

Efeitos de modelização no input: o caso da aquisição de conectores¹

Ana Costa, Nélia Alexandre, Ana Lúcia Santos & Nuno Soares
FLUL / CLUL

Neste trabalho, discutimos algumas hipóteses sobre o papel do input na aquisição, em particular sobre os efeitos da frequência no input. Levantaremos um problema metodológico aos trabalhos que usam a frequência no input como hipótese explicativa da ordenação dos estádios na aquisição: ao contrário do que é comumente assumido, o input não é estável, verificando-se efeitos de adaptação do discurso ao estágio em que a criança se encontra.

1. O problema: hipóteses sobre o papel na aquisição da frequência no *input*

Na investigação em aquisição, discute-se muitas vezes até que ponto a ordem de emergência de determinadas estruturas decorre da sua frequência no *input*, sugerindo-se que as estruturas mais frequentes serão adquiridas mais cedo. Em concreto, em trabalhos que assumem o modelo de abordagem “usage based”, defende-se que a emergência de certas estruturas linguísticas está positivamente correlacionada com a frequência dessas estruturas (Bybee, 1995; Langacker, 1988 e.o.). Partindo deste género de hipóteses, existe um interesse renovado pela análise do *input*, em particular do *input* dos pais (Diessel, 2004; Tomasello, 2003), sendo o estudo de Diessel (2004) um exemplo recente de investigação que defende a importância da análise do “ambiente linguístico” para a explicação da ordem de aquisição de conectores.

Diessel (2004), analisando a aquisição de conectores num *corpus* do Inglês, verifica que os conectores mais precocemente produzidos pelas crianças são *and* > *because* > *so* > *but* > *when* > *if*, pela ordem apresentada. De acordo com o autor, que inscreve o seu trabalho na linha da linguística cognitiva e funcional, conjugando a abordagem “construction grammar” com o modelo “usage-based”, a relação entre frequência no *input* e precocidade na aquisição é mediada ainda por uma noção de complexidade em que se assume que as expressões linguísticas e os padrões gramaticais mais frequentes na fala espontânea são os mais profundamente intrincados na gramática

¹ Agradecemos a Luísa Louira (FCUL) o apoio no tratamento estatístico dos dados, sem o qual este trabalho não poderia ter sido realizado. Agradecemos ainda a Nina Hyams a discussão de algumas hipóteses teóricas e a um avaliador anónimo as sugestões de clarificação do texto. Este trabalho foi realizado com o apoio do programa **POCTI-SFA-17-745** (Onset-CEL) da Fundação para a Ciência e a Tecnologia.

mental. Nesse sentido, e como admite o autor, os pesos relativos da frequência e da complexidade são difíceis de ponderar.²

A avaliação de explicações do processo de aquisição baseadas na frequência no *input* tem uma importância particular: no caso de se aceitar uma explicação deste tipo, qualquer explicação independente interna à teoria linguística sairá francamente enfraquecida. E, no entanto, não poderemos excluir a hipótese de que quer a ordem de emergência quer a frequência no *input* se encontrem correlacionadas por poderem ser explicadas por uma causa comum, essa sim interna à teoria linguística. Este é, na realidade, o tipo de posição assumido por Diessel (2004). Mas, crucialmente, não será possível prová-lo: podemos colocar a hipótese de que uma estrutura é mais frequente na fala espontânea por razões de complexidade sintáctica, semântica ou outra mas não será fácil provar que a frequência não se deve a outros factores, como a frequência da própria situação pragmática ou discursiva que justifica a sua ocorrência.

As abordagens centradas no *input* têm sido alvo de algumas críticas. Em particular, Evers-Vermeul (2005) aponta um problema metodológico a este tipo de trabalhos: estes estudos, que avaliam *corpora* longitudinais e apreciam a evolução da frequência de uma determinada estrutura na produção das crianças, consideram apenas números globais de ocorrências e frequência quando se trata de apreciar o *input*. Isto é, toma-se como seguro que o *input* é estável e que, assim, a frequência global seja um indicador do que a criança ouve em qualquer momento do processo de aquisição. No entanto, Evers-Vermeul, que observa a aquisição de conectores em Holandês e Inglês, sugere, com base no estudo de uma mãe e duas crianças gémeas, que o *input* não é estável porque sofre adaptações ao estágio em que a criança se encontra³. Esta hipótese, a confirmar-se, é particularmente relevante: por um lado, é possível que as generalizações que têm sido feitas sobre fala espontânea não sejam as mais correctas, o que, por si só, enfraquece abordagens centradas no *input*; por outro lado, a hipótese de Evers-Vermeul deixa em aberto a possibilidade de que a evolução da frequência no *input* seja causada (não a causa) pela própria ordem de emergência na aquisição. Em qualquer dos casos, justifica-se, então, a necessidade de uma explicação da ordem de emergência das estruturas definida de forma interna à teoria linguística.

² Citam-se as palavras do autor: “Since the earliest complex sentences are not only simple but also frequent, complexity and frequency are difficult to disentangle; both correlate very closely with the order of acquisition.” (Diessel, 2004: 6)

“Linguistic expressions and grammatical patterns that occur with high frequency in language use are more deeply entrenched in mental grammar than expressions that are infrequent.” (Diessel, 2004: 23)

³ Aliás, outros estudos sugerem esta hipótese. Os resultados de McCabe & Peterson (1997), referidos em Evers-Vermeul (2005), ainda que indirectamente, apontam para a existência de efeitos de modelização dos estímulos da linguagem dos pais: *They found that parents predate their children’s so-called spontaneous explicit expression of causal links by approximately five months. According to McCabe & Peterson (1997: 151), parents “scaffold their children’s emergent causal language by asking questions, repeating or revising such questions, occasionally answering those questions themselves, or abandoning them. Children respond to parental causal questions before they make statements of causal connection without such prompts (...).”* Evers-Vermeul (2005, 204).

Neste trabalho, apreciamos as hipóteses de Diessel e de Evers-Vermeul através da análise de um *corpus* da aquisição do Português Europeu a que aplicámos um modelo estatístico de derivação regressiva de forma a verificar até que ponto e em que direcção se verificam relações de causa-efeito entre o discurso das crianças e a fala que lhes é dirigida.

2. O caso da aquisição de conectores num *corpus* do Português Europeu

Costa (2006), estudando a produção escrita de conexões contrastivas por crianças e adolescentes (24 crianças entre os 9;3 e os 10;2 e 17 adolescentes entre os 14;3 e os 15;4), sugere que a aquisição de nexos concessivos, semanticamente contíguos aos adversativos e aos condicionais, é tardia. Costa coloca a hipótese de que o carácter tardio da aquisição de conectores concessivos possa ser explicado pela complexidade estrutural destas unidades, representadas no léxico por configurações complexas de traços.

No estudo de Costa (em prep.), os resultados relatados em Costa (2006) são comparados com a produção espontânea de conectores representativos de nexos contrastivos, causais e condicionais por três crianças entre os 1;5 e 3;11 (*corpus* de Santos, 2006, apresentado na Tabela 1). Também em estádios iniciais de aquisição se encontra uma escala clara de emergência de conectores que deve ser explicada à luz de hipóteses como a de Costa (2006) ou à luz de hipóteses centradas em correlações com a frequência no *input*. São estas as hipóteses que avaliaremos.

Tabela 1 – *Corpus* de produção espontânea

Criança	Idade	MLUw	Número de ficheiros	Número de enunciados produzidos pelas crianças
I. ⁴	1;6.6– 3;11.12	1.527– 3.815	21	6591
T.	1;6.18 – 2;9.7	1.286– 2.954	16	6800
I.M.	1;5.9 – 2;7.24	1.315– 2.370	15	5101

2.1. Um primeiro olhar sobre aquisição de conectores

Tendo em conta que o trabalho se centra na expressão de nexos contrastivos e outros semanticamente contíguos, foram pesquisados diversos conectores com os valores em análise no *corpus* de Santos (2006).⁵ De acordo com os resultados obtidos,

⁴ Dados recolhidos por M. J. Freitas (Freitas, 1997), projecto PCSH/C/LIN/524/93 – Laboratório de Psicolinguística, FLUL.

⁵ Esta pesquisa incluiu conjunções, locuções conjuncionais e advérbios conectivos aditivos, concessivos, adversativos, contrastivos (como *enquanto*), causais, explicativos e condicionais. A pesquisa foi feita automaticamente, usando complementarmente os comandos *freq* e *kwal* (este último permite pesquisar as palavras em contextos) no programa CLAN do sistema Childes (MacWhinney, 2000).

os conectores produzidos nos primeiros estádios são *e*, *mas*, *porque* e *se*. Se observarmos as idades a que correspondem as primeiras produções de cada conector em cada uma das três crianças observadas, é óbvia a possibilidade de estabelecer uma escala de emergência que corresponderá a *e* > *mas* > *porque* > *se*. As idades das primeiras ocorrências de cada um destes conectores na produção de cada uma das três crianças são apresentadas na tabela 2.

Tabela 2 – Emergência de conectores no *corpus* de Santos (2006)

	I.	T.	I.M.
e	1;1.29	1;8.16	1;9.8
mas	2;4.19	2;1.7	2;2.18
porque	2;3.8/ 2;5.24	2;6.6	2;3.22
se	2;7.16	2;9.7	2;5.23

Em (1) a (4) são apresentadas as primeiras ocorrências dos conectores em estudo no discurso de uma das crianças, Inês M.

- (1) *INM: e mai(s)? (I.M. 1;9.8)
- (2) *INM: pa(r)ece ma(s) u [: não] é . (I.M. 2;2.18)
- (3) *INM: porque não . (I.M. 2;3.22)
- (4) *INM: se (calh)ar (es)tá na carrinha . (I.M. 2;5.23)

Na verdade, a escala encontrada para o Português (*e* > *mas* > *porque* > *se*) aproxima-se bastante da escala apresentada por Diessel (2004) para o inglês: *and* > *because* > *so* > *but* > *when* > *if*.⁶

Interessará, então, observar como se distribuem os dados em termos de número de ocorrências. Também desse ponto de vista se confirma a escala encontrada: nos dados observados, *e* > *mas* > *porque* > *se* corresponde igualmente a uma escala de ordenação dos conectores dos mais frequentes para os menos frequentes (Tabela 3). A ideia de que *e* e *mas* são adquiridos mais precocemente do que *porque* e *se* é confirmada pelo número relativo de ocorrências dos diferentes conectores, sendo, como esperado, as estruturas mais estabilizadas as que ocorrem mais frequentemente.

⁶ O conector *quando* e os conectores conclusivos não foram analisados no Português, por não se encontrarem dentro do escopo do trabalho de Costa (2006) e Costa (em prep.).

Tabela 3 – Número de ocorrências dos conectores *e*, *mas*, *porque* e *se*

	T	I	I.M.
e	300	302	230
mas	15	84	12
porque	3	50	3
se	2	6	2

Será então de procurar replicar com os dados do Português as observações feitas por Diessel (2004). Sabendo que o autor encontra uma correlação clara entre ordem de emergência de conectores no discurso das crianças e número de ocorrências e frequência dos mesmos conectores na fala espontânea das mães (que o autor toma como exemplo do *input*), tentámos estabelecer o mesmo tipo de comparação. Na tabela 4, são apresentados os números de ocorrências dos diferentes conectores em estudo na fala espontânea das mães.

Tabela 4 – Número de ocorrências dos conectores *e*, *mas*, *porque* e *se*

	Mãe T	Mãe I	Mãe I.M.
e	1331	1268	371
mas	288	690	137
porque	28	91	27
se	16	13	14

A observação destes dados mostra, corroborando as observações de Diessel, que o número de ocorrências dos diferentes conectores no discurso das mães estabelece uma escala que espelha a ordem de emergência dos mesmos conectores no discurso das crianças: *e* > *mas* > *porque* > *se*. No entanto, haverá apenas uma explicação para este facto?

Na verdade, ao provarmos a coincidência das duas escalas (de emergência e de número de ocorrências), no *input* e no discurso das crianças, temos uma indicação de uma correlação, mas não *provamos* a existência de uma relação de causalidade entre os dois tipos de facto. Como foi já dito na secção anterior, se é verdade que as crianças podem ser sensíveis à frequência de uma estrutura na fala a que estão expostas, também é verdade que não excluimos a hipótese de que uma escala de complexidade sintáctica e / ou semântica explique quer a ordem de aquisição quer a frequência na fala espontânea. Se numa abordagem “usage-based” a questão da frequência não se distingue da medida de complexidade, em abordagens no âmbito da gramática generativa essas noções são independentes (veja-se a escala de complexidade proposta por Jakubowicz & Nash, 2001, Jakubowicz, 2004 ou ainda em Soares, 2006).

Por outro lado, a única forma de provar uma relação de causalidade entre as duas ordens de factos será considerar (i) não o número global de ocorrências, mas a frequência (número de ocorrências / número de enunciados), (ii) a frequência ao longo

das diferentes observações, tendo em conta a possibilidade de o *input* não ser estável (Evers-Vermeul, 2005) e (iii) um modelo estatístico que permita determinar o valor preditivo da frequência no *input* relativamente à frequência no discurso das crianças.

2.2. Hipóteses

Pretendemos verificar se a proporção de conectores na fala espontânea das crianças é de facto correlacionável com a proporção dos mesmos conectores na fala espontânea das mães. Por outro lado, e sendo relevante provar que os números globais de frequência apresentam, como se espera em aquisição, uma curva de crescimento reveladora da estabilização de uma estrutura, pretendemos verificar até que ponto há crescimento da proporção de conectores na fala espontânea ao longo do tempo.

Contudo, tendo em conta a hipótese levantada por Evers-Vermeul (2005), segundo a qual o *input* revela efeitos de adaptação ao estádio em que as crianças se encontram, interessou verificar se, ao contrário do que é do senso-comum, há crescimento da proporção de conectores no *input* ao longo do tempo e se a proporção de conectores na fala espontânea das mães é, em cada um desses momentos no tempo, correlacionável com a proporção dos mesmos conectores na fala espontânea das crianças.

Para responder a estas perguntas, foi feito um cálculo de frequências dos diferentes conectores em cada um dos momentos de observação (número de ocorrências / número de enunciados). Uma vez que p (proporção de conectores) se situa entre 0 e 1, foi feita uma transformação logit aos dados, de acordo com a fórmula apresentada em (5).

$$(5) \quad \text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \log(p) - \log(1-p).$$

Foi então aplicado aos dados um modelo de regressão linear. Interessava aferir se:

(i) a frequência de conectores na criança é uma variável preditora da frequência de conectores na mãe;

(ii) o Tempo é uma variável preditora da frequência de conectores na mãe, caso em que se provaria que o *input* não é estável;

(i) o Tempo é uma variável preditora da frequência de conectores na criança, provando o crescimento;

(ii) a frequência de conectores na mãe é uma variável preditora da frequência de conectores na criança.

Assim, foram introduzidas como variáveis, num modelo de regressão múltipla:

– variável T (Tempo, 1 unidade = 1 mês);

– variável T²;

– (i) no estudo da produção das crianças, a proporção de conectores produzidos pela mãe;

(ii) no estudo da produção das mães, a proporção de conectores produzidos pela criança.

Quando relevante, foi aplicado um modelo de regressão linear usando como única variável preditora a proporção do conector no discurso da mãe / da criança. Esta última permitiria verificar a eventual existência de uma correlação deste tipo que poderia não ser detectada num modelo com múltiplas variáveis por o seu valor preditivo ser redundante tendo em conta o valor preditivo de outras variáveis. O trabalho estatístico foi feito com SPSS versão 15.0 para o Windows.

2.3. Resultados

Os resultados obtidos na avaliação da produção das crianças revelam, de forma não surpreendente, que o Tempo tem uma capacidade preditiva significativa ($p < .001$) da ocorrência dos conectores *e*, *mas*, *porque* e *se*, sendo capaz de explicar entre 48,8% e 79,6% da variação. A observação dos dados mostra que os diferentes conectores se tornam progressivamente mais frequentes ao longo das observações. No entanto, o modelo não confirma que a frequência dos conectores no discurso das mães tenha capacidade preditiva relativamente à frequência dos mesmos conectores no discurso das crianças. Na realidade, nestes casos, o estudo é inconclusivo: não foi possível obter resultados na aplicação de um modelo de regressão linear que tivesse como única variável preditora a proporção do conector no discurso das mães, já que não se verificavam, nestes casos, os pressupostos do modelo estatístico.⁷

Tabela 5 – Resultados da análise do discurso das crianças: regressão linear

Conector	Capacidade preditora do tempo (T / T^2)	Capacidade preditora do discurso da mãe
e	✓ (79,6% da variação, $p < .001$)	-----
mas	✓ (61,7% da variação, $p < .001$)	-----
porque	✓ (48,8% da variação, $p < .001$)	-----
se	-----	

Já a observação dos resultados da análise das produções das mães traz maiores surpresas. Veja-se a tabela 6.

⁷ A notação ----- nas tabelas indica que as hipóteses não foram confirmadas nem infirmadas.

Tabela 6 – Resultados da análise do discurso das mães: regressão linear

Conector	Capacidade preditora do tempo (T / T ²)	Capacidade preditora do discurso da criança
e	-----	Hipótese rejeitada
mas	✓ (30,1% da variação, $p < .001$)	✓ (12% da variação, $p < .05$)
porque	✓ (26,2% da variação, $p < .001$)	-----
se	-----	-----

Muito embora seja possível rejeitar a hipótese de que a proporção do conector *e* no discurso das crianças tenha valor preditivo da proporção do mesmo conector no discurso das mães, os resultados obtidos com os conectores *mas* e *porque* justificam a formulação de hipóteses diferentes.

Por um lado, os resultados da análise revelam que, no caso do conector *mas*, foi possível determinar que a frequência do conector no discurso das crianças é capaz de explicar 12% da variação no discurso das mães, sendo esta correlação significativa ($p < .05$).

Por outro lado, confirmando a hipótese de Evers-Vermeul (2005), verifica-se que o discurso das mães não é estável, sofrendo um crescimento ao longo do tempo: o tempo é capaz de explicar entre 26,2% e 30,1% da variação, tendo uma capacidade preditiva significativa ($p < .001$). Por si só, este resultado é bastante interessante: como explicar o crescimento da produção de um conector no discurso de um adulto, que não passa por um processo de desenvolvimento linguístico? A única hipótese possível parece de facto ser a de que este crescimento seja um efeito da adaptação do discurso das mães ao estágio em que as crianças se encontram.

Nas Figuras 1 e 2, apresenta-se graficamente a evolução da produção dos conectores *mas* e *porque* no discurso das mães de cada uma das crianças. Na Figura 3, apresenta-se a evolução da produção do conector *se* no discurso das mesmas mães – neste caso, embora não seja possível apresentar resultados de análise, visto que não se verificam os pressupostos do modelo estatístico, a simples observação do gráfico mostra que há evolução na produção de *se* ao longo do tempo.⁸

⁸ Em concreto, ajustando a variável tempo, seria provavelmente possível obter um modelo que descrevesse correctamente a curva de crescimento deste conector no discurso das mães (no modelo utilizado foram apenas utilizadas as medidas T e T²).

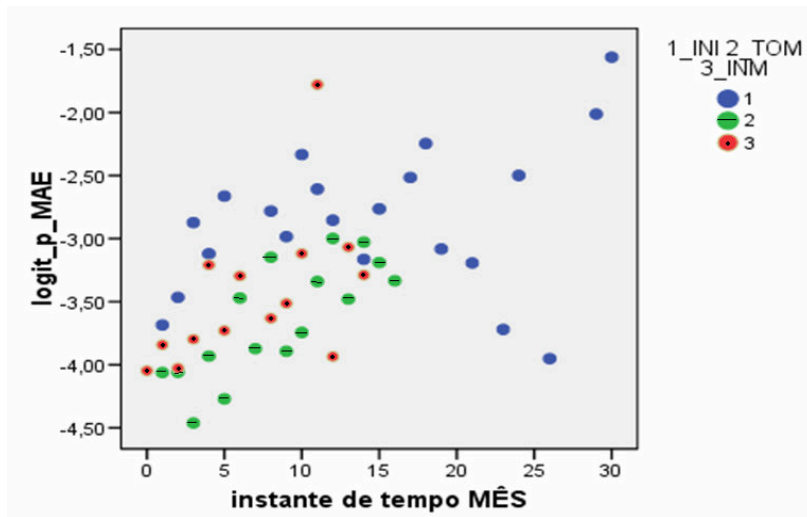


Fig. 1 – Produção de *mas* pelas mães

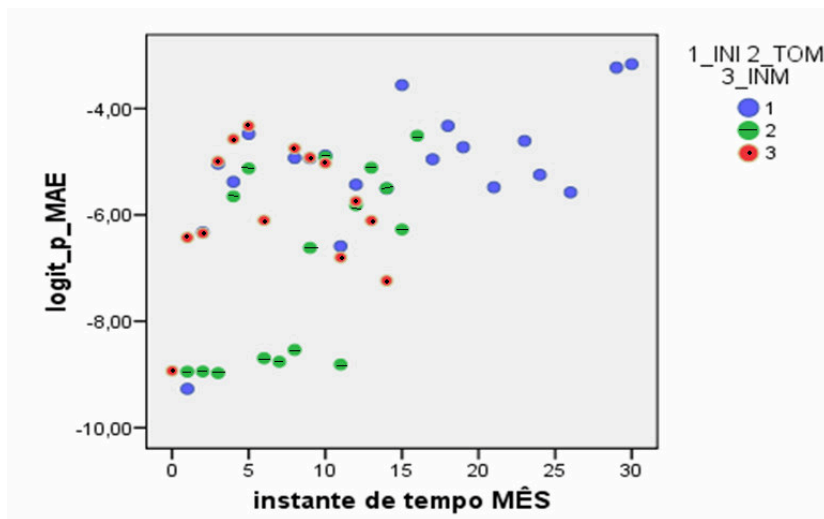
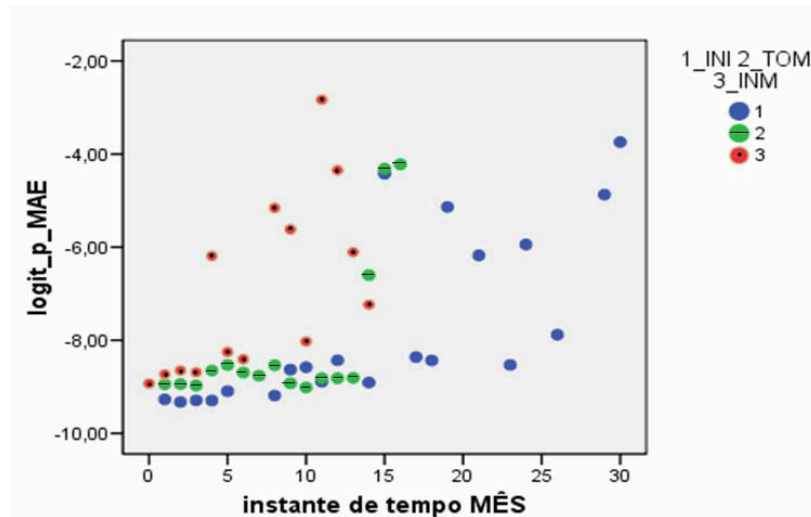


Fig. 2 – Produção de *porque* pelas mães

Fig. 3 – Produção de *se* pelas mães

A aplicação do modelo de regressão linear permitiu, portanto, observar que:

- (i) O Tempo é a variável com maior valor preditivo, quer da produção de conectores pela criança, quer da produção de conectores pela mãe;
- (ii) Num caso, a proporção de conectores no discurso da criança é uma variável explicativa da frequência no discurso da mãe (no caso da produção de *mas*, 12% da variação, $p < .05$);
- (iii) Em nenhum caso se prova o contrário, i.e., em nenhum caso a proporção de conectores no discurso da mãe é uma variável explicativa da proporção de conectores no discurso da criança.

3. Conclusões

Deste trabalho, resultam duas grandes conclusões. Em primeiro lugar, não foi possível confirmar a hipótese de que a frequência de conectores no *input* determina a frequência e a ordem de emergência dos mesmos conectores no discurso das crianças, ao contrário do que se poderia esperar no quadro de análises como a que é proposta em Diessel (2004). Em segundo lugar, e esta é certamente a conclusão mais relevante deste trabalho, confirma-se a hipótese de Evers-Vermeul (2005): o *input* não é estável, podendo a frequência dos conectores no discurso das mães ser descrita através de uma linha de crescimento. Assim, é possível admitir, com Evers-Vermeul, que as mães adaptam o seu discurso ao estágio de desenvolvimento em que a criança se encontra.

Este último facto constitui, por si só, um problema metodológico importante a apontar aos estudos que baseiam as suas explicações em efeitos de frequência global no *input*: estes estudos negligenciam uma análise detalhada da distribuição das frequências ao longo do tempo, o que só seria legítimo se o *input* ou fosse estável ou variasse de forma aleatória. Por outro lado, este tipo de resultados, ao retirar força explicativa ao papel da frequência no *input*, justifica a necessidade de explicações dos factos linguísticos internas à teoria linguística. Naturalmente, não é nosso objectivo sugerir que o *input* ou a frequência no *input* é irrelevante na aquisição ou que o discurso das mães é o único tipo de experiência linguística a que as crianças estão expostas. O que queremos sugerir é que (i) as conclusões a que se chegou em estudos como o de Diessel (2004) são baseadas em pressupostos errados e que, (ii) independentemente do papel da experiência linguística na explicação do processo de aquisição, há certamente necessidade de se recorrer a parâmetros como escalas de complexidade definidas em termos puramente sintácticos ou semânticos para explicar a ordem de emergência das diferentes estruturas.

Finalmente, será ainda necessário considerar um possível problema a levantar à análise aqui apresentada: como explicamos nós que, embora se verifique crescimento na produção de conectores no discurso das mães, possa não se verificar nos mesmos casos valor preditivo do discurso das crianças relativamente à produção das mães? Isto é, por que será que a proporção de conectores no discurso das mães cresce ao longo do tempo, mas sem espelhar necessariamente o número de conectores efectivamente produzidos pelas crianças? Na realidade, é possível que o efeito de modelização do discurso, por parte da mãe, seja um efeito de adaptação àquilo que ela acha ser a capacidade de *compreensão* da criança e não necessariamente ao que a criança produz já efectivamente.

Esta e as restantes hipóteses aqui sugeridas levam a formular outras perguntas e esse conjunto de novas perguntas é a verdadeira conclusão deste trabalho: teremos intuições sobre complexidade? Como será que temos intuições sobre complexidade? O que será a base dessas intuições?⁹

4. Referências

- Bybee, J. (1995) Regular Morphology and the Lexicon. *Language and Cognitive Processes* 10.5, pp. 425-45
- Costa, A. L. (2006) Complexidade Estrutural de Conectores Concessivos. *Textos Seleccionados do XXII Encontro da Associação Portuguesa de Linguística*. Lisboa: APL.
- Costa, A. L. (em prep) *Estruturas Contrastivas: Desenvolvimento do Conhecimento Explícito e da Competência de Escrita*. Dissertação de doutoramento a apresentar à Universidade de Lisboa
- Diessel, H. (2004) *The Acquisition of Complex Sentences*. Cambridge: Cambridge U. P.
- Evers-Vermeul, J. (2005) *The Development of Dutch Connectives: Change and Acquisition as Windows on Form-Function Relations*. Utrecht: LOT

⁹ Nina Hyams (p.c.).

- Freitas M. J. (1997) *Aquisição da Estrutura Silábica do Português Europeu*. Dissertação de Doutoramento. Universidade de Lisboa.
- Jakubowicz, C. (2004). Is Movement Costly? The Grammar and the Processor in Language Acquisition. Comunicação apresentada ao JEL'2004, Nantes, France, 5-7 mai.
- Jakubowicz, C e Nash, L. (2001). Functional categories and syntactic operations in (ab)normal language acquisition. *Brain and Language*, 77.3: 321-339.
- Langacker (1988) A Usage-Based Model. In B. Rudzka-Ostyn (ed.) *Topics in Cognitive Linguistics*. Amsterdam: John Benjamins, pp. 127-161
- MacWhinney, B. (2000) *The CHILDES Project: Tools for Analyzing Talk*. 3ª edição. Mahwah, NJ: Lawrence Erlbaum Associates
- Santos, A. L. (2006) *Minimal Answers. Ellipsis, syntax and discourse in the acquisition of European Portuguese*. Dissertação de Doutoramento. Universidade de Lisboa.
- Soares, C. (2006) *La Syntaxe de la Périphérie Gauche en Portugais Européen et son Acquisition*. Dissertação de Doutoramento. Université Paris 8.
- Tomasello, M. (2003) *Constructing a Language. A Usage-Based Theory of Language Acquisition*. Cambridge, Mass.: Harvard University Press.