Revista de Estudos da Linguagem, Belo Horizonte, v.23, n.1, p. 211-244, 2015

Estudo sobre um parâmetro de tarefa e um parâmetro amostral para experimentos com julgamentos de aceitabilidade temporalizados[1]

Inquiry of a Task Parameter and a Sampling Parameter for Speeded Acceptability Judgments Experiments

Ricardo Augusto de Souza
Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Minas Gerais, Brasil.
ricsouza.ufmg@gmail.com

Cândido Samuel Fonseca de Oliveira
Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG), Contagem, Minas Gerais, Brasil.
coliveira.ufmg@gmail.com

Jesiel Soares-Silva
Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Minas Gerais, Brasil.
fassiedojeb@yahoo.com.br

Alberto Gallo Araújo Penzin
Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Minas Gerais, Brasil.
albertopenzin@hotmail.com

Alexandre Alves Santos
Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Minas Gerais, Brasil.
alexandrealves_santos@yahoo.com

[1] Este estudo contou com financiamentos da CAPES (BEX 4087/10-0), do CNPq (485285/2013-4) e da FAPEMIG (APQ 20038). Os autores manifestam seus agradecimentos.

eISSN: 2237-2083
DOI: 10.17851/2237.2083.23.1.211-244

Resumo: A tarefa de julgamento de aceitabilidade de sentenças temporalizadas é uma das técnicas de eliciação de julgamentos na qual se impõem restrições temporais aos juízes. Propõe-se que essa técnica proporciona observações mais fidedignas de conhecimento implícito e processos automatizados. Este estudo explora a definição de tetos temporais mínimos para a execução dessa tarefa por falantes nativos das línguas dos estímulos, assim como avalia o impacto do recrutamento de amostras de conveniência, formadas por participantes com treinamento em estudos de linguagem, sobre esse tipo de experimento. Os resultados indicam não haver impacto crítico dessa forma de amostragem de conveniência, e que efeitos de gramaticalidade são detectáveis com janelas temporais de 4 segundos para cada sentença.

Palavras-chave: Aceitabilidade; Gramaticalidade; Julgamentos temporalizados; Amostragem de conveniência.

Abstract: The speeded sentence acceptability judgment task is a technique for the elicitation of judgments in which temporal constraints are imposed on judges. It is suggested that such technique provides more reliable observations of implicit knowledge and automatic processes. This study explored the setting of minimal temporal ceilings for performance in the speeded acceptability judgment task by native speakers of the stimuli languages, and it also assessed the impact of convenience sampling where participants with language studies backgrounds are recruited. The results show that there is no critical impact of this kind of convenience sampling, and they also show that grammaticality effects are detectable within a time window of 4 seconds per sentence.

Keywords: Acceptability; Grammaticality; Speeded Judgments; Convenience Sampling.

Recebido em 18 de março de 2014.
Aprovado em 28 de maio de 2014.

Introdução

A demonstração de que construções e arranjos de unidades linguísticas potencialmente realizáveis tendem a ser rejeitadas pelos falantes oferece pistas sobre as restrições em operação no conhecimento linguístico. Assim, não é surpreendente que a observação de julgamentos sobre o estatuto de gramaticalidade, ou aceitabilidade, de dados linguísticos constitui uma matriz metodológica importante em várias abordagens sobre a natureza do conhecimento linguístico, conhecimento esse cujo desvendamento é uma das tarefas primeiras da ciência linguística. A partir especialmente do surgimento do movimento gerativista, a aceitação dessa matriz atingiu por vezes patamares extremos, passíveis de caracterização como pleno introspeccionismo subjetivista, com os julgamentos individuais do próprio proponente de análises como dados apresentados, sem problematização, como suficientes para a justificativa dessas propostas (HARRIS, 1995; WASOW; ARNOLD, 2005). Não obstante, tal como foi relatado em Schütze (1996), questionamentos sobre a validade e a confiabilidade dos julgamentos de aceitabilidade individuais, como base empírica para o avanço da teoria linguística, têm sido levantados desde os primórdios da proliferação de estudos neles prioritariamente fundamentados.

A busca de refinamento metodológico, em bases empíricas consistentes, da coleta e tratamento de dados oriundos da observação de julgamentos de aceitabilidade gerou, em anos recentes, um intenso debate sobre seus limites e potencialidades. Gibson e Fedorenko (2013), por exemplo, argumentam haver evidências de que a ausência de rigor metodológico no tratamento de julgamentos de gramaticalidade resultou em propostas espúrias em teoria da gramática. Os autores defendem a utilização de desenhos experimentais rigorosos, assim como tratamento estatístico criterioso, como estratégia indispensável para que estudos baseados em tais julgamentos atinjam níveis aceitáveis de confiabilidade, validade e cientificidade. Esses argumentos vêm ao encontro de outras propostas de adequação do julgamento de aceitabilidade ao método experimental, o que é por vezes denominado "sintaxe experimental" (COWART, 1997; MAIA, 2012).

O presente estudo tem por objetivo geral trazer contribuições ao empreendimento de fomento à articulação de bases rigorosas para a investigação das representações linguísticas dos falantes através da

observação sistemática de julgamentos sobre o estatuto de gramaticalidade de dados linguísticos. Entendemos, portanto, que se trata de um estudo alinhado ao movimento direcionado à busca de construção do julgamento de aceitabilidade como desenho experimental para a pesquisa linguística.

