

Análise de *corpus*, a intuição do lingüista e metodologia experimental na pesquisa sobre as orações relativas do PB e do PE. *

Eduardo Kenedy (UFF) **

RESUMO: Este artigo toma o caso das orações relativas do português do Brasil e de Portugal para exemplificar as limitações impostas às pesquisas baseadas exclusivamente em análise de corpus e na intuição do lingüista. Argumenta-se que a metodologia experimental da pesquisa em psicolingüística é uma solução adequada tanto para testar a realidade psicológica de modelos teóricos em competição quanto para superar contradições advindas de análises de corpora.

Palavras-chave: metodologia de pesquisa em linguística, análise de corpus, psicolingüística experimental.

Introdução

Desde o trabalho seminal de Fernando Tarallo (1983), as orações relativas da língua portuguesa têm sido objeto de intensa controvérsia entre lingüistas brasileiros e portugueses. Na análise original de Tarallo, as relativas do português do Brasil (PB) pertenceriam a um sistema gramatical muito diferente daquele responsável pela relativização em português europeu (PE). Para o autor, relativas do PB seriam geradas sem a aplicação de regra de Movimento e com a aplicação de uma regra de retenção ou apagamento de pronome, enquanto, em PE, a aplicação sistemática de regra de Movimento sobre pronomes relativos seria responsável pela derivação de orações relativas. Essa diferença estrutural seria a explicação para o fato de, nos dados analisados por Tarallo, relativas *pied-piping* (a estratégia padrão para as relativas preposicionadas), como em (1a), serem praticamente inexistentes em PB, mas de alta produtividade em PE, bem como para a farta ocorrência de relativas resumptivas e cortadoras, como em, respectivamente, (1b) e (1c), nos dados do PB, por contraste à sua baixa produtividade ou inexistência em PE.

- 1) a. A pessoa com quem eu conversei ficou doente.
- b. A pessoa que eu conversei com ela ficou doente.
- c. A pessoa que eu conversei ficou doente.

* Este artigo resume as partes essenciais dos capítulos III e IV da tese de Kenedy (2007). O autor agradece a seu orientador, Dr. Marcus Maia (UFRJ), a sua orientadora no exterior, Dra. Armanda Costa (Univ. de Lisboa) e à CAPES (Bolsa PDEE, processo BEX2955/06-7), sem os quais a pesquisa não teria sido possível.

** Contato: www.eduardokenedy.net

A hipótese, levantada por Tarallo (1983, 1985), de uma separação entre o sistema de relativização do PB e do PE ficou conhecida como *hipótese paramétrica* e foi estendida a outras áreas da gramática, como a pronominalização, a concordância verbo-nominal, a topicalização etc. (cf. Roberts & Kato, 1993). Para fundamentar as suas hipóteses, Tarallo baseou-se em sua própria intuição sobre o português – o tradicional julgamento de gramaticalidade emitido pelo falante nativo – e em dados retirados de *corpora* representativos de diferentes sincronias da língua.

Inspirados pelo estudo original de Tarallo, seguiram-se, nos últimos 25 anos, diversos trabalhos que ora confirmavam, alteravam ou negavam a hipótese paramétrica. Por exemplo, Kato (1993), com base nos estudos de *corpus* de Pontes (1987) sobre a topicalização, propôs a hipótese da Left-Deslocation (LD) para as relativas do PB, por oposição ao PE, em que tal LD não se aplicaria. Essa hipótese seria assumida mais tarde por Galves (2001), sem o suporte dos dados de *corpus*, mas com base exclusivamente na intuição linguística da autora. Já estudos como o de Arin, Ramilo e Freitas (2005) encontraram, em dados de língua oral e escrita da imprensa portuguesa, evidências de que relativas resumptivas e cortadoras seriam produtivas em PE, numa contradição à hipótese paramétrica. Os dados desses autores se afastavam das conclusões de Alexandre (2000), que, na análise da fala culta do *Corpus de Referência do Português Contemporâneo – Oral* (CRPC), encontrou ocorrências muito baixas das relativas não-padrão. Por sua vez, a pesquisa de Varejão (2006), sustentada nos dados do *Corpus Dialectal para o Estudo da Sintaxe* (Cordial-Sin), com a fala de indivíduos analfabetos ou semi-alfabetizados portugueses, detectou padrão inverso ao da pesquisa de Alexandre (2000): relativas *pied-piping* praticamente não são encontradas no Cordial-Sin, por oposição à vasta ocorrência no *corpus* de cortadoras e resumptivas.

Essa intensa variabilidade nos resultados de pesquisas dedicadas à natureza das relativas em PB *versus* PE pode ser explicada em função da metodologia adotada nos estudos

citados. A análise de *corpus*, às vezes aliada à intuição do lingüista, é quase sempre a fonte de informação privilegiada nas pesquisas sobre a querela PB e PE. No presente artigo, demonstraremos que, embora essenciais para apontar tendências no uso de uma língua, as análises de *corpus* apresentam limitado poder explanatório para as pesquisas que pretendem fazer generalizações sobre a competência linguística dos falantes (no sentido de Chomsky, 1965), uma vez que seus resultados podem ser fortemente enviesados pelo perfil sociocultural dos informantes entrevistados e/ou pelo gênero textual característico do *corpus*. Quer isso dizer que *corpora* com perfis socioculturais e gêneros textuais diferentes normalmente apresentarão dados muito distintos, mesmo em relação a um único fenômeno gramatical variável como a relativização, seja no cotejo entre PB e PE ou entre variedades internas ao próprio PB ou ao próprio PE. Apontaremos neste artigo que, para superação de impasses metodológicos em estudos comparativos, como o da oposição paramétrica entre PB e PE, a metodologia experimental da psicolingüística pode trazer importante contribuição. O trabalho de Kenedy (2007) é uma ilustração de pesquisa que permite a comparação controlada de falantes do PB e do PE que possuem perfis socioculturais equivalentes e são submetidos a experimentos psicolingüísticos idênticos, cujos resultados são submetidos a tratamento estatístico rigoroso, permitindo generalizações mais estáveis sobre a gramática mental dos sujeitos do PB e do PE. Os resultados do autor indicaram que, no momento reflexo do uso da linguagem – capturado em experimentos *on-line* e *off-line* controlados –, o desempenho de falantes do PB e do PE com orações relativas não é estatisticamente distinguível.

