96

Valiant, L. G. (1984). A theory of the learnable. *Commun. ACM*, 27:1134–1142. 44

Vijay-Shanker, K. & Weir, D. J. (1994). The equivalence of four extensions of context-free grammars. *Mathematical Systems Theory*, 27:27–511. 39, 41

Villavicencio, A. (2002). *The acquisition of a unification-based generalised categorial grammar*. Doctoral dissertation, University of Cambridge. xii, xiii, 2, 52, 53, 85, 87, 92, 176

Wintner, S. (2002). Formal language theory for natural language processing. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 71–76, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics. 32

Wolff, J. G. (1975). An algorithm for the segmentation of an artificial language analogue. *British Journal of Psychology*, 66(1):79–90. 52, 53

Yang, C. (2002). *Knowledge and learning in natural language*. Oxford University Press. 2, 42, 43, 52

Yang, C. (2011). Computational models of syntactic acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*. 16, 43, 50, 51, 53, 92

Apêndice A

Exemplo de gramática adquirida

A.1 Nota preliminar

Conjunto de regras gramaticais adquiridas numa simulação envolvendo o corpus mínimo. As regras estão ordenadas por tipo (regras ANEXE e depois COMBINE) e por nome. Os valores "C" (indicados após o nome das regras) se referem a quantidade de vezes que a regra foi aplicada em análises. A entrada "Sense" (em regras do tipo COMBINE) indica a natureza do sintagma a ser formado pela regra de combinação. As indicações "head-s" e "head-p" em regras do tipo ANEXE indicam, respectivamente, anexação de núcleo em *merge* por substituição ("set-Merge") e por adjunção ("pair-Merge"). Finalmente, o símbolo "@" que acompanham certos atributos indicam subespecificação.

A.2 A gramática

```
------
Regras do tipo ANEXE
------

Regra: A0004 (C:12)
CYC: None, ACT: N0122, (adjuncts)
Células (símbolo de entrada: posição 1):
 • (1) N0013, [a=[cf=[f=[+animacy, person=1, -plural, +quantizable, -wh],
               m=[]], f=[+definite], m=[]], f=[relation=at-poss], m=[]]
   (2) *
   (3) None

Regra: A0005 (C:32)
CYC: *, ACT: N0122, (head-p)
Células (símbolo de entrada: posição 1):
 • (1) N0122, [f=[animacy=@animacy, concept=@concept, person=3,
               plural=@plural, quantizable=@quantizable, -wh], m=@m]
   (2) *
   (3) *
```

Regra: A0011 (C:191)
CYC: *, **ACT:** N0312, (comp)
Células (símbolo de entrada: posição 1):
• (1) N0122, [f=[animacy=@animacy, concept=@concept, person=3,
plural=@plural, quantizable=@quantizable, -wh], m=@m]
(2) *
(3) *

Regra: A0016 (C:44)
CYC: N0312, **ACT:** N0122, (head-p)
Células (símbolo de entrada: posição 2):
(1) [f=[concept=@concept, -wh], m=[]]
• (2) N0122, [f=[animacy=@animacy, concept=@concept, person=3,
plural=@plural, quantizable=@quantizable, -wh], m=@m]
(3) *

Regra: A0024 (C:11)
CYC: None, **ACT:** N0122, (head-p)
Células (símbolo de entrada: posição 2):
(1) [a=[cf=[f=[+animacy, person=1, -plural, +quantizable, -wh],
m=[]], f=[+definite], m=[]], f=[relation=at-poss], m=[]]
• (2) N0122, [f=[animacy=@animacy, concept=@concept, person=3,
plural=@plural, quantizable=@quantizable, -wh], m=@m]
(3) None

Regra: A0055 (C:3)
CYC: N0312, **ACT:** N0177, (head-s)
Células (símbolo de entrada: posição 1):
• (1) N0171, [cf=[], f=[concept=former, -wh], m=[]]
(2) [f=[+animacy, concept=@concept, person=3, plural=@plural,
+quantizable, -wh], m=[]]
(3) None

Regra: A0056 (C:3)
CYC: N0312, **ACT:** N0177, (comp)
Células (símbolo de entrada: posição 1):
• (1) N0122, [f=[animacy=@animacy, concept=@concept, person=3,
plural=@plural, quantizable=@quantizable, -wh], m=@m]
(2) None
(3) None

Regra: A0057 (C:3)
CYC: None, **ACT:** N0312, (comp)
Células (símbolo de entrada: posição 1):
• (1) N0177, [cf=[f=[+animacy, concept=@concept, person=3,
plural=@plural, +quantizable, -wh], m=[]],
f=[concept=former, -wh], m=[]]
(2) None
(3) None

Regra: A0060 (C:3)
CYC: N0312, **ACT:** N0122, (head-p)
Células (símbolo de entrada: posição 1):
• (1) N0122, [f=[animacy=@animacy, concept=@concept, person=3,

```
              plural=@plural, quantizable=@quantizable, -wh], m=@m]
   (2) [a=[], f=[relation=at-poss], m=[]]
   (3) [cf=[], f=[-definite], m=[]]

Regra: A0061 (C:27)
CYC: N0122, ACT: N0197, (head-s)
Células (símbolo de entrada: posição 1):
 • (1) N0179, [a=[], f=[relation=at-poss], m=[]]
   (2) [f=@f, m=[]]
   (3) *

Regra: A0064 (C:23)
CYC: N0122, ACT: N0197, (comp)
Células (símbolo de entrada: posição 1):
 • (1) N0312, [cf=[f=[animacy=@animacy, concept=@concept,
              person=3, plural=@plural, quantizable=@quantizable,
              -wh], m=[]], f=[definite=@definite], m=[]]
   (2) None
   (3) None

Regra: A0065 (C:25)
CYC: N0312, ACT: N0122, (adjuncts)
Células (símbolo de entrada: posição 1):
 • (1) N0197, [a=[cf=[f=[+animacy, concept=@concept, person=3,
              plural=@plural, +quantizable, -wh], m=@m],
              f=[definite=@definite], m=[]], f=[relation=at-poss], m=[]]
   (2) None
   (3) None

Regra: A0075 (C:51)
CYC: N0312, ACT: N0122, (adjuncts)
Células (símbolo de entrada: posição 1):
 • (1) N0024, [f=[concept=@concept, -wh], m=[]]
   (2) *
   (3) *

Regra: A0077 (C:174)
CYC: *, ACT: N0312, (head-s)
Células (símbolo de entrada: posição 1):
 • (1) N0141, [cf=[], f=[definite=@definite], m=[]]
   (2) [f=[concept=@concept, -wh], m=[]]
   (3) *

Regra: A0093 (C:4)
CYC: N0122, ACT: N0197, (comp)
Células (símbolo de entrada: posição 1):
 • (1) N0220, [+*abs*, cf=[f=[+animacy, concept=@concept,
              person=3, +plural, +quantizable, -wh], m=[]],
              f=[-definite], m=[]]
   (2) None
   (3) None

Regra: A0094 (C:2)
CYC: N0312, ACT: N0122, (adjuncts)
```

Células (símbolo de entrada: posição 1):
• (1) N0222, [a=[+*abs*, cf=[f=[+animacy, concept=@concept,
person=3, +plural, +quantizable, -wh], m=[]],
f=[-definite], m=[]], f=[relation=at-poss], m=[]]
(2) None
(3) None

Regra: A0104 (C:2)
CYC: None, **ACT:** N0312, (comp)
Células (símbolo de entrada: posição 1):
• (1) N0122, [f=[animacy=@animacy, concept=@concept, person=3,
plural=@plural, quantizable=@quantizable, -wh], m=@m]
(2) [cf=[cf=[a=[], cf=[], f=[-change], m=[]], f=[-finite,
+realis], m=[]], f=[force=decl], m=[]]
(3) [f=[concept=red, -wh], m=[]]

Regra: A0105 (C:11)
CYC: None, **ACT:** N0319, (head-s)
Células (símbolo de entrada: posição 2):
(1) [cf=[f=[animacy=@animacy, concept=@concept, person=3, -plural,
quantizable=@quantizable, -wh], m=[]], f=[definite=@definite],
m=[]]
• (2) N0231, [cf=[cf=[a=[], cf=[], f=[-change], m=[]], f=[-finite,
+realis], m=[]], f=[force=decl], m=[]]
(3) [f=[concept=@concept, -wh], m=[]]

Regra: A0115 (C:23)
CYC: None, **ACT:** N0319, (comp)
Células (símbolo de entrada: posição 2):
(1) [cf=[f=[animacy=@animacy, concept=@concept, person=3, -plural,
quantizable=@quantizable, -wh], m=[]], f=[definite=@definite],
m=[]]
• (2) N0024, [f=[concept=@concept, -wh], m=[]]
(3) None

Regra: A0116 (C:26)
CYC: None, **ACT:** N0278, (head-s)
Células (símbolo de entrada: posição 1):
• (1) N0319, [cf=[cf=[a=@a, cf=[f=[concept=@concept, -wh], m=[]],
f=[change=@change], m=[]], f=[finite=@finite, +realis],
m=[]], f=[force=@force], m=[]]
(2) [f=@f, m=[]]
(3) None

Regra: A0131 (C:2)
CYC: None, **ACT:** N0319, (comp)
Células (símbolo de entrada: posição 3):
(1) [cf=[f=[+animacy, concept=mary, person=3, -plural, +quantizable,
-wh], m=[]], f=[+definite], m=[]]
(2) [f=[concept=often, +iteration, -location, -wh], m=[]]
• (3) N0024, [f=[concept=@concept, -wh], m=[]]

Regra: A0134 (C:4)
CYC: None, **ACT:** N0251, (head-p)

Células (símbolo de entrada: posição 1):
• (1) N0278, [cf=[cf=[a=[cf=[f=[animacy=@animacy, concept=@concept,
person=3, -plural, quantizable=@quantizable, -wh], m=[]],
f=[definite=@definite], m=[]], cf=[f=[concept=@concept,
-wh], m=[]], f=[-change], m=[]], f=[finite=@finite, +realis],
m=[]], f=[force=@force], m=[]]
(2) [f=[concept=often, +iteration, -location, -wh], m=[]]
(3) None

Regra: A0135 (C:4)
CYC: None, **ACT:** N0251, (adjuncts)
Células (símbolo de entrada: posição 1):
• (1) N0250, [f=[concept=often, +iteration, -location, -wh], m=[]]
(2) None
(3) None

Regra: A0137 (C:1)
CYC: None, **ACT:** N0319, (comp)
Células (símbolo de entrada: posição 3):
(1) [cf=[f=[-animacy, concept=car, person=3, -plural, +quantizable,
-wh], m=[]], f=[-definite], m=[]]
(2) [f=[concept=often, +iteration, -location, -wh], m=[]]
• (3) N0024, [f=[concept=@concept, -wh], m=[]]

Regra: A0138 (C:24)
CYC: None, **ACT:** N0278, (comp)
Células (símbolo de entrada: posição 1):
• (1) N0312, [cf=[f=[animacy=@animacy, concept=@concept, person=3,
plural=@plural, quantizable=@quantizable, -wh], m=[]],
f=[definite=@definite], m=[]]
(2) *
(3) *

Regra: A0139 (C:1)
CYC: None, **ACT:** N0319, (head-s)
Células (símbolo de entrada: posição 3):
(1) [f=[concept=often, +iteration, -location, -wh], m=[]]
(2) [cf=[f=[+animacy, concept=mary, person=3, -plural, +quantizable,
-wh], m=[]], f=[+definite], m=[]]
• (3) N0231, [cf=[cf=[a=[], cf=[], f=[-change], m=[]], f=[-finite,
+realis], m=[]], f=[force=decl], m=[]]

Regra: A0140 (C:1)
CYC: None, **ACT:** N0319, (comp)
Células (símbolo de entrada: posição 3):
(1) [f=[concept=often, +iteration, -location, -wh], m=[]]
(2) [cf=[f=[+animacy, concept=mary, person=3, -plural, +quantizable,
-wh], m=[]], f=[+definite], m=[]]
• (3) N0024, [f=[concept=@concept, -wh], m=[]]

Regra: A0141 (C:1)
CYC: None, **ACT:** N0278, (head-s)
Células (símbolo de entrada: posição 2):
(1) [f=[concept=often, +iteration, -location, -wh], m=[]]

- (2) N0319, [cf=[cf=[a=@a, cf=[f=[concept=@concept, -wh], m=[]],
  f=[change=@change], m=[]], f=[finite=@finite, +realis],
  m=[]], f=[force=@force], m=[]]
  (3) [cf=[f=[+animacy, concept=mary, person=3, -plural, +quantizable,
  -wh], m=[]], f=[+definite], m=[]]

Regra: A0142 (C:1)
CYC: None, **ACT:** N0278, (comp)
Células (símbolo de entrada: posição 2):
  (1) [f=[concept=often, +iteration, -location, -wh], m=[]]
- (2) N0312, [cf=[f=[animacy=@animacy, concept=@concept, person=3,
  plural=@plural, quantizable=@quantizable, -wh], m=[]],
  f=[definite=@definite], m=[]]
  (3) None

Regra: A0143 (C:5)
CYC: None, **ACT:** N0319, (head-s)
Células (símbolo de entrada: posição 2):
  (1) [cf=[f=[-animacy, concept=@concept, person=3, -plural, +quantizable,
  -wh], m=[]], f=[definite=@definite], m=[]]
- (2) N0259, [cf=[cf=[a=[], cf=[], f=[-change], m=[]], f=[+finite, +realis],
  m=[]], f=[force=decl], m=[]]
  (3) [f=[concept=@concept, -wh], m=[]]

Regra: A0155 (C:2)
CYC: None, **ACT:** N0278, (comp)
Células (símbolo de entrada: posição 1):
- (1) N0063, [cf=[f=[-animacy, concept=ball, person=3, -plural, +quantizable,
  -wh], m=[]], f=[-definite], m=[]]
  (2) None
  (3) None

Regra: A0156 (C:7)
CYC: None, **ACT:** N0319, (head-s)
Células (símbolo de entrada: posição 1):
- (1) N0270, [cf=[cf=[a=[], cf=[], f=[-change], m=[]], f=[-finite, +realis],
  m=[]], f=[force=int], m=[]]
  (2) [cf=@cf, f=[+definite], m=[]]
  (3) [f=[concept=@concept, -wh], m=[]]

Regra: A0160 (C:3)
CYC: None, **ACT:** N0319, (head-s)
Células (símbolo de entrada: posição 1):
- (1) N0284, [cf=[cf=[a=[], cf=[], f=[-change], m=[]], f=[+finite, +realis],
  m=[]], f=[force=int], m=[]]
  (2) [cf=[], f=[definite=@definite], m=[]]
  (3) [f=[-animacy, concept=@concept, person=3, -plural, +quantizable, -wh],
  m=[]]

Regra: A0163 (C:5)
CYC: None, **ACT:** N0319, (head-s)
Células (símbolo de entrada: posição 1):
- (1) N0292, [cf=[cf=[], f=[+finite, +realis], m=[]], f=[force=int], m=[]]
  (2) [cf=[], f=[+definite], m=[]]

```
   (3) [f=[animacy=@animacy, concept=@concept, person=3, plural=@plural,
        +quantizable, -wh], m=[]]

Regra: A0164 (C:5)
CYC: N0319, ACT: N0303, (head-s)
Células (símbolo de entrada: posição 2):
   (1) [cf=[f=[animacy=@animacy, concept=@concept, person=3, plural=@plural,
        +quantizable, -wh], m=[]], f=[+definite], m=[]]
 • (2) N0293, [a=[], cf=[], f=[+change], m=[]]
   (3) [f=[concept=@concept, -wh], m=[]]

Regra: A0165 (C:1)
CYC: N0319, ACT: N0303, (comp)
Células (símbolo de entrada: posição 2):
   (1) [cf=[f=[-animacy, concept=ball, person=3, -plural, +quantizable, -wh],
        m=[]], f=[+definite], m=[]]
 • (2) N0024, [f=[concept=@concept, -wh], m=[]]
   (3) None

Regra: A0166 (C:1)
CYC: N0319, ACT: N0311, (head-s)
Células (símbolo de entrada: posição 1):
 • (1) N0303, [a=[], cf=[f=[concept=@concept, -wh], m=[]], f=[+change], m=[]]
   (2) [cf=[f=[-animacy, concept=ball, person=3, -plural, +quantizable, -wh],
        m=[]], f=[+definite], m=[]]
   (3) None

Regra: A0167 (C:5)
CYC: N0319, ACT: N0311, (comp)
Células (símbolo de entrada: posição 1):
 • (1) N0312, [cf=[f=[animacy=@animacy, concept=@concept, person=3,
              plural=@plural, quantizable=@quantizable, -wh], m=[]],
             f=[definite=@definite], m=[]]
   (2) None
   (3) None

Regra: A0168 (C:5)
CYC: None, ACT: N0319, (comp)
Células (símbolo de entrada: posição 1):
 • (1) N0311, [a=[cf=[f=[animacy=@animacy, concept=@concept, person=3,
              plural=@plural, +quantizable, -wh], m=[]], f=[+definite],
              m=[]], cf=[f=[concept=@concept, -wh], m=[]], f=[+change], m=[]]
   (2) None
   (3) None

Regra: A0170 (C:2)
CYC: N0319, ACT: N0312, (comp)
Células (símbolo de entrada: posição 1):
 • (1) N0297, [f=[-animacy, concept=book, person=3, -plural, +quantizable,
              -wh], m=[]]
   (2) [a=[], cf=[], f=[+change], m=[]]
   (3) [f=[concept=small, -wh], m=[]]

Regra: A0172 (C:2)
```

CYC: N0319, **ACT:** N0303, (comp)
Células (símbolo de entrada: posição 2):
   (1) [cf=[f=[-animacy, concept=book, person=3, -plural, +quantizable, -wh],
        m=[]], f=[+definite], m=[]]
 • (2) N0298, [f=[concept=small, -wh], m=[]]
   (3) None

Regra: A0177 (C:2)
CYC: N0319, **ACT:** N0303, (comp)
Células (símbolo de entrada: posição 2):
   (1) [cf=[f=[+animacy, concept=@concept, person=3, plural=@plural,
        +quantizable, -wh], m=[]], f=[+definite], m=[]]
 • (2) N0306, [f=[concept=violent, -wh], m=[]]
   (3) None

Regra: A0178 (C:1)
CYC: N0319, **ACT:** N0311, (head-s)
Células (símbolo de entrada: posição 1):
 • (1) N0303, [a=[], cf=[f=[concept=@concept, -wh], m=[]], f=[+change], m=[]]
   (2) [cf=[f=[+animacy, concept=kid, person=3, +plural, +quantizable, -wh],
        m=[]], f=[+definite], m=[]]
   (3) None

Regra: A0181 (C:3)
CYC: N0319, **ACT:** N0311, (head-s)
Células (símbolo de entrada: posição 1):
 • (1) N0303, [a=[], cf=[f=[concept=@concept, -wh], m=[]], f=[+change], m=[]]
   (2) [cf=[f=[animacy=@animacy, concept=@concept, person=3, -plural,
        +quantizable, -wh], m=[]], f=[+definite], m=[]]
   (3) None

Regras do tipo COMBINE

Regra: M0002 (C:12)
CYC: None, **ACT:** None, (seletor: ('m',))
Sense: [f=[animacy=@animacy, concept=@concept, person='3', plural=@plural,
        quantizable=@quantizable, -wh], m=@m]
Células (símbolo de entrada: posição 2):
   (1) [a=[cf=[f=[+animacy, person=1, -plural, +quantizable, -wh], m=[]],
        f=[+definite], m=[]], f=[relation=at-poss], m=[]]
 • (2) N0122, [f=[animacy=@animacy, concept=@concept, person=3, plural=@plural,
              quantizable=@quantizable, -wh], m=@m]
   (3) None

Regra: M0003 (C:20)
CYC: None, **ACT:** None, (seletor: cf)
Sense: [cf=[f=[animacy=@animacy, concept=@concept, person='3', plural=@plural,
        quantizable=@quantizable, -wh], m=[]], f=[definite=@definite], m=[]]
Células (símbolo de entrada: posição 1):
 • (1) N0122, [f=[animacy=@animacy, concept=@concept, person=3, plural=@plural,
              quantizable=@quantizable, -wh], m=@m]
   (2) None

```
   (3) None

Regra: M0008 (C:43)
CYC: None, ACT: N0312, (seletor: ('m',))
Sense: [f=[animacy=@animacy, concept=@concept, person='3', plural=@plural,
        quantizable=@quantizable, -wh], m=@m]
Células (símbolo de entrada: posição 2):
   (1) [f=[concept=@concept, -wh], m=[]]
 • (2) N0122, [f=[animacy=@animacy, concept=@concept, person=3, plural=@plural,
              quantizable=@quantizable, -wh], m=@m]
   (3) *

Regra: M0026 (C:3)
CYC: None, ACT: N0312, (seletor: ('cf',))
Sense: [cf=[f=[+animacy, concept=@concept, person='3', plural=@plural,
        +quantizable, -wh], m=[]], f=[concept='former', -wh], m=[]]
Células (símbolo de entrada: posição 1):
 • (1) N0171, [cf=[], f=[concept=former, -wh], m=[]]
   (2) [f=[+animacy, concept=@concept, person=3, plural=@plural, +quantizable,
        -wh], m=[]]
   (3) None

Regra: M0029 (C:27)
CYC: N0312, ACT: N0122, (seletor: ('a',))
Sense: [a=[cf=[f=[+animacy, concept=@concept, person='3', plural=@plural,
        +quantizable, -wh], m=@m], f=[definite=@definite], m=[]],
        f=[relation='at-poss'], m=[]]
Células (símbolo de entrada: posição 1):
 • (1) N0179, [a=[], f=[relation=at-poss], m=[]]
   (2) [f=@f, m=[]]
   (3) *

Regra: M0036 (C:22)
CYC: N0122, ACT: N0197, (seletor: ('cf',))
Sense: [cf=[f=[animacy=@animacy, concept=@concept, person='3', plural=@plural,
        quantizable=@quantizable, -wh], m=[]], f=[definite=@definite], m=[]]
Células (símbolo de entrada: posição 1):
 • (1) N0141, [cf=[], f=[definite=@definite], m=[]]
   (2) [f=[concept=@concept, -wh], m=[]]
   (3) *

Regra: M0041 (C:8)
CYC: N0197, ACT: N0312, (seletor: ('m',))
Sense: [f=[+animacy, concept=@concept, person='3', plural=@plural,
        +quantizable, -wh], m=@m]
Células (símbolo de entrada: posição 2):
   (1) [f=[concept=@concept, -wh], m=[]]
 • (2) N0122, [f=[animacy=@animacy, concept=@concept, person=3, plural=@plural,
              quantizable=@quantizable, -wh], m=@m]
   (3) None

Regra: M0042 (C:27)
CYC: None, ACT: N0312, (seletor: ('m',))
Sense: [f=[animacy=@animacy, concept=@concept, person='3', plural=@plural,
```

```
       quantizable=@quantizable, -wh], m=@m]
Células (símbolo de entrada: posição 1):
 • (1) N0122, [f=[animacy=@animacy, concept=@concept, person=3, plural=@plural,
               quantizable=@quantizable, -wh], m=@m]
   (2) [a=[], f=[relation=at-poss], m=[]]
   (3) [f=@f, m=[]]

Regra: M0045 (C:4)
CYC: N0122, ACT: N0197, (seletor: cf)
Sense: [+*abs*, cf=[f=[+animacy, concept=@concept, person='3', +plural,
       +quantizable, -wh], m=[]], f=[-definite], m=[]]
Células (símbolo de entrada: posição 1):
 • (1) N0122, [f=[animacy=@animacy, concept=@concept, person=3, plural=@plural,
               quantizable=@quantizable, -wh], m=@m]
   (2) None
   (3) None

Regra: M0047 (C:11)
CYC: None, ACT: None, (seletor: ('cf', 'cf', 'cf'))
Sense: [cf=[cf=[a=@a, cf=[f=[concept=@concept, -wh], m=[]], f=[change=@change],
       m=[]], f=[finite=@finite, +realis], m=[]], f=[force=@force], m=[]]
Células (símbolo de entrada: posição 2):
   (1) [cf=[f=[animacy=@animacy, concept=@concept, person=3, -plural,
       quantizable=@quantizable, -wh], m=[]], f=[definite=@definite], m=[]]
 • (2) N0231, [cf=[cf=[a=[], cf=[], f=[-change], m=[]], f=[-finite, +realis],
               m=[]], f=[force=decl], m=[]]
   (3) [f=[concept=@concept, -wh], m=[]]

Regra: M0048 (C:2)
CYC: None, ACT: None, (seletor: ('cf', 'cf', 'a'))
Sense: [cf=[cf=[a=[cf=[f=[animacy=@animacy, concept=@concept, person='3',
       -plural, quantizable=@quantizable, -wh], m=[]], f=[definite=@definite],
       m=[]], cf=[f=[concept=@concept, -wh], m=[]], f=[-change], m=[]],
       f=[finite=@finite, +realis], m=[]], f=[force=@force], m=[]]
Células (símbolo de entrada: posição 1):
 • (1) N0319, [cf=[cf=[a=@a, cf=[f=[concept=@concept, -wh], m=[]],
               f=[change=@change], m=[]], f=[finite=@finite, +realis], m=[]],
               f=[force=@force], m=[]]
   (2) [cf=[f=[-animacy, concept=@concept, person=3, -plural,
       quantizable=@quantizable, -wh], m=[]], f=[+definite], m=[]]
   (3) None

Regra: M0055 (C:1)
CYC: None, ACT: None, (seletor: ('cf', 'cf', 'a'))
Sense: [cf=[cf=[a=[cf=[f=[animacy=@animacy, concept=@concept, person='3',
       -plural, quantizable=@quantizable, -wh], m=[]], f=[definite=@definite],
       m=[]], cf=[f=[concept=@concept, -wh], m=[]], f=[-change], m=[]],
       f=[finite=@finite, +realis], m=[]], f=[force=@force], m=[]]
Células (símbolo de entrada: posição 1):
 • (1) N0319, [cf=[cf=[a=@a, cf=[f=[concept=@concept, -wh], m=[]],
               f=[change=@change], m=[]], f=[finite=@finite, +realis], m=[]],
               f=[force=@force], m=[]]
   (2) [cf=[f=[+animacy, concept=girl, person=3, -plural, +quantizable, -wh],
       m=[]], f=[+definite], m=[]]
```

```
   (3) None

Regra: M0058 (C:2)
CYC: None, ACT: None, (seletor: ('cf', 'cf', 'a'))
Sense: [cf=[cf=[a=[cf=[f=[animacy=@animacy, concept=@concept, person='3',
        -plural, quantizable=@quantizable, -wh], m=[]], f=[definite=@definite],
        m=[]], cf=[f=[concept=@concept, -wh], m=[]], f=[-change], m=[]],
        f=[finite=@finite, +realis], m=[]], f=[force=@force], m=[]]
Células (símbolo de entrada: posição 2):
   (1) [cf=[f=[+animacy, concept=mary, person=3, -plural, +quantizable, -wh],
        m=[]], f=[+definite], m=[]]
 • (2) N0319, [cf=[cf=[a=@a, cf=[f=[concept=@concept, -wh], m=[]],
              f=[change=@change], m=[]], f=[finite=@finite, +realis], m=[]],
              f=[force=@force], m=[]]
   (3) [f=[concept=often, +iteration, -location, -wh], m=[]]

Regra: M0059 (C:4)
CYC: None, ACT: None, (seletor: ('cf', 'cf', 'm'))
Sense: [cf=[cf=[a=[cf=[f=[animacy=@animacy, concept=@concept, person='3',
        -plural, +quantizable, -wh], m=[]], f=[definite=@definite], m=[]],
        cf=[f=[concept=@concept, -wh], m=[]], f=[-change], m=@m], f=[-finite,
        +realis], m=[]], f=[force='decl'], m=[]]
Células (símbolo de entrada: posição 1):
 • (1) N0278, [cf=[cf=[a=[cf=[f=[animacy=@animacy, concept=@concept, person=3,
              -plural, quantizable=@quantizable, -wh], m=[]],
              f=[definite=@definite], m=[]], cf=[f=[concept=@concept, -wh],
              m=[]], f=[-change], m=[]], f=[finite=@finite, +realis], m=[]],
              f=[force=@force], m=[]]
   (2) [f=[concept=often, +iteration, -location, -wh], m=[]]
   (3) None

Regra: M0060 (C:139)
CYC: None, ACT: None, (seletor: ('cf',))
Sense: [cf=[f=[animacy=@animacy, concept=@concept, person='3', plural=@plural,
        quantizable=@quantizable, -wh], m=[]], f=[definite=@definite], m=[]]
Células (símbolo de entrada: posição 1):
 • (1) N0141, [cf=[], f=[definite=@definite], m=[]]
   (2) [f=[concept=@concept, -wh], m=[]]
   (3) *

Regra: M0062 (C:1)
CYC: None, ACT: None, (seletor: ('cf', 'cf', 'a'))
Sense: [cf=[cf=[a=[cf=[f=[animacy=@animacy, concept=@concept, person='3',
        -plural, quantizable=@quantizable, -wh], m=[]], f=[definite=@definite],
        m=[]], cf=[f=[concept=@concept, -wh], m=[]], f=[-change], m=[]],
        f=[finite=@finite, +realis], m=[]], f=[force=@force], m=[]]
Células (símbolo de entrada: posição 2):
   (1) [cf=[f=[-animacy, concept=car, person=3, -plural, +quantizable, -wh],
        m=[]], f=[-definite], m=[]]
 • (2) N0319, [cf=[cf=[a=@a, cf=[f=[concept=@concept, -wh], m=[]],
              f=[change=@change], m=[]], f=[finite=@finite, +realis], m=[]],
              f=[force=@force], m=[]]
   (3) [f=[concept=often, +iteration, -location, -wh], m=[]]
```

Regra: M0064 (C:1)
CYC: None, **ACT:** None, (seletor: ('cf', 'cf', 'cf'))
Sense: [cf=[cf=[a=@a, cf=[f=[concept=@concept, -wh], m=[]], f=[change=@change],
m=[]], f=[finite=@finite, +realis], m=[]], f=[force=@force], m=[]]
Células (símbolo de entrada: posição 3):
(1) [f=[concept=often, +iteration, -location, -wh], m=[]]
(2) [cf=[f=[+animacy, concept=mary, person=3, -plural, +quantizable, -wh],
m=[]], f=[+definite], m=[]]
• (3) N0231, [cf=[cf=[a=[], cf=[], f=[-change], m=[]], f=[-finite, +realis],
m=[]], f=[force=decl], m=[]]

Regra: M0065 (C:1)
CYC: None, **ACT:** None, (seletor: ('cf', 'cf', 'a'))
Sense: [cf=[cf=[a=[cf=[f=[animacy=@animacy, concept=@concept, person='3',
-plural, quantizable=@quantizable, -wh], m=[]], f=[definite=@definite],
m=[]], cf=[f=[concept=@concept, -wh], m=[]], f=[-change], m=[]],
f=[finite=@finite, +realis], m=[]], f=[force=@force], m=[]]
Células (símbolo de entrada: posição 2):
(1) [f=[concept=often, +iteration, -location, -wh], m=[]]
• (2) N0319, [cf=[cf=[a=@a, cf=[f=[concept=@concept, -wh], m=[]],
f=[change=@change], m=[]], f=[finite=@finite, +realis], m=[]],
f=[force=@force], m=[]]
(3) [cf=[f=[+animacy, concept=mary, person=3, -plural, +quantizable, -wh],
m=[]], f=[+definite], m=[]]

Regra: M0066 (C:5)
CYC: None, **ACT:** None, (seletor: ('cf', 'cf', 'cf'))
Sense: [cf=[cf=[a=@a, cf=[f=[concept=@concept, -wh], m=[]], f=[change=@change],
m=[]], f=[finite=@finite, +realis], m=[]], f=[force=@force], m=[]]
Células (símbolo de entrada: posição 2):
(1) [cf=[f=[-animacy, concept=@concept, person=3, -plural, +quantizable, -wh],
m=[]], f=[definite=@definite], m=[]]
• (2) N0259, [cf=[cf=[a=[], cf=[], f=[-change], m=[]], f=[+finite, +realis],
m=[]], f=[force=decl], m=[]]
(3) [f=[concept=@concept, -wh], m=[]]

Regra: M0067 (C:18)
CYC: None, **ACT:** None, (seletor: ('cf', 'cf', 'a'))
Sense: [cf=[cf=[a=[cf=[f=[animacy=@animacy, concept=@concept, person='3',
-plural, quantizable=@quantizable, -wh], m=[]], f=[definite=@definite],
m=[]], cf=[f=[concept=@concept, -wh], m=[]], f=[-change], m=[]],
f=[finite=@finite, +realis], m=[]], f=[force=@force], m=[]]
Células (símbolo de entrada: posição 1):
• (1) N0319, [cf=[cf=[a=@a, cf=[f=[concept=@concept, -wh], m=[]],
f=[change=@change], m=[]], f=[finite=@finite, +realis], m=[]],
f=[force=@force], m=[]]
(2) [cf=[f=[animacy=@animacy, concept=@concept, person=3, -plural,
+quantizable, -wh], m=[]], f=[definite=@definite], m=[]]
(3) None

Regra: M0073 (C:7)
CYC: None, **ACT:** None, (seletor: ('cf', 'cf', 'cf'))
Sense: [cf=[cf=[a=@a, cf=[f=[concept=@concept, -wh], m=[]], f=[change=@change],
m=[]], f=[finite=@finite, +realis], m=[]], f=[force=@force], m=[]]

Células (símbolo de entrada: posição 1):
- (1) N0270, [cf=[cf=[a=[], cf=[], f=[-change], m=[]], f=[-finite, +realis], m=[]], f=[force=int], m=[]
  (2) [cf=@cf, f=[+definite], m=[]]
  (3) [f=[concept=@concept, -wh], m=[]]

Regra: M0074 (C:1)
CYC: None, **ACT:** N0319, (seletor: ('cf',))
Sense: [cf=[f=[animacy=@animacy, concept=@concept, person='3', plural=@plural, quantizable=@quantizable, -wh], m=[]], f=[definite=@definite], m=[]]
Células (símbolo de entrada: posição 1):
- (1) N0141, [cf=[], f=[definite=@definite], m=[]]
  (2) [f=[-animacy, concept=cake, person=3, -plural, -quantizable, -wh], m=[]]
  (3) [f=[concept=red, -wh], m=[]]

Regra: M0075 (C:1)
CYC: None, **ACT:** None, (seletor: ('cf', 'cf', 'a'))
Sense: [cf=[cf=[a=[cf=[f=[animacy=@animacy, concept=@concept, person='3', -plural, quantizable=@quantizable, -wh], m=[]], f=[definite=@definite], m=[]], cf=[f=[concept=@concept, -wh], m=[]], f=[-change], m=[]], f=[finite=@finite, +realis], m=[]], f=[force=@force], m=[]]
Células (símbolo de entrada: posição 1):
- (1) N0319, [cf=[cf=[a=@a, cf=[f=[concept=@concept, -wh], m=[]], f=[change=@change], m=[]], f=[finite=@finite, +realis], m=[]], f=[force=@force], m=[]]
  (2) [cf=[f=[-animacy, concept=cake, person=3, -plural, -quantizable, -wh], m=[]], f=[+definite], m=[]]
  (3) None

Regra: M0077 (C:1)
CYC: None, **ACT:** N0319, (seletor: ('cf',))
Sense: [cf=[f=[animacy=@animacy, concept=@concept, person='3', plural=@plural, quantizable=@quantizable, -wh], m=[]], f=[definite=@definite], m=[]]
Células (símbolo de entrada: posição 1):
- (1) N0141, [cf=[], f=[definite=@definite], m=[]]
  (2) [f=[+animacy, concept=cat, person=3, -plural, +quantizable, -wh], m=[]]
  (3) [f=[concept=big, -wh], m=[]]

Regra: M0082 (C:1)
CYC: None, **ACT:** N0319, (seletor: ('cf',))
Sense: [cf=[f=[animacy=@animacy, concept=@concept, person='3', plural=@plural, quantizable=@quantizable, -wh], m=[]], f=[definite=@definite], m=[]]
Células (símbolo de entrada: posição 1):
- (1) N0141, [cf=[], f=[definite=@definite], m=[]]
  (2) [f=[+animacy, concept=girl, person=3, -plural, +quantizable, -wh], m=[]]
  (3) [f=[concept=big, -wh], m=[]]

Regra: M0083 (C:3)
CYC: None, **ACT:** None, (seletor: ('cf', 'cf', 'cf'))
Sense: [cf=[cf=[a=@a, cf=[f=[concept=@concept, -wh], m=[]], f=[change=@change], m=[]], f=[finite=@finite, +realis], m=[]], f=[force=@force], m=[]]
Células (símbolo de entrada: posição 1):
- (1) N0284, [cf=[cf=[a=[], cf=[], f=[-change], m=[]], f=[+finite, +realis], m=[]], f=[force=int], m=[]]

(2) [cf=[], f=[definite=@definite], m=[]]
(3) [f=[-animacy, concept=@concept, person=3, -plural, +quantizable, -wh], m=[]]

Regra: M0084 (C:1)
CYC: None, **ACT:** N0319, (seletor: ('cf',))
Sense: [cf=[f=[animacy=@animacy, concept=@concept, person='3', plural=@plural, quantizable=@quantizable, -wh], m=[]], f=[definite=@definite], m=[]]
Células (símbolo de entrada: posição 1):
• (1) N0141, [cf=[], f=[definite=@definite], m=[]]
(2) [f=[-animacy, concept=ball, person=3, -plural, +quantizable, -wh], m=[]]
(3) [f=[concept=red, -wh], m=[]]

Regra: M0085 (C:1)
CYC: None, **ACT:** None, (seletor: ('cf', 'cf', 'a'))
Sense: [cf=[cf=[a=[cf=[f=[animacy=@animacy, concept=@concept, person='3', -plural, quantizable=@quantizable, -wh], m=[]], f=[definite=@definite], m=[]], cf=[f=[concept=@concept, -wh], m=[]], f=[-change], m=[]], f=[finite=@finite, +realis], m=[]], f=[force=@force], m=[]]
Células (símbolo de entrada: posição 1):
• (1) N0319, [cf=[cf=[a=@a, cf=[f=[concept=@concept, -wh], m=[]], f=[change=@change], m=[]], f=[finite=@finite, +realis], m=[]], f=[force=@force], m=[]]
(2) [cf=[f=[-animacy, concept=ball, person=3, -plural, +quantizable, -wh], m=[]], f=[+definite], m=[]]
(3) None

Regra: M0087 (C:1)
CYC: None, **ACT:** N0319, (seletor: ('cf',))
Sense: [cf=[f=[animacy=@animacy, concept=@concept, person='3', plural=@plural, quantizable=@quantizable, -wh], m=[]], f=[definite=@definite], m=[]]
Células (símbolo de entrada: posição 1):
• (1) N0141, [cf=[], f=[definite=@definite], m=[]]
(2) [f=[-animacy, concept=car, person=3, -plural, +quantizable, -wh], m=[]]
(3) [f=[concept=big, -wh], m=[]]

Regra: M0089 (C:1)
CYC: None, **ACT:** N0319, (seletor: ('cf',))
Sense: [cf=[f=[-animacy, concept='ball', person='3', -plural, +quantizable, -wh], m=[]], f=[-definite], m=[]]
Células (símbolo de entrada: posição 1):
• (1) N0141, [cf=[], f=[definite=@definite], m=[]]
(2) [f=[-animacy, concept=ball, person=3, -plural, +quantizable, -wh], m=[]]
(3) [f=[concept=big, -wh], m=[]]

Regra: M0090 (C:5)
CYC: None, **ACT:** None, (seletor: ('cf', 'cf'))
Sense: [cf=[cf=[a=@a, cf=[f=[concept=@concept, -wh], m=[]], f=[change=@change], m=[]], f=[finite=@finite, +realis], m=[]], f=[force=@force], m=[]]
Células (símbolo de entrada: posição 1):
• (1) N0292, [cf=[cf=[], f=[+finite, +realis], m=[]], f=[force=int], m=[]]
(2) [cf=[], f=[+definite], m=[]]
(3) [f=[animacy=@animacy, concept=@concept, person=3, plural=@plural, +quantizable, -wh], m=[]]

Regra: M0091 (C:1)
CYC: None, **ACT:** N0319, (seletor: ('cf',))
Sense: [cf=[f=[animacy=@animacy, concept=@concept, person='3', plural=@plural,
quantizable=@quantizable, -wh], m=[]], f=[definite=@definite], m=[]]
Células (símbolo de entrada: posição 1):
- (1) N0141, [cf=[], f=[definite=@definite], m=[]]
  (2) [f=[-animacy, concept=ball, person=3, -plural, +quantizable, -wh], m=[]]
  (3) [a=[], cf=[], f=[+change], m=[]]

Regra: M0092 (C:5)
CYC: None, **ACT:** N0319, (seletor: ('cf',))
Sense: [a=[], cf=[f=[concept=@concept, -wh], m=[]], f=[+change], m=[]]
Células (símbolo de entrada: posição 2):
  (1) [cf=[f=[animacy=@animacy, concept=@concept, person=3, plural=@plural,
+quantizable, -wh], m=[]], f=[+definite], m=[]]
- (2) N0293, [a=[], cf=[], f=[+change], m=[]]
  (3) [f=[concept=@concept, -wh], m=[]]

Regra: M0093 (C:1)
CYC: None, **ACT:** N0319, (seletor: ('a',))
Sense: [a=[cf=[f=[animacy=@animacy, concept=@concept, person='3', plural=@plural,
+quantizable, -wh], m=[]], f=[+definite], m=[]], cf=[f=[concept=@concept,
-wh], m=[]], f=[+change], m=[]]
Células (símbolo de entrada: posição 1):
- (1) N0303, [a=[], cf=[f=[concept=@concept, -wh], m=[]], f=[+change], m=[]]
  (2) [cf=[f=[-animacy, concept=ball, person=3, -plural, +quantizable, -wh],
m=[]], f=[+definite], m=[]]
  (3) None

Regra: M0099 (C:1)
CYC: None, **ACT:** N0319, (seletor: ('cf',))
Sense: [cf=[f=[animacy=@animacy, concept=@concept, person='3', plural=@plural,
quantizable=@quantizable, -wh], m=[]], f=[definite=@definite], m=[]]
Células (símbolo de entrada: posição 1):
- (1) N0141, [cf=[], f=[definite=@definite], m=[]]
  (2) [f=[+animacy, concept=kid, person=3, +plural, +quantizable, -wh], m=[]]
  (3) [a=[], cf=[], f=[+change], m=[]]

Regra: M0101 (C:1)
CYC: None, **ACT:** N0319, (seletor: ('a',))
Sense: [a=[cf=[f=[animacy=@animacy, concept=@concept, person='3', plural=@plural,
+quantizable, -wh], m=[]], f=[+definite], m=[]], cf=[f=[concept=@concept,
-wh], m=[]], f=[+change], m=[]]
Células (símbolo de entrada: posição 1):
- (1) N0303, [a=[], cf=[f=[concept=@concept, -wh], m=[]], f=[+change], m=[]]
  (2) [cf=[f=[+animacy, concept=kid, person=3, +plural, +quantizable, -wh],
m=[]], f=[+definite], m=[]]
  (3) None

Regra: M0102 (C:5)
CYC: None, **ACT:** N0319, (seletor: ('cf',))
Sense: [cf=[f=[animacy=@animacy, concept=@concept, person='3', plural=@plural,
quantizable=@quantizable, -wh], m=[]], f=[definite=@definite], m=[]]

Células (símbolo de entrada: posição 1):
- (1) N0141, [cf=[], f=[definite=@definite], m=[]]
  (2) [f=[animacy=@animacy, concept=@concept, person=3, -plural, +quantizable, -wh], m=[]]
  (3) [f=@f, m=[]]

Regra: M0103 (C:3)
CYC: None, **ACT:** N0319, (seletor: ('a',))
Sense: [a=[cf=[f=[animacy=@animacy, concept=@concept, person='3', plural=@plural, +quantizable, -wh], m=[]], f=[+definite], m=[]], cf=[f=[concept=@concept, -wh], m=[]], f=[+change], m=[]]
Células (símbolo de entrada: posição 1):
- (1) N0303, [a=[], cf=[f=[concept=@concept, -wh], m=[]], f=[+change], m=[]]
  (2) [cf=[f=[animacy=@animacy, concept=@concept, person=3, -plural, +quantizable, -wh], m=[]], f=[+definite], m=[]]
  (3) None