Especificamente, um outro objetivo buscado através deste estudo foi a exploração de um parâmetro da tarefa experimental de julgamento de aceitabilidade: a imposição de uma restrição do teto temporal dentro do qual um participante deve emitir seu julgamento. Ao ser configurada com tal restrição, a tarefa fica caracterizada como uma variante que é denominada neste artigo "julgamento de aceitabilidade temporalizado". Tal variante é comumente empregada em estudos psicolinguísticos, nos quais, além do julgamento em si, há interesse na investigação do grau de dificuldade, ou custo cognitivo, imposto pela própria tarefa de julgamento de tipos particulares de estímulos (COOK, 1994). Nosso interesse na exploração desse parâmetro de configuração da tarefa foi a determinação de uma janela temporal tão restrita quanto possível para a emissão de julgamentos confiáveis. Nossa motivação para essa exploração foi a hipótese de que uma tarefa de eliciação de julgamentos rápidos, em associação à recomendação de Schütze (1996) sobre o treinamento prévio dos participantes para a tarefa, permitiria a aproximação mais confiável de respostas convergentes com a materialidade mais estritamente linguística dos estímulos; portanto, tais respostas mais adequadamente filtradas de impressões oriundas de conhecimentos não relevantes.[2]

Por fim, um outro objetivo específico deste estudo foi a exploração de um parâmetro amostral, ou seja, relativo ao recrutamento de participantes para estudos experimentais. Tal parâmetro amostral diz respeito à prática, por nós conjecturada como sendo de razoável frequência, de recrutamento de amostras de conveniência, compostas por estudantes e pessoal das próprias instâncias onde os pesquisadores atuam. Assim, acreditamos ser prevalente nos estudos linguísticos a

[2] Em um estudo com julgamentos de aceitabilidade não temporalizados, realizado há alguns anos por um dos autores do presente trabalho, os dados de um dos participantes foram descartados porque tal participante anotou em folha de respostas que seu julgamento de uma das sentenças como inaceitável fora motivado por sua interpretação de que o conteúdo proposicional da sentença refletia uma situação de abuso de poder. É a possibilidade de sobreposição desse tipo de conhecimento à resposta eliciada pela estruturação gramatical do estímulo que acreditamos poder diminuir com restrição temporal e treinamento para a execução da tarefa.

seleção de participantes que recebem ou receberam instrução explícita em análise e descrição linguística, ou seja, participantes com formação na área de Letras e Linguística. Se tal conjectura for representativa dos fatos, então há provavelmente a seleção de amostras que não poderiam ser caracterizadas como sujeitos plenamente inocentes, ou leigos, em relação aos objetos dos estudos. No estudo ora relatado, manipulamos diretamente o perfil de formação dos participantes, com vistas a colocar em teste a hipótese de que os participantes com formação na área de Letras e Linguística fariam julgamentos de aceitabilidade diferentes dos participantes com formação em áreas diferentes.

Na próxima seção, apresentaremos mais detalhes sobre as questões suscitadas pela adoção dos julgamentos de aceitabilidade como base empírica da pesquisa linguística, assim como de propostas de sua adequação aos controles típicos dos métodos experimentais. Em seguida, explicitaremos o desenho metodológico de dois experimentos a partir dos quais buscamos atingir os objetivos acima delineados. A terceira seção do artigo se destina à análise e à discussão das observações feitas em relação a esses dois experimentos. Por fim, concluiremos com a retomada dos objetivos que nortearam este estudo e com as considerações sobre a modalidade de tarefa de eliciação de julgamentos de aceitabilidade ora investigada.

1 Julgamentos do estatuto de aceitabilidade como metodologia experimental na pesquisa linguística

Tal como foi comentado acima, dentre as diversas formas de se estudar a linguagem humana, os métodos introspectivos são empregados com alta frequência em algumas subáreas da linguística (SCHÜTZE, 1996; FERREIRA, 2005; MYERS, 2009a). Dentre tais métodos, julgamentos informais acerca da aceitabilidade de sentenças são amplamente utilizados como fonte de dados em campos como a sintaxe e a semântica. Geralmente, as hipóteses são defendidas a partir de dados oriundos da introspecção do próprio linguista/autor sobre o *status* de aceitabilidade de um par mínimo de sentenças.

Tal prática tem sido alvo de uma série de discussões metodológicas no campo dos estudos linguísticos. Por um lado, alguns autores argumentam que os próprios linguistas podem fornecer dados mais confiáveis devido ao conhecimento técnico acerca da linguagem (PHILIPS; WAGERS, 2007

apud MYERS, 2009a; DEVITT, 2006 *apud* CULBERTSON; GROSS, 2009; PHILIPS, 2009 *apud* GIBSON; FEDORENKO, 2013). Além da confiabilidade que tal conhecimento poderia conferir a esses dados, esses autores ainda afirmam que os métodos introspectivos são mais práticos e funcionais, do ponto de vista da implementação e do gerenciamento. Por outro lado, existem autores que questionam tal procedimento, salientando os riscos do interpretativismo subjetivo para a confiabilidade das bases de construção do conhecimento científico (SCHÜTZE, 1996; COWART, 1997; FERREIRA, 2005; GIBSON; FEDORENKO, 2013). Myers (2009a, p. 406) assevera que os "julgamentos informais inspiram horror em muitos pesquisadores, devido ao fato de que eles claramente violam os protocolos metodológicos que são padrão no resto da ciência cognitiva empírica".[3]

Do ponto de vista da ciência experimental, o método da introspecção individual apresenta problemas como: (i) número reduzido e falta de controle de itens; (ii) número reduzido de participantes; e (iii) participantes familiarizados com os itens (CULBERTSON; GROSS, 2009; GIBSON; FEDORENKO, 2013). Assim, há críticos que afirmam que a utilização da introspecção como fonte principal de dados tem resultado no enfraquecimento dos laços que uniam algumas subáreas da linguística. De acordo com Ferreira (2005), por exemplo, tal prática foi um dos motivos pelos quais a psicolinguística e a sintaxe formal se distanciaram, principalmente após o surgimento do Programa Minimalista (PM). Segundo a autora, além de o PM ser incompatível com alguns fatos que já foram investigados por estudos psicolinguísticos – os processos de reanálise e, entre outros, o parseamento incremental da esquerda para a direita –, a principal fonte de dados que corroboram seus princípios são oriundos de julgamentos de aceitabilidade realizados pelos próprios teóricos.