1. A hipótese paramétrica

Tarallo (1983), em pesquisa diacrônica, apresentou a hipótese de que as relativas cortadoras teriam sido criadas, no século XIX, pelo PB, sendo essas desconhecidas ou irrelevantes na gramática do PE desde sempre. Os dados a seguir demonstram como o autor compreende que o PB tenha invertido completamente o sistema de relativização do PE. As

relativas *pied-piping* teriam deixado de ser a estratégia mais produtiva na língua, realidade encontrada nos textos mais antigos, e teriam passado a ser pouco recorrentes nos textos mais atuais, ao passo que as relativas cortadoras praticamente inexisteriam em textos do século XVIII e passariam a ser responsáveis por quase 60% das relativizações em textos do século XIX.

	circa 1725	circa 1770	circa 1825	circa 1880
Pied-piping	89.2%	88.1%	91.3%	35.4%
Resumptivas	9.9%	7.9%	1.3%	5.1%
Cortadoras	0.9%	4%	7.5%	59.5%

Tabela 1: ocorrência das estratégias de relativização ao longo de textos dos séculos XVIII e XIX, adaptado de Tarallo (1985: 371).

Para o lingüista, essa mudança histórica estaria associada a uma mudança na gramática mental dos falantes do português – o que se conhece como *hipótese paramétrica*. No PB, o sistema de relativização teria perdido a aplicação sistemática do Move sobre o constituinte *wh-* e teria passado a derivar as relativas na base, sem regras de Movimento, através da aplicação de regras de retenção ou elipse pronominal (pro-drop).

Do ponto de vista sincrônico, Tarallo (1983: 368) abandona os textos escritos (de diferentes gêneros) de sua pesquisa diacrônica e aproxima-se do uso natural da linguagem, concentrando-se na língua oral espontânea, e demonstra que, em PB, as relativas cortadoras são muito mais produtivas que *pied-piping*. No entanto, ao contrário do que é indicado pelos seus dados diacrônicos, a cortadora não é a relativização mais produtiva no *corpus* da atual sincronia do PB, já que perde numericamente para a resumptiva, que se apresenta com praticamente o dobro das ocorrências registradas para as cortadoras. O autor analisou um *corpus* de língua falada bastante heterogêneo, que continha falantes cultos e falantes com média e baixa escolarização, incluindo trabalhadores da classe operária do estado de São Paulo e também tipos de fala muito díspares, como a conversa casual e a dissertação livre. Nesse *corpus*, Tarallo identificou 324 relativas preposicionadas, retiradas de um universo de

1700 orações relativas. Dessas 324, apenas 21 (6,6%) eram *pied-piping*. As cortadoras registram 101 ocorrências (31,7%) e seguem as resumptivas, com 200 ocorrências (61,7%).

O descompasso entre a baixíssima produtividade das resumptivas no *corpus* histórico e a sua preponderância no *corpus* sincrônico pode ser explicado pela diferença na tipologia textual entre os dados analisados por Tarallo. No seu *corpus* da atual sincronia, encontramos dados de língua oral em situações que reproduzem aproximadamente os diálogos naturais, contexto em que as reiteraões com uso de pronomes são muito freqüentes em qualquer língua humana (cf. Dooley & Levinsohn, 2001). Ao levarmos em conta especialmente a fala de pessoas iletradas, não treinadas nos artificialismos da língua mais formal, será natural encontrarmos um grande número de relativas resumptivas. Por outro lado, os textos escritos que compõem o *corpus* diacrônico são representativos de um uso mais monitorado e artificial da língua, em que reiteraões com resumptivos são eliminadas de maneira sistemática, a não ser nos casos em que, por alguma razão, são intencional ou descuidadamente preservados. Eis aqui um grave problema que, se não for controlado, pode comprometer os resultados de pesquisas baseadas exclusivamente em análise de *corpus*: a variabilidade sociocultural dos falantes e a grande variedade de gêneros do discurso.

O estudo de Tarallo (1983), inobstante sua grande importância para a história da lingüística brasileira e de sua originalidade para a época, apresenta o problema metodológico de trabalhar com textos de escritos em diversos gêneros (formal, informal, anúncios de jornal, documentos legais, peças de teatro etc.), sacados de diferentes momentos históricos, e usar os dados estatísticos ali coletados para fazer generalizaões a respeito da gramática natural de uma língua. Parece correto interpretarmos que textos escritos formais ou artísticos tipicamente não apresentam fenômenos da oralidade – como as relativas cortadoras –, a não ser quando tais textos procuram intencionalmente reproduzir certo tipo de fala, ou ainda quando esses registram o descuido de um escritor que deixou escapar um uso lingüístico menos elaborado.