Não obstante, a reflexão do linguista teórico e seus postulados sobre a organização da linguagem costumam ser os únicos pontos de partida para a explicitação de hipóteses cujas previsões serão, subsequentemente, submetidas a teste em bases empíricas mais consistentes e generalizáveis. Vários experimentos já confirmaram

[3] Nossa tradução para: "Informal judgments inspire horror in many scholars, as they so clearly violate the methodological protocols standard in the rest of the empirical cognitive sciences."

hipóteses oriundas da teoria da gramática de base introspeccionista que testavam, sendo exemplos Cowart (1997), Clifton *et al.* (2006) e Myers (2009b). Por outro lado, tal como foi demonstrado por Myers (2009a) e Gibson; Fedorenko (2013), várias dessas hipóteses, quando testadas empiricamente, foram falseadas.

Tais resultados contraditórios, em nosso entender, somente realçam a relevância do aprimoramento dos métodos de eliciação de julgamentos sobre dados linguísticos dos falantes. Ao se conferir a tais métodos máximo rigor e controle, acreditamos que será oferecida aos pesquisadores das mais diversas áreas e subáreas da linguística uma ferramenta poderosa e que é suplementar a outras metodologias relevantes para a edificação da linguística como ciência empírica. Portanto, é pertinente uma delimitação de características da abordagem experimental dos julgamentos de aceitabilidade, assim como de suas especificações.

Inicialmente, chamamos a atenção sobre a natureza do construto que as técnicas de eliciação de julgamentos visam a capturar. A denominação dessas técnicas oscila entre julgamento de aceitabilidade e julgamento de gramaticalidade. Embora ambas as terminologias por vezes sejam tratadas como sinônimas, elas não se referem necessariamente aos mesmos construtos. O termo *gramaticalidade* se refere a um construto teórico dentro da teoria linguística, ou seja, uma característica postulada como inerente às construções ou arranjos de unidades linguísticas, e da qual resulta a boa formação das mesmas, de acordo com tal postulação teórica. Já o termo *aceitabilidade* faz referência a um construto perceptual, ou seja, que se manifesta na percepção de um falante, acerca das construções ou arranjos de unidades linguísticas (BARD; ROBERTSON; SORACE, 1996). Em outras palavras, a aceitabilidade se refere a uma sensação consciente do participante, enquanto a gramaticalidade, a uma consequência lógica das premissas da teoria linguística, não sendo necessariamente acessível através da consciência (MYERS, 2009a).

A diferença entre esses dois construtos é comumente exemplificada em frases em inglês com múltiplas orações subordinadas, tais como "*the mouse the cat the dog saw chased ate*" (o rato que o gato que o cachorro viu perseguiu comeu). Do ponto de vista de alguns quadros da teoria da gramática, tal sentença é classificada como gramatical, uma vez que pode ser logicamente derivada de premissas desses quadros sobre a boa formação de arranjos sintáticos. Por outro lado, essa sentença dificilmente

será percebida como aceitável, provavelmente devido ao seu alto custo de processamento (CULBERTSON; GROSS, 2009).

Assim, o termo *julgamento de aceitabilidade* parece descrever de maneira mais acurada o procedimento metodológico no qual são as repostas dos falantes, a sua percepção sobre unidades linguísticas, que são efetivamente observadas. Myers (2009a) demonstra que os termos *julgamento de gramaticalidade* e *julgamento de aceitabilidade* surgiram, com a mesma frequência, na literatura linguística em língua inglesa, entre meados da década de 1970 e meados da década de 1990. Porém, a autora demonstra que em anos mais recentes parece haver emergido uma nítida prevalência do termo *julgamento de aceitabilidade* na literatura científica em língua inglesa. Em consonância com essa tendência, neste estudo, utilizar-se-à doravante apenas o termo *julgamento de aceitabilidade*.

Em termos gerais, o julgamento de aceitabilidade de sentenças como paradigma experimental consiste na observação, em uma amostra representativa de uma comunidade de fala, das avaliações intuitivas sobre a boa formação de sentenças, em uma determinada língua (KELLER, 1998). Os participantes do experimento são tipicamente apresentados a um conjunto de sentenças, devendo manifestar-se sobre o quão aceitável cada uma delas é. Em outras palavras, usualmente os participantes devem dizer se cada uma das sentenças soa bem ou mal (CULBERTSON; GROSS, 2009).

Obviamente, tal paradigma experimental pressupõe que a emissão da percepção sobre o bem ou mal "soar" de uma sentença está pautada em algum critério comum a toda a amostra de participantes, critério esse que pode ser delimitado através de instruções e treinamento sobre o que deve ser o foco da atenção de cada participante. Uma outra forma de realização do pressuposto de que um critério básico permeará as manifestações da amostra de participantes é a apresentação de vários itens representativos de condições linguísticas que modulam a hipotética diferença de gramaticalidade entre as sentenças. Adota-se, então, procedimentos analíticos de estatística inferencial, considerando-se se há diferenças entre medidas de tendência central (ex. média), em associação a medidas de dispersão (ex. desvio padrão) das manifestações observadas, o que revelaria efeitos da manipulação das condições linguísticas e, portanto, um efeito da gramaticalidade.

De acordo com Gibson e Fedorenko (2013), é imprescindível que os estudos experimentais baseados em julgamentos de aceitabilidade

sigam alguns parâmetros importantes. Um desses parâmetros diz respeito à configuração da amostra de participantes. Segundo os autores, é importante incluir na amostra diversos tipos de participantes, inclusive aqueles que são "ingênuos", ou leigos, em relação à linguagem como objeto de estudo.

Um dos efeitos possíveis da experiência com os estudos da linguagem é a perda de sensibilidade à agramaticalidade. Por exemplo, Barile e Maia (2008) e Maia (2013) relatam um estudo no qual julgamentos foram obtidos experimentalmente de um grupo de participantes formado por estudantes de Letras que haviam concluído um curso de teoria de sintaxe, no qual um tema específico foi precisamente o tipo de sentenças alvo dos julgamentos. Os julgamentos desse grupo foram comparados com julgamentos emitidos por estudantes de outros cursos. Nesse estudo, foi observado que os sujeitos que tiveram exposição prévia ao tipo de violação gramatical presente nas sentenças do experimento recusaram tais sentenças significativamente menos que os participantes "ingênuos".

Não obstante, é importante notar que tampouco tal orientação quanto à amostragem é plenamente consensual. Culbertson e Gross (2009), por exemplo, argumentam que não há diferenças relevantes entre participantes com conhecimento de linguística ou não, mas sim diferenças relacionadas à familiaridade com a tarefa à qual são apresentados os participantes.

Compreendemos que tal parâmetro amostral pode constituir-se em anátema para parte da comunidade dos pesquisadores em linguística, uma vez que percebemos como relativamente frequente a realização de estudos com amostragem de conveniência, ou amostragem acidental (COZBY, 2009). A amostragem por conveniência é uma técnica não probabilística de amostragem,[4] caracterizada pelo recrutamento de participantes no qual o critério é a disponibilidade mais fácil e pronta dos mesmos. A amostragem por conveniência guarda genericamente um risco considerável de viés, uma vez que exclui do recrutamento participantes que não são membros das comunidades específicas onde ele ocorre. Essa situação pode trazer prejuízos severos para a possibilidade de generalização dos resultados de uma pesquisa experimental.

[4] As técnicas de amostragem probabilística asseguram que todos os membros de uma população ou estrato populacional tenham chances iguais de ser selecionados para a composição da amostra.

Entendemos que a amostragem por conveniência é prevalente quando o pesquisador recruta seus participantes nas instituições de ensino superior onde atua. Na situação de pesquisa por nós evocada, o risco mais iminente de viés é o fato de que o recrutamento de estudantes e egressos de cursos de graduação da área de Letras e Linguística acarreta o risco de que as amostras observadas tenham maior treinamento, capacidade metalinguística e acuidade na avaliação de casos de agramaticalidade do que a população geral de pessoas com nível educacional semelhante.

Ainda do ponto de vista das decisões metodológicas que devem ser tomadas, há também um parâmetro da tarefa de julgamento de aceitabilidade que pode ter impacto sobre a fidedignidade das inferências sobre as observações apoiadas nesse paradigma experimental: a temporalização da tarefa. A temporalização dessa tarefa é manipulada em uma variante da mesma, na qual impõem-se aos participantes limitações no tempo de exposição aos estímulos e/ou na emissão dos julgamentos. Ainda é possível apenas demandar aos participantes que cheguem a um julgamento tão rápido quanto possível. Essa variante é denominada julgamento de aceitabilidade temporalizado (*speeded acceptability judgment* ou *timed acceptability judgment*, em inglês).

Normalmente subjaz ao julgamento de aceitabilidade temporalizado a concepção de que a percepção de aceitabilidade de uma sentença pressupõe o processamento da mesma. Assim, segundo Jiang (2012), diferenças no tempo de emissão de julgamentos podem, segundo os proponentes do paradigma experimental, refletir a interpolação de diferentes estratégias e mecanismos de processamento (por exemplo, a reanálise de uma sentença ambígua). Além disso, tal como sugerem Ellis (2005), Bowles (2011) e Gutiérrez (2013), o julgamento de aceitabilidade temporalizado pode ser considerado um instrumento psicométrico que captura a ativação de conhecimento implícito e rotinas de ativação de representações linguísticas automatizadas, contrariamente ao julgamento de aceitabilidade não temporalizado, que melhor reflete conhecimento explícito e reflexão metalinguística.

Assim, entendemos que a manipulação do teto temporal para a emissão de julgamentos acarreta um aumento da possibilidade de que essa percepção se aproxime do processamento linguístico. Especificamente, entendemos que através da imposição de um teto temporal na tarefa, busca-se restringir o escopo de informações disponíveis aos participantes para a emissão de seus julgamentos, na tentativa de maximamente

limitá-las àquelas estritamente associadas à representação mental do sistema linguístico e, portanto, ao estatuto de gramaticalidade teoricamente conferido à sentença julgada. Em outras palavras, busca-se, com a temporalização, evitar a possibilidade de que os julgamentos sejam motivados por impressões aleatórias, oriundas de reflexões idiossincráticas[5] e livres associações em torno do estímulo linguístico, o que poderia ocorrer quando períodos de tempo suficientemente amplos transcorrem antes da emissão dos julgamentos.

Considerando a clara relevância do parâmetro amostral e dos parâmetros das tarefas ora discutidos para a validade das interpretações sobre a competência linguística advindas de experimentos com julgamentos de aceitabilidade, neste estudo, realizamos uma investigação que buscou definir uma mínima janela temporal para a emissão de julgamentos de aceitabilidade temporalizados, bem como o impacto da seleção de amostras formadas por estudantes ou egressos de cursos superiores na área de Letras e Linguística.

Passamos, a seguir, à descrição dos métodos e, em seguida, aos resultados deste estudo.

2 Métodos

No presente estudo, tendo em vista nosso objetivo geral de exploração dos limites temporais mínimos para a resolução de um julgamento sobre o estatuto gramatical de sentenças, assim como a exploração do impacto da amostragem por conveniência dentre segmentos populacionais caracterizados por formação na área de Letras e Linguística, conduzimos dois experimentos. As tarefas empregadas em ambos os experimentos foram variantes da tarefa de julgamento de aceitabilidade temporalizado. Os dois experimentos foram conduzidos em contexto monolíngue, o que implicou a seleção de participantes falantes nativos das línguas dos estímulos.

O primeiro experimento (doravante "experimento um") teve por objetivo específico a estimativa da janela temporal média mínima para a formação de julgamentos. Caracterizou-se, portanto, como um estudo exploratório do parâmetro de tarefa almejado neste estudo. Para tal estudo

[5] Entendemos que postulados de quadros teóricos em linguística certamente poderiam estar no rol das idiossincrasias que poderiam enviesar a emissão de um julgamento.

exploratório, os participantes realizaram uma única tarefa de julgamento de aceitabilidade em computador, sendo a variável independente por nós controlada a tipologia das sentenças que compunham o *corpus* de estímulos experimentais. Nessa tarefa, a primeira variável dependente de interesse foi a latência temporal para reação a cada um dos estímulos, ou seja, o tempo de reação (TR) para a emissão de julgamentos para cada sentença. A segunda variável dependente por nós observada no experimento um foi a convergência dos julgamentos emitidos pelos participantes com a previsão de aceitabilidade ou não-aceitabilidade dos itens, feita com base na literatura linguística.

A tarefa do experimento um eliciou julgamentos binários (aceitação ou rejeição da sentença), através do acionamento de botões específicos do teclado do computador. A opção por julgamentos binários foi motivada pelo fato de nosso interesse majoritário ter sido a exploração do TR mínimo médio para a emissão de um julgamento. É nossa compreensão que uma tarefa que estimulasse a emissão de julgamentos graduais, ainda que potencialmente melhor adaptada à percepção de aceitabilidade, seria, por outro lado, uma tarefa cuja maior complexidade mascararia nossa intenção de obtenção de estimativas sobre o tempo mínimo de exposição a uma sentença simples para a formação de julgamentos sobre ela.

O segundo experimento (doravante "experimento dois") eliciou julgamentos graduais, ou seja, de grau ou nível de aceitabilidade, através de uma escala de tipo Likert de 5 pontos. Esse experimento teve por objetivo específico testar a hipótese de que o desempenho de participantes com formação na área de Letras e Linguística discrepa do desempenho de participantes sem tal treinamento específico em tarefas de julgamento de aceitabilidade, haja vista seu treinamento específico em modelos e métodos de análise linguística. Portanto, no experimento dois a hipótese de que diferenças de temporalização da tarefa de julgamento de aceitabilidade causam diferenças nos resultados de tais julgamentos foi igualmente testada. Desse modo, foi com o experimento dois que efetivamos o estudo do parâmetro de tarefa (temporalização) e do parâmetro amostral (perfil de treinamento profissional dos participantes selecionados) ora proposto, tendo sido o experimento um o seu preâmbulo.

A seguir, detalharemos o perfil dos participantes, os materiais e os procedimentos de cada um dos dois experimentos.

2.1 Participantes

2.1.1 Participantes do experimento um

Participaram do experimento um 16 falantes monolíngues do inglês americano, estudantes de graduação no Queens College da City University of New York. Todos eram residentes na cidade de Nova Iorque, nos EUA. A média de idade dos participantes era 19,1 anos de idade. Esse grupo de participantes foi recrutado entre alunos de uma disciplina de introdução à Psicologia, em nível de graduação, ofertada na instituição de ensino superior na qual eles eram estudantes. A disciplina compõe o currículo básico de trajetórias de formação em várias áreas de ciência básica e aplicada, assim em como artes e humanidades. Esses fatos asseguram que a amostra de participantes ora descrita não era composta unicamente por estudantes em formação específica em área correlata aos estudos em Letras/Linguística no Brasil.

2.1.2 Participantes do experimento dois

Participaram do experimento dois um total de 48 falantes nativos do português do Brasil, recrutados nas cidades de Belo Horizonte, Goiânia e Brasília. Desse total de participantes do experimento dois, 24 eram pessoas com formação na área de Letras e Linguística (estudantes de graduação, graduados, estudantes de pós-graduação e pós-graduados) e 24 pessoas com formação em áreas diferentes de Letras e Linguística. Os participantes com formação na área de Letras e Linguística eram estudantes ou egressos de cursos oferecidos na Universidade de Brasília, Universidade Federal de Goiás e Universidade Federal de Minas Gerais. A média de idade do grupo de participantes com formação na área de Letras e Linguística era de 25,6 anos de idade. A média de idade do grupo de participantes com formação em áreas diferentes de Letras e Linguística era de 30,7 anos de idade.

2.2 Materiais e procedimentos

2.2.1 Materiais e procedimentos do experimento um

Os materiais empregados constituíram um *corpus* experimental contendo 56 sentenças em inglês, das quais 16 apresentavam violações

gramaticais na língua inglesa. Tais violações agrupavam-se em duas categorias, cada uma contendo 8 sentenças: violações morfossintáticas e violações de subcategorização verbal. Na primeira categoria, agruparam-se instâncias de ausência de concordância e instâncias de violações de dependências de longa distância entre palavras WH e possíveis antecedentes (HAEGEMAN, 1991; HAEGEMAN; GUÉRON, 1999). Na segunda categoria, houve instâncias de violação na transitividade verbal, nas quais falsas sentenças transitivas foram configuradas a partir de verbos inergativos.

Para a produção de contrastes com essas categorias de sentenças agramaticais, realizamos comparações diretas com sentenças que incluíam verbos que podem ocorrer no inglês, tanto em construções transitivas quanto em construções intransitivas: verbos de modo de movimento[6] em construções causativas denominadas por Levin (1993) alternância de movimento induzido; e verbos de mudança de estado, que tipicamente não têm restrições para ocorrência em construções transitivas causativas, segundo Levin; Rappaport Hovav (1995). A opção por verbos de transitividade ambígua no inglês foi motivada por nossa expectativa de que tal ambiguidade poderia acarretar alguma dubiedade sobre a aceitabilidade da sentença em análise, evitando-se assim um contraste excessivamente óbvio.

As sentenças (1) a (4) são exemplos dos quatro tipos de sentenças críticas do experimento um. Cada exemplo é seguido de tradução para o português e é marcado de acordo com o estatuto de gramaticalidade esperado a partir da literatura linguística.

(1) The instructor ran the boys around the park.
"O instrutor fez os meninos correrem ao redor do parque."
(TIPO: Alternância de movimento induzido)

[6] Nos estímulos do experimento foram empregados verbos de modo de movimento cuja participação na alternância de movimento induzido foi proposta por Levin (1993) e Ritter e Rosen (2000). Trata-se de verbos que, na proposta dessas autoras e também na de Levin e Rappaport Hovav (1995), podem ocorrer em construções transitivas com leitura causativa, desde que denotem leitura télica, o que pode ser obtido através de adjuntos.

(2) The girls melted the cheese in the bowl.
"As meninas derreteram o queijo na vasilha."
(TIPO: Causativa com verbo de mudança de estado)

(3) *The farmer fell the apple from the tree.
"O fazendeiro caiu a maçã (sic) da árvore."
(TIPO: Falsa causativa com verbo inergativo)

(4) *What did Steven read the book that Helen talked about?
"O que Steven leu o livro do qual Helen falou?"
(TIPO: Violação de dependência de longa distância)

As sentenças eram apresentadas em ordem aleatória, para julgamentos binários, ou seja, elas eram julgadas apenas como bem ou mal formadas. A apresentação do conjunto de sentenças era contínua, sendo os julgamentos emitidos através de acionamento de dois botões específicos do teclado do computador destinado à tarefa, que foram modificados através da colagem de adesivos nas cores verde (botão situado mais à direita do teclado) e vermelho (botão situado à esquerda do teclado). O botão verde representava a aceitação da sentença como bem formada, ao passo que o botão vermelho representava a rejeição da sentença, ou seja, seu julgamento como mal formada. A sessão era iniciada com instruções através das quais os participantes eram orientados a julgarem a forma das sentenças, e não se elas poderiam ou não ter sentido em algum contexto, seguidas de fase de treinamento. Após as instruções e a fase de treinamento, os participantes executavam a tarefa em seu próprio ritmo, sem a presença do experimentador. O tempo médio para conclusão da sessão era de 4 minutos. A apresentação dos estímulos, o gerenciamento da randomização de itens e o registro dos tempos de reação para cada item foram feitos através do software DMDX, funcionando em um computador portátil (*laptop*) com sistema operacional produzido pela empresa Microsoft (Windows).

2.2.2 Materiais e procedimentos do experimento dois

Os materiais empregados constituíram um *corpus* experimental, contendo 60 sentenças em português, 21 das quais constituíram as

sentenças alvo do experimento dois, divididas em 3 categorias de violação da língua portuguesa. As demais sentenças eram gramaticais.

As 3 categorias contendo violações da gramática da língua portuguesa eram formadas por: (1) sentenças que forçavam para a língua portuguesa a alternância de movimento induzido, apresentando, portanto, verbos de modo de movimento em construção transitiva de provável leitura causativa; (2) sentenças falsas causativas com verbos intransitivos inergativos em construções com objeto direto; e (3) sentenças com violações de concordância e de movimento de ilhas. Portanto, as categorias de foco do experimento um encontravam-se reproduzidas, com a modificação do estatuto de gramaticalidade das sentenças com verbos de modo de movimento em construções transitivas, de gramatical (no inglês) para agramatical (no português).[7] Os exemplos (5) a (8) ilustram as 3 categorias de agramaticalidade, e uma sentença distratora (gramatical) do *corpus* experimental do experimento dois, sendo os tipos e o estatuto de gramaticalidade indicados.

(5) *Os pesquisadores correram os ratos pelo labirinto.
(TIPO: Alternância de movimento induzido)

(6) *O bebê sorriu a mulher na loja de eletrodomésticos.
(TIPO: Falsa causativa com verbo inergativo)

(7) *Quem o vendedor de carro conversou com o cliente?
(TIPO: Violação de dependência de longa distância)

(8) Vou enviar os dois trabalhos para ela imprimir.
(TIPO: Sentença distratora).

O experimento dois foi delineado como um estudo entre-sujeitos com dois fatores amostrais e dois fatores de constituição temporal da tarefa experimental. Portanto, para executá-lo selecionamos dois grupos de participantes, diferenciados por se tratarem de indivíduos com formação na

[7] A agramaticalidade de verbos de modo de movimento em construções transitivas no português do Brasil é analisada em Cambrussi (2009), e atestada por dados empíricos em Souza (2011).

área de Letras/Linguística (grupo 1) ou não (grupo 2). A tarefa de julgamento, por sua vez, foi dividida em duas variantes, divisão através da qual foi manipulado o parâmetro temporal e a exposição a treinamento. Assim, em uma das condições a tarefa era ora realizada com a exposição a cada sentença em tempo mínimo para emissão de julgamentos (estimado como sendo 4 segundos através do experimento um, como será discutido adiante) e, após sessão de treinamento, com 10 sentenças. Na outra condição, os julgamentos eram emitidos com o dobro de tempo de exposição a cada sentença e após a realização do mesmo tipo de tarefa com exposição de 4 segundos às sentenças, ou seja, após uma longa sessão de treinamento.

Antes do início das sessões de treinamento, os participantes passavam por uma sessão de instruções através de texto apresentado em tela de computador, com a presença do experimentador para a resolução de eventuais dúvidas. Na sessão de instruções, os participantes eram informados de que a tarefa em questão requeria julgamentos sobre a ordem e a seleção de palavras que compunham as sentenças. Eles eram instruídos a apenas considerar esses dois fatores, buscando ignorar, em seus julgamentos, se as sentenças poderiam ou não fazer sentido, ou ser ouvidas ou lidas em circunstâncias excepcionais.

As sentenças eram apresentadas em ordem aleatória e em apresentação contínua. Para a emissão de seus julgamentos, os participantes selecionavam as teclas numéricas de 1 a 5 de um teclado de computador. A codificação dos graus de agramaticalidade era enfatizada na sessão de instruções, na qual os participantes eram expostos, em dois momentos distantes, ao esquema de codificação reproduzido na Figura 1.

Tecla numérica	Julgamento
1	Totalmente inaceitável.
2	Bastante mal formada, quase inaceitável.
3	Mal formada, mas talvez aceitável.
4	Ligeiramente mal formada, quase perfeita.
5	Totalmente perfeita.

Figura 1 - Esquema de codificação para julgamentos de aceitabilidade do experimento dois

As sessões de instrução e de treinamento, a apresentação dos estímulos, o gerenciamento da randomização de itens e o registro dos julgamentos para cada item foram feitos por meio do software DMDX, funcionando em computadores portáteis (*laptops*) com sistema

operacional produzido pela empresa Microsoft (o Windows).

Passemos, agora, aos dados observados nos dois experimentos deste estudo.

3 Análises e discussão de resultados

Os dados obtidos através dos arquivos de registros gerados pelo DMDX foram tabulados em planilhas geradas pelo Excel, da empresa Microsoft. As análises foram realizadas através dos pacotes estatísticos SPSS versão 21, da empresa IBM, e do software livre R. Os gráficos foram gerados através do Graphpad Prism 6, da empresa Graphpad.

3.1 Experimento um

As médias dos tempos de reação (TRs) para a emissão de julgamentos dos 16 sujeitos, para cada um dos quatro tipos críticos de sentenças do experimento um, foram submetidos ao teste de Kolmogorov-Smirnov para verificação de sua adequação à distribuição normal. Os resultados do teste atestaram a normalidade das médias de TR. Essas médias, assim como os desvios-padrão observados, encontram-se expostas da Tabela 1.

Tabela 1 - Médias e desvio-padrão de tempo de reação para a emissão de julgamentos, por tipo de sentença.

Tipos de Sentenças	Médias de TR (em milissegundos)	DP
Movimento induzido	3231	709
Falsa transitiva	3294	474
Causativas com verbos de mudança de estado	3181	607
Violações sintáticas	3804	472

A análise de variância dos TRs observados nos quatro grupos de sentenças revelou um efeito principal do tipo de sentença, com significância no tratamento dos sujeitos como fator aleatório ($F1(3,45)=6,59$, $p<0,01$), e com significância marginal no tratamento

dos itens como fator aleatório (F2 (3,21)=2,92, p=0,058). Pós-testes pareados, ajustados com a correção de Bonferroni, demonstraram não haver diferença estatisticamente significativa entre os tempos médios observados nos julgamentos das sentenças com verbos de modo de movimento na alternância de movimento induzido e nas sentenças causativas com verbos de mudança de estado, ambos os tipos de sentenças previstos como gramaticais no inglês. Não houve tampouco diferença estatisticamente significativa entre os TRs médios observados nos julgamentos das sentenças gramaticais e nos julgamentos emitidos para as sentenças que eram falsas transitivas com verbos inergativos, essas últimas agramaticais. Entretanto, houve diferença estatisticamente significativa entre os TRs médios para a emissão de julgamentos para as sentenças com violações de concordância e dependências de longa distância e as sentenças com a alternância de movimento induzido ($p<0,05$), as falsas transitivas com verbos inergarivos ($p<0,05$) e as sentenças causativas com verbos de mudança de estado ($p<0,01$).

Observamos, portanto, que as sentenças apresentando violações morfossintáticas produziram latências para emissão de julgamentos maiores que as sentenças com violações estritamente relacionados a restrições de âmbito semântico-lexical, assim como maiores que as latências para a emissão de julgamentos de sentenças aceitáveis.

Tendo tal observação em vista, conduzimos análises com vistas à estimativa do TR médio mínimo para a resolução de julgamentos, a partir de uma nova amostra das sentenças usadas no experimento. Essa nova amostra de sentenças foi composta pelas oito sentenças com violações de natureza morfossintática e oito sentenças plenamente gramaticais que instanciavam diversos tipos de construção, retiradas aleatoriamente entre os dois tipos de sentenças gramaticais críticas da primeira análise e entre as sentenças distratoras que compunham o restante do conjunto dos estímulos usados no experimento um.

Procedemos com a comparação do tamanho médio das sentenças gramaticais e das sentenças agramaticais, com vistas a verificar se uma discrepância de TR entre os dois grupos poderia ser motivada por diferenças sistemáticas entre o tamanho dos estímulos verbais utilizados. A unidade de medida de tal tamanho foi o caractere alfabético, sem inclusão de espaços na contagem. As sentenças agramaticais tiveram tamanho médio de 36,62 caracteres (DP= 3,77); as sentenças gramaticais tiveram tamanho médio de 37 caracteres (DP= 6,80). A comparação das médias

não revelou diferenças estatisticamente significativas (t=0,13; GL=14, p>0,05), indicando, portanto, tratar-se de um único grupo amostral do ponto de vista do tamanho das sentenças em número de caracteres. Não obstante, o TR médio para emissão de julgamentos das sentenças gramaticais foi de 2.939 milissegundos (DP= 646 milissegundos), enquanto o TR médio de emissão de julgamentos das sentenças agramaticais foi de 3.804 milissegundos (DP= 472 milissegundos). Essa diferença foi significativa levando-se em consideração tanto os sujeitos como fator aleatório (t1=4,32; GL=30, p<0,001), quanto os itens como fator aleatório (t2=3,45; GL=14, p<0,01).

Observamos, portanto, a manutenção de um padrão confiável de diferenciação entre o TR médio para a emissão de julgamentos de sentenças com violações morfossintáticas e o TR médio para emissão de julgamentos de sentenças gramaticais. Tal como anteriormente verificado, esse padrão é caracterizado pela tendência da latência entre a visualização do estímulo e a emissão de um julgamento de aceitabilidade se a sentença com esse tipo de violação for maior do que aquela envolvida na decisão sobre a aceitabilidade de uma sentença. Esse não é um padrão surpreendente se considerarmos o TR como um indicador de custo de processamento. Essa consideração levaria logicamente à previsão de que, efetivamente, o julgamento de rejeição da sentença somente se daria após tentativas de processamento bem sucedido da mesma, através de reanálises.

A estimativa final da média de tempo para a emissão de julgamentos de aceitabilidade, considerando-se as sentenças gramaticais e agramaticais aglomeradas por sujeitos, foi de 3.372 milissegundos (DP=709 milissegundos). Ao considerarmos as sentenças gramaticais e agramaticais aglomeradas por itens, a média observada foi de 3,375 milissegundos (DP=653 milissegundos). Levando-se em conta a incorporação de um desvio-padrão para cima nesses dois cálculos de TRs médios, propomos que 4 segundos seria tempo suficiente, em média, para a emissão de julgamentos de aceitabilidade para sentenças redigidas com cerca de 40 caracteres alfabéticos, por um falante nativo adulto e com nível instrucional posterior à conclusão da educação básica, em sua língua materna.

3.2 Experimento dois

Na primeira análise do experimento dois, buscou-se examinar

se a percepção de aceitabilidade era um fator observável nos dois grupos de perfis de sujeitos, participantes com formação na área de Letras e Linguística e participantes com formação em áreas diferentes. Igualmente, buscou-se examinar se esse fator seria observável tanto na condição na qual a janela temporal para a resolução da tarefa era de 4 segundos, quanto na condição na qual essa janela temporal era ampliada para 8 segundos. Para esta análise, os quatro tipos de sentenças descritos anteriormente foram tomados como variáveis independentes. Assim, houve a possibilidade de investigar não somente se os estímulos construídos com violações da gramática da língua portuguesa eram percebidos como distintos dos estímulos sem essas violações, mas também se haveria diferenciação entre os dois tipos de violação relacionados a configurações semântico-lexicais e à violação primariamente sintática.

Como já foi mencionado, as tarefas do experimento dois requeriam que os participantes emitissem julgamentos de aceitabilidade através de uma escala Likert de 5 pontos, em que o ponto 1 representava a percepção de completa inaceitabilidade e o ponto 5 representava completa aceitabilidade. Há divergências na literatura estatística sobre a adequação do tratamento de dados obtidos através de escalas Likert, que podem ser interpretados como medidas ordinais, através de testes estatísticos paramétricos, tais como o teste t e a ANOVA, que se adequam a medidas de grandezas contínuas. Porém, autores como Dancey e Reidy (2006) e Norman (2010) defendem esse tipo de tratamento, argumentando que a robustez desses testes os torna eficientes mesmo para os dados das escalas Likert. Com base nesses autores, as médias dos pontos na escala atribuídos a cada um dos quatro tipos de sentença usados neste experimento, pelos sujeitos dos dois grupos, foram submetidos ao teste de Kolmogorov-Smirnov para verificação de sua adequação à distribuição normal. Os resultados do teste atestaram a normalidade