Portanto, é possível interpretar que os dados da pesquisa diacrônica de Tarallo indiquem o momento em que as cortadoras passaram a ser registradas na língua escrita, considerando-se certos tipos de documentos, escritos sob certas circunstâncias sócio-históricas, mas não parece correto que esse registro possa dizer algo diretamente relacionável à gramática natural de uma língua, como o surgimento de uma estrutura na gramática ou o desaparecimento de outra.

2. Confirmando a hipótese paramétrica com dados de *corpus* do PE

A dissertação de mestrado de Alexandre (2000) foi especialmente dedicada às relativas resumptivas, com o objetivo de caracterizá-las de um ponto de vista formal, sob a perspectiva do Programa Minimalista, com base num extenso banco de dados de língua falada, o *Corpus de Referência do Português Contemporâneo – Oral* (CRPC). Desse *corpus*, a autora extraiu 18.500 ocorrências de orações relativas, em todos os tipos de encaixamento sintático (sujeito, objeto direto, objeto indireto etc.) e procurou identificar a produtividade da estratégia resumptiva no PE ao longo dessas funções. Alexandre (2000: 14-15) reconhece que as resumptivas são pouco produtivas na língua, se comparadas às relativas preposicionadas padrão e cortadora, mas assume que “o facto de os falantes utilizarem com alguma frequência este tipo de estratégia não-canónica, porque ela se encontra disponível na gramática desses falantes, motivou o interesse pela descrição e compreensão dos mecanismos que subjazem a este recurso”. De fato, ao considerarmos todas as funções sintáticas catalogadas pela autora, as relativas resumptivas ocorrem em apenas 0,36% dos casos, sendo sensivelmente mais frequentes nas posições sintáticas mais baixas, conforme a hierarquia de acessibilidade de Keenan & Comrie (1977). Também as relativas cortadoras são apresentadas, na análise da autora, com um reduzidíssimo percentual de ocorrência: 1,21%. No entanto, o trabalho de Alexandre (2000) incorreu no erro metodológico de computar os percentuais de ocorrência de *pied-piping*, cortadoras e resumptivas juntamente aos de relativas em função de sujeito, objeto direto, genitivo etc. Ora, nos dados do CRPC, se analisarmos somente os ambientes sintáticos

preposicionados, única posição estrutural em que *pied-piping* e cortadoras podem ocorrer, as relativas cortadoras chegam a 34% de ocorrência. Do total de 18.500 relativas analisadas por Alexandre, apenas 3,5% correspondem à extração de constituinte preposicionado. São, portanto, 648 relativas preposicionadas, das quais 392 eram *pied-piping*, 222 eram cortadoras e 34 eram resumptivas.¹ Trata-se de um quadro muito diferente do apresentado pela autora em sua dissertação. Como se vê, metade das 68 resumptivas encontradas no *corpus* ocorre em ambiente preposicionado, no qual a relativização cortadora foi empregada em relevantes 34% dos casos. Naturalmente, *pied-piping* é ainda a relativização preposicionada mais produtiva no *corpus*, fato que pode ser explicado por se tratar de dados colhidos entre pessoas letradas em situação de fala mais ou menos monitorada, contexto sociolinguístico que torna ainda mais significativo os quase 40% de relativas não-padrão catalogados. Mais uma vez, o trato com dados de *corpus* pode levar a diferentes conclusões, a depender do perfil sociocultural e textual do *corpus* e a depender da atenção ao trabalho estatístico dispensado aos dados.

2. Enfraquecendo a hipótese paramétrica com dados de *corpus* do PE

A pesquisa de Arin, Ramilo e Freitas (2005) tratou das relativas preposicionadas presentes no *corpus* REDIP (*Rede de Difusão Internacional do Português – rádio, televisão, jornais e revistas*). Os autores passaram em revista os trabalhos de Peres e Mória (1995) e Alexandre (2000) para defender que as relativas resumptivas apresentam ocorrência muito baixa no REDIP, por oposição à media recorrência das cortadoras e à alta produtividade de *pied-piping* naquele *corpus*, conforme representado na tabela a seguir.

¹ A autora, em comunicação pessoal, admitiu que, em seu trabalho de 2000, não chegou a analisar cortadoras e resumptivas exclusivamente em função preposicionada, isolando-as, para o cálculo estatístico, das relativas não-preposicionadas. Alexandre reconheceu que a subida percentual de relativas cortadoras e resumptivas que ocorre quando analisadas em separado corresponde mais fielmente à sua intuição como falante do PE e também ao que ela vê ocorrer no seu atual estudo sobre crioulos de base portuguesa, em que relativas Ppp nunca ocorrem. Como a autora já não dispõe de seu excerto com as 18.500 do CRPC, tivemos de reconstituí-lo recorrendo diretamente ao *corpus*. Ao catalogarmos a relativa número 18.500, chegávamos também à 222ª cortadora, que correspondia a 1,21% do total de relativas, exatamente como reportou Alexandre (2000: 24). Portanto, acreditamos que o *corpus* do CRPC que reanalisamos corresponde ao analisado por Alexandre (2000).

Tipo de relativa	Casos atestados	%
Pied-piping	189	71
Cortadora	74	28
Resumptiva	2	1
Total	265	100

Tabela 2: dados das relativas do REDIP (Arin, Ramilo e Freitas, 2005: 70).

Os quase 30% de ocorrência de relativas cortadoras no REDIP podem ser interpretados como altos, se considerarmos as características do *corpus*: registro de língua escrita da mídia (jornal e revista), de língua escrita para leitura (noticiários de TV) e de língua falada mais ou menos monitorada (programas de auditório e entrevistas). Todas essas são modalidades em que o treinamento com a língua formal ensinada pela escola melhor se manifesta. Ainda assim, ao analisarmos no REDIP os casos de língua oral (mesmo que formal, como em textos lidos em telejornal), os dados são comparáveis aos resultados do PB da pesquisa de Tarallo (1983): as cortadoras atingem o considerável percentual de 41%, por oposição aos baixos 3% registrados na língua escrita. É interessante notar que, em certas circunstâncias, o REDIP registra cortadoras em maior percentual que o de *pied-piping*. Por exemplo, quando o verbo da relativa é “precisar”, que seleciona a preposição “de”, as cortadoras ocorrem em 100% dos casos, mesmo na língua escrita. Com o verbo “gostar”, a preposição “de” é cortada em 83%, mesmo percentual de corte que ocorre com a preposição “a”, subcategorizada por “chamar”. Com o verbo “falar”, a supressão de “de” acontece em 74% dos casos.

Dados como esses parecem evidenciar que a relativa cortadora parece ser natural no uso espontâneo do PE, por oposição ao uso normativizado de *pied-piping* – algo incompatível com a hipótese paramétrica. Mais uma vez, devemos levar em consideração que os dados do *corpus* em questão são enviesados e favorecem uma produção mais monitorada, próxima do padrão idealizado da escrita, o que, entretanto, não impede a alta produtividade das cortadoras. Em resumo, certos dados de *corpus* do PE podem apresentar dados idênticos aos dados do PB explorados pelos defensores da hipótese paramétrica.

3. Refutando a hipótese paramétrica com dados de *corpus* do PE

A tese de doutoramento de Varejão (2006) analisou fenômenos de variação morfossintática no PE (concordância verbo-nominal e estratégias de relativização), com o objetivo de verificar se os fenômenos típicos da variação dialetal do PB (como a falta de concordância e as relativas cortadoras) não seriam também encontráveis na variação dialetal do PE. Os dados utilizados por Varejão foram extraídos do *Corpus Dialectal para o Estudo da Sintaxe* (Cordial-Sin), que registra, em 4.500 horas de gravação, a fala de indivíduos analfabetos e indivíduos com baixa escolarização espalhados por 200 localidades não-urbanas do território português. As gravações desse *corpus* foram feitas ao longo das décadas de 80 e 90, e contaram com a participação de 120 informantes portugueses (75 homens e 45 mulheres), que travavam diálogos livres ou semi-dirigidos com os entrevistadores.

Do total das 3315 relativas registradas no Cordial-Sin, apenas 108 eram relativas preposicionadas. Relativas oblíquas não-preposicionadas (mediadas pelos pronomes *onde*, *aonde* ou *quando*) compõem um universo à parte, com 102 ocorrências. Entre as 108 relativas preposicionadas de seus dados, Varejão (2006: 131) catalogou 74 cortadoras, 29 resumptivas e apenas 5 *pied-piping*. Isso quer dizer que as relativas cortadoras alcançam quase 70% de ocorrência entre os portugueses com baixa escolarização ou analfabetos, seguidas das resumptivas com 27%. Os notáveis 3% de ocorrência de Ppp são explicáveis por se tratar de estruturas estereotipadas (“altura em que”, “momento em que”, “lugar em que”), que devem ser melhor analisadas como construções fixas ou conjunções complexas, e não como legítimas orações relativas. As relativas *pied-piping* chegaram a apenas 5 ocorrências entre as 3315 relativas do Cordial-Sin, e não por acaso se encontram num diálogo sobre a vida escolar, no registro de um informante que apresentava algum nível de instrução (não indicado precisamente no *corpus*).

A tese de Varejão indica que as relativas *pied-piping* parecem não existir entre indivíduos portugueses analfabetos ou com pouca escolarização. A autora nos traz forte evidência de que as relativas preposicionadas padrão em PE são uma realidade dos centros urbanos, entre indivíduos escolarizados, à semelhança do que ocorre, *mutatis mutandis*, no PB. Os dados de Varejão indicam que a realidade do PB apontada pelos estudiosos da hipótese paramétrica não é essencialmente diferente do que ocorre em PE, desde que o fator letramento/escolarização seja adequadamente isolado na análise dos dados extraídos de *corpus* de língua oral.

4. Análise experimental das relativas do PB e do PE

Conforme procuramos sustentar nas seções anteriores, a hipótese paramétrica sobre as relativas do PB e do PE parece estar sem solução quando conduzida exclusivamente por pesquisas baseadas em *corpus* ou em intuições lingüísticas. Como vimos, a depender do *corpus* tomado para análise, há evidências em favor da hipótese da separação ou da unificação entre a gramática da relativização do PB e do PE. Nesta seção, traremos a problemática das *relativas do português* para o campo da pesquisa experimental. Sustentamos que as contradições das análises de *corpus* citadas podem ser superadas por testes experimentais, já que esses permitem o controle objetivo de inúmeras variáveis envolvidas no desempenho lingüístico e, por isso, são capazes de reduzir a ação da aleatoriedade nos dados disponíveis ao pesquisador, possibilitando uma análise mais apurada das representações lingüísticas subjacentes aos dados da *performance*. Para ilustrar essa argumentação, descreveremos a pesquisa de Kenedy (2007), que testou experimentalmente a hipótese paramétrica por meio do confronto entre o desempenho de portugueses e de brasileiros nos experimentos psicolingüísticos de percepção (*Julgamento imediato de gramaticalidade*) e de processamento (*Leitura automonitorada*) de orações relativas. Esses experimentos procuraram identificar se brasileiros e portugueses manifestam reações diferentes a estímulos lingüísticos como

relativas *pied-piping* ou cortadoras. Por questões de espaço, descreveremos a seguir os experimentos de *Leitura automonitorada* – remetemos o leitor a Kenedy (2008) para um resumo dos experimentos de *Julgamento imediato de gramaticalidade*.

4.1. Experimento de Leitura Automonitorada

Nesse experimento, os sujeitos que participaram da pesquisa foram levados a ler frases na tela do computador e responder perguntas interpretativas a respeito delas. Cada frase era apresentada em quatro partes, chamadas segmentos, que surgiam na tela conforme fossem solicitadas pelos sujeitos, mediante o acionamento de uma tecla. Para o início da leitura, com a apresentação do primeiro segmento, o sujeito deveria pressionar uma barra branca no teclado do computador. Após a leitura do primeiro segmento, deveria novamente pressionar a barra branca para que o segundo fosse apresentado, e assim procederia até que o quarto e último segmento fosse lido – após o qual uma pergunta interpretativa sobre a frase em leitura seria feita. Essa técnica experimental é tradicionalmente interpretada como *on-line*, já que permite a medição dos tempos de leitura dos segmentos de maneira muito próxima a seu processamento em tempo real na mente dos sujeitos. Assume-se que quanto mais rápida seja a leitura de um dado segmento, mais fácil será o seu processamento, sendo o inverso também verdadeiro: quanto maior for o tempo de leitura de certo segmento, mais custoso será o seu processamento psicolinguístico.

Criamos um desenho experimental de tal forma que um dos segmentos a ser apresentado aos sujeitos contivesse apenas duas palavras, a seqüência *preposição + pronome*, como */de que/*, */com quem/*, */para onde/* etc., que caracteriza as estruturas “*prepositional pied-piping*” em português. Essa seqüência (Ppp) ora se inseria no contexto sintático de uma oração interrogativa, ora, numa oração relativa. Como, em orações interrogativas, Ppp é uma estrutura natural para falantes do PB e do PE, não estando em variação em nenhuma das duas grandes modalidades continentais do português, independente da escolarização dos falantes,

espera-se que os tempos de leitura de Ppp em interrogativas sejam mais rápidos do que os de Ppp em relativas, já que, nessas, a seqüência *preposição + pronome* deve ser considerada problemática, como sugerem os estudos de *corpus* aqui citados. Com essa hipótese em mente, elaboramos o experimento, que pode ser caracterizado como um teste bastante simples, com apenas uma variável independente, o encaixamento de Ppp (oração relativa ou interrogativa).

Na tabela 3 a seguir, ilustram-se as duas condições do experimento, com destaque para o segmento crítico, em negrito, no qual se encontra a estrutura Ppp, e para os demais segmentos, separados por barras oblíquas. Abaixo das duas condições, apresenta-se a pergunta interpretativa relativa a ambas.

(a) Ppp EM INTERROGATIVAS: A tia olhou para o sobrinho: / de quem / o menino / tinha recebido o chocolate?
(b) Ppp EM RELATIVAS: A tia não conhecia o senhor / de quem / o menino / tinha recebido o chocolate.
Pergunta interpretativa para ambas as condições: A tia deu um chocolate para o menino?

Tabela 3: condições experimentais da *Leitura automonitorada*.

Se Ppp for um objeto sintático estranho para os brasileiros, devemos esperar que os tempos de leitura de Ppp na condição (b) sejam significativamente maiores do que em (a). Se, para os portugueses, Ppp é um construção natural, como assume a hipótese paramétrica, então não deve haver diferenças entre os tempos de (a) e (b). A principal variável dependente do experimento é, portanto, o tempo de leitura do segmento crítico, que interpretamos como uma *medida on-line*. Não obstante, também a resposta à pergunta interpretativa foi considerada como uma variável, que corresponde a uma *medida off-line*.

(a) TEMPO DE LEITURA DO SEGMENTO CRÍTICO (<i>medida on-line</i>)
(b) ÍNDICE DE ERROS NAS PERGUNTAS INTERPRETATIVAS (<i>medida off-line</i>)

Tabela 4: variáveis dependentes da *Leitura automonitorada*.

4.1. Método

Procurou-se assegurar no experimento o máximo de controle na metodologia aplicada na condução dos testes, seja na observação das variáveis envolvidas na construção das frases, seja na seleção dos sujeitos. Brasileiros e portugueses foram submetidos a um questionário

sociocultural, para que os marcadores *urbano* e *antecedentes de escolaridade na família* fossem confirmados. A observação desse perfil é necessária para que desempenhos assimétricos na tarefa não possam ser atribuídos a diferenças no nível de letramento/escolarização dos sujeitos dos conjuntos do PE e do PB.

4.2. Participantes

4.2.1. Sujeitos do PE

Para a realização do experimento, os sujeitos portugueses e brasileiros foram divididos em 3 grupos de escolaridade, correspondentes ao ensino fundamental (9 anos de escolarização), médio (12 anos de escolarização) e superior (mínimo de 16 anos de escolarização). Essa segmentação teve o objetivo de permitir a comparação entre os níveis de escolaridade mais baixos e mais altos, testando a hipótese de que Ppp em relativas seja uma estrutura mais problemática no nível fundamental, ao passo que Ppp em interrogativas não representaria problema para nenhum dos três grupos.

Os 20 sujeitos portugueses com o correspondente ao nível fundamental foram selecionados entre os alunos da Escola Secundária do Lumiar, instituição pública de ensino situada em Lisboa, onde cursavam o 10º ano escolar. 12 deles eram do sexo feminino e 8, do sexo masculino. A idade média registrada foi de 14 anos.

Participaram dos testes no nível médio, 20 estudantes de primeiro ano do curso de Lingüística da Faculdade de Letras da Universidade de Lisboa, 10 do sexo masculino e 10 do sexo feminino, com idade média de 19 anos. Os 20 sujeitos de nível superior eram estudantes do último ano de diferentes cursos de graduação do Instituto de Ciências Social da Universidade de Lisboa, e também do curso de Lingüística, da Faculdade de Letras da mesma Universidade. Dentre esses sujeitos, 11 eram mulheres e 9 homens, com idade média de 24 anos.

Todos os sujeitos de todos os níveis declararam ter nascido e serem residentes em Lisboa, bem como terem sido criados em ambiente letrado, com pais e/ou responsáveis formados no nível médio ou superior.

4.2.2. Sujeitos do PB

Os sujeitos brasileiros com nível fundamental foram selecionados entre os alunos da oitava série da Escola Municipal Paulo Freire, localizada na cidade de Niterói (Região Metropolitana do Rio de Janeiro). Foram 10 rapazes e 10 moças com idade média de 15 anos. No nível médio, foram testados calouros do 1º semestre de 2007 do curso de Letras da Universidade Estadual do Rio de Janeiro. 8 homens e 12 mulheres, com faixa etária média de 19 anos de idade. Os sujeitos de nível superior foram selecionados entre os formandos do curso de Letras, do 1º semestre de 2007, também da Universidade do Estado do Rio de Janeiro. 7 do sexo feminino e 13 do sexo masculino, com idade média de 25 anos. Em todas as faixas de escolaridade os sujeitos declararam ser nascidos e residentes na região urbana de Niterói ou São Gonçalo, bem como confirmaram criação em ambiente de fala escolarizada, tendo sido seus pais e/ou responsáveis pessoas com nível de instrução médio ou superior.

4.3. Materiais

10 frases experimentais compuseram o experimento. Com esse número, foi possível que cada sujeito de cada nível testado fosse exposto 5 vezes às duas condições do teste. Adotou-se a distribuição *between subjects*, de forma que um mesmo sujeito não tivesse acesso à versão *relativa* e *interrogativa* de uma mesma frase. Assim, um sujeito que, por exemplo, fosse exposto à condição (2) abaixo, não veria a versão relativa dessa, mas, sim, outra relativa não-relacionada, como (3).

2) A tia olhou para o sobrinho: / **de quem** / o menino / tinha recebido o chocolate?

Pergunta: A tia deu um chocolate para o menino?

3) O aluno conhece os rapazes / **a quem** / o professor / tinha entregado o livro.

Pergunta: O professor ainda tinha o livro consigo?

Às 10 frases experimentais, juntaram-se 30 frases distratoras, que tinham a função de impedir que o sujeito percebesse que relativas e interrogativas Ppp eram o objeto de análise do teste. Todas as 10 frases experimentais deveriam receber *não* como resposta correta. Dessa forma, respostas *sim* para essas frases seriam interpretadas como incorretas, e seriam tomadas como evidência de problemas de interpretação. Para acontecer o balanceamento entre o total de respostas *sim* e *não* esperadas no experimento, 20 frases distratoras deveriam ser respondidas com *sim* e 10 deveriam receber *não* como resposta.

4.4. Procedimentos

Cada sujeito recebeu, individualmente, instruções para a realização do experimento, que eram apresentadas oralmente, na interação direta com o experimentador, e também por escrito, ao início da tarefa, na tela do computador. Todos participavam, também, de um pré-teste (treinamento), realizado diante do experimentador, em que ocorriam apenas frases distratoras e que possuía o mesmo *design* do experimento. Esse pré-teste tinha intenção de confirmar o perfeito entendimento, por parte do sujeito, da tarefa a ser desempenhada. O experimento real só tinha início quando o sujeito demonstrasse ter compreendido completamente a tarefa a que se submeteria.

Ao pressionar-se a barra branca destacada no teclado, surgia na tela do computador o primeiro dos quatro segmentos de cada frase a ser lida pelo sujeito. Com a conclusão da leitura de um segmento, o sujeito deveria pressionar novamente a barra branca para autorizar a apresentação do próximo, e assim deveria proceder até que todos os segmentos tivessem sido lidos. Após a leitura do último segmento, a pergunta interpretativa era apresentada na tela. O sujeito deveria pressionar, no teclado do computador, a tecla verde, caso considerasse

que a resposta para a pergunta fosse *sim*, e a tecla vermelha caso considerasse *não* como a resposta correta.

Em Portugal, o experimento foi aplicado nos meses de fevereiro e março de 2007, no Laboratório de Psicolinguística da Universidade de Lisboa, para os sujeitos do ensino médio e superior, e na Sala dos Professores da Escola Secundária do Lumiar, para os sujeitos do ensino fundamental. Todas as salas apresentavam ambiente confortável e silencioso, adequado para a concentração dos sujeitos na tarefa. No Brasil, os sujeitos de nível médio e superior foram testados nos meses de julho de 2007, na sala 309 do Departamento de Letras da Faculdade de Formação de Professores (Universidade do Estado do Rio de Janeiro), ambiente adequado para a realização do experimento. No nível fundamental, o experimento foi realizado nos meses de julho e agosto de 2007, na sala de reuniões da Escola Municipal Paulo Freire, que apresentava condições adequadas de iluminação e silêncio necessários para a concentração dos sujeitos.

O experimento foi elaborado e rodado nos seguintes equipamentos: Psycope versão X – B46, Sistema Operacional X do computador PowerBook G4 (*laptop* da Apple, Macintosh), tela LCD de 15". Palavras apresentadas em fonte 25, na cor preta, com fundo branco. O relógio interno do computador registrou os tempos de leitura de cada segmento das frases.

4.5. Resultados

4.5.1. Dados do PE

4.5.1.1. Ensino Fundamental

No que diz respeito à medida *off-line*, os sujeitos do PE com o equivalente ao ensino fundamental apresentaram 39% de erros nas perguntas interpretativas quando as frases lidas continham relativas Ppp. Esse percentual foi considerado significativamente superior aos 9% de erros registrados quando se liam frases em que ocorriam interrogativas Ppp [$F_1(1,20) = 13,64; p < .01$].

Quanto ao tempo de leitura do segmento crítico, foi detectada diferença relevante entre os 1021 msg médios demandados para a leitura de Ppp em interrogativas e os 1996 msg despendidos nas relativas [$F_1(1,20) = 188,54; p < .01$; $F_2(1,5) = 219,65; p < .01$].

4.5.1.2. Ensino Médio

Entre os sujeitos portugueses com escolaridade média, a medida *off-line* também se mostrou relevante na distinção entre relativas e interrogativas. Enquanto interrogativas apresentaram apenas 3% de erro nas respostas às perguntas de interpretação, relativas provocaram 26% de respostas equivocadas, diferença estatisticamente significativa [$F_1(1,20) = 27,65; p < .01$]. Na medida *on-line*, o segmento Ppp apresentou tempos de leitura mais lentos nas relativas (1314 msg) no cotejo com o mesmo segmento nas interrogativas (1088 msg) – diferença detectada como relevante no teste estatístico: [$F_1(1,20) = 15,66 p < .01$; $F_2(1,5) = 17,36; p < .01$].

4.5.1.3. Ensino Superior

As diferenças de reação a frases que continham relativas ou interrogativas persistiram mesmo entre sujeitos com nível superior. Na medida *off-line*, as relativas registraram 19% de erros, por oposição aos 5% de respostas equivocadas registradas nas interrogativas [$F_1(1,20) = 22,65; p < .01$]. Com relação à medida *on-line*, o segmento crítico em interrogativas foi lido com a média de 1027 msg, tempo bastante inferior aos 1670 msg médios registrados na leitura de Ppp em relativas [$F_1(1,20) = 79,49; p < .01$; $F_2(1,5) = 512,66; p < .01$].

4.5.2. Dados do PB

4.5.2.1. Nível Fundamental

Os sujeitos brasileiros com escolaridade fundamental apresentaram sensível diferença nas respostas interpretativas, considerando-se as condições *relativa X interrogativa*. As perguntas feitas quando orações interrogativas haviam sido lidas receberam apenas 11% de respostas erradas, por oposição aos 36% de erros verificados nas respostas a perguntas cujas

frases continham orações relativas Ppp, diferença importante para a análise estatística [$F_1(1,20) = 13,64; p < .01$]. A medida *on-line* do experimento também se mostrou significativa. Os sujeitos brasileiros com escolaridade fundamental apresentaram leitura do segmento crítico em interrogativas com a média de 734 msg, um tanto inferior aos 927 msg médios registrados na leitura de Ppp em relativas. Tal diferença foi interpretada como significativa pela ANOVA bivariada [$F_1(1,20) = 13,42; p < .05; F_2(1,5) = 13,06; p < .05$].

4.5.2.2. Nível Médio

Os erros nas respostas interpretativas cometidos pelos sujeitos do PB com nível médio foram menores quando a frase alvo da respectiva pergunta continha uma oração interrogativa, em comparação ao encaixamento de Ppp que envolvia uma oração relativa. Interrogativas registraram 10% de erros, ao passo que, para as relativas, registraram-se 24%. Mais uma vez, esses resultados foram submetidos à análise estatística, cujos resultados indicam que a diferença dificilmente pode ser aleatória: [$F_1(1,20) = 13,64; p < .01$]. Nos tempos de leitura do segmento crítico, foi verificado que relativas consumiam em média 1164 msg, enquanto interrogativas registraram média de 942 msg, diferença significativa [$F_1(1,20) = 14,62; p < .01; F_2(1,5) = 17,64; p < .01$].

4.5.2.3. Nível Superior

Nos sujeitos do PB representantes do nível superior, os erros nas repostas interpretativas chegaram a 22% nas frases com orações relativas, exatamente o dobro dos 11% ocorridos em frases com orações interrogativas [$F_1(1,20) = 35,28; p < .01$]. Também na leitura do segmento crítico, as diferenças entre relativas e interrogativas mostraram-se significativas. Ppp em relativas registrou tempo médio de 1061 msg, enquanto em interrogativas a média foi de 802 msg [$F_1(1,20) = 18,45; p < .01; F_2(1,5) = 19,98; p < .01$].

4.6. Discussão

Os resultados do experimento de *Leitura automonitorada* parecem indicar que os sujeitos portugueses e brasileiros apresentam o mesmo tipo de reação Ppp. Ambos manifestaram um estranhamento quando essa estrutura é encaixada numa oração relativa, por oposição à naturalidade com que reagem ao encaixe de Ppp em orações interrogativas.

Com relação à medida *off-line*, verificou-se que o número de erros nas respostas às perguntas interpretativas é significativamente maior quando tais perguntas têm como alvo frases em que relativas Ppp estão encaixadas, por contraste a frases com o encaixamento de interrogativas Ppp. As relativas superam as interrogativas no percentual de respostas erradas em todos os segmentos escolares testados, tanto no PE quanto no PB. É importante notar que os erros cometidos nas frases com encaixamento de interrogativa Ppp não são significativamente diferentes entre os níveis de escolaridade analisados: no caso do PB, os erros chegam a 11% nos ensinos fundamental e superior, e a 10% no ensino médio, enquanto, no PE, não passam de 9% no ensino fundamental, 5% no ensino médio e 3% no nível superior (todos os $p > .05$). Como são praticamente os mesmos em todos os níveis, parece correto interpretar que tais erros encontram-se dentro da faixa de aleatoriedade previsível em testes com respostas *sim* ou *não*. Já as relativas Ppp apresentam um quadro bastante diferente. Os erros na interpretação de relativas decrescem de maneira significativa conforme os sujeitos apresentem maior grau de escolaridade. No caso do PE, os 39% registrados no ensino fundamental caem para 26% no ensino médio e para 19% no nível superior. No PB, os 36% de erros registrados no ensino fundamental são reduzidos para 24% e 22% nos ensinos médio e superior, respectivamente ($p < .05$).

No que diz respeito à medida *on-line*, foi verificado um efeito principal da variável *estrutura* nas diferenças entre os tempos de leitura de Ppp em relativas e em interrogativas,

tanto nos sujeitos do PE quanto do PB. Independente do nível de escolaridade, portugueses e brasileiros lêem Ppp mais lentamente quando inserido em relativas.

Esses dados parecem indicar que as relativas Ppp são igualmente problemáticas para brasileiros quanto para portugueses, tanto em termos *qualitativos* quanto *quantitativos*. Os dados do experimento sustentam a hipótese de relativas Ppp são uma criação artificial da escola, que os falantes do português podem ou não vir a dominar e tornar-se fluentes, a depender de seu perfil sociocultural e de seu treinamento nos gêneros mais formais e artificiais. Ou seja, com a aplicação dessa metodologia experimental, a hipótese paramétrica parece ceder lugar à hipótese da antinaturalidade das relativas Ppp (cf. Kenedy, 2007).

5. Considerações finais

Com este artigo, argumentamos que as pesquisas baseadas em *corpus*, embora desempenhem crucial importância para a identificação de padrões no uso de uma língua, podem não ser uma fonte segura para generalizações a respeito da competência linguística dos falantes. *Corpora* são tão variáveis quanto a heterogeneidade dos seres humanos e dos gêneros textuais pelos quais se expressam, o que significa que cotejar *corpora* com características muito diferentes pode levar a erros. Apresentamos o recurso à pesquisa experimental, conforme a metodologia da psicolinguística, como um bom recurso à linguística comparativa, especialmente com relação à querela PB *versus* PE.

Corpus analysis, the linguist's intuition and experimental methodology in researches on BP and EP relative clauses

ABSTRACT: This article takes into consideration the case of relative clauses of BP and EP to show the limitations of researches based solely on corpus analysis or on linguist's intuitions. It is argued that the methodology of experimental research in psycholinguistics is an appropriate way both to test the psychological reality of theoretical models as to overcome the contradictions in corpora analysis.

Keywords: research methodology in linguistics, corpus analysis, experimental psycholinguistics.

Referências

ALEXANDRE, N. 2000. *A Estratégia resumptiva em relativas restritivas do português europeu*. Faculdade de Letras, Universidade de Lisboa. (Dissertação de mestrado)

ARIM, E., RAMILO, M. & FREITAS, T. 2005. Mudança em curso e os média: o caso das relativas. IN.: MATEUS, M. & Do NASCIMENTO, F. (eds) *A língua portuguesa em mudança*. Lisboa: Caminho.

CHOMSKY, N. 1965. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

DOOLEY, R. & LEVINSOHN, S. 2001. *Analyzing discourse: a manual of basic concepts*. Dallas: SIL International.

GALVES, C. 2001. *Ensaio sobre as gramáticas do português*. Campinas: Editora da Unicamp.

KATO, M. 1993. Recontando a história das relativas em uma perspectiva paramétrica. In. ROBERTS, I. & KATO, M. (orgs.) *Português Brasileiro: uma viagem diacrônica*. Homenagem a Fernando Tarallo. Campinas: Unicamp. ed. 1996

KEENAN, E. & COMRIE, B. 1977. Noun phrase accessibility and universal grammar. *Linguistic Inquiry* 8.63–99.

KENEDY, E. 2007. *A hipótese da antinaturalidade de pied-piping em orações relativas*. RJ: UFRJ (Tese de doutorado).

KENEDY, E. 2008. As orações relativas preposicionadas e a hipótese da antinaturalidade de pied-piping. *Veredas (UFJF)*, v. 1, p. 92-111.

PERES, J. & MÓIA, T. 1995. *Áreas Críticas da Língua Portuguesa*. Lisboa: Caminho.

PONTES, E. 1987. *O tópico no português do Brasil*. Campinas: Pontes.

ROBERTS, I. & KATO, M. (orgs.) 1993. *Português Brasileiro: uma viagem diacrônica*. Homenagem a Fernando Tarallo. Campinas: Unicamp. ed. 1996

TARALLO, F. 1983. *Relativization Strategies in Brazilian Portuguese*. Philadelphia Univ. of Pennsylvania (Doctoral dissertation).

TARALLO, F. 1985. The filling of the gap: Pro-drop rules in Brazilian Portuguese. In KING, L. & MALEY, A. (eds.) *Selected Papers from the XIIIth Linguistic Symposium on Romance Languages*. Amsterdam/Philadelphia: John Benjamins Publishing Co.

VAREJÃO, F. 2006. *Variação em estruturas de concordância verbal e em estratégias de relativização no português europeu popular*. RJ: UFRJ. (Tese de doutorado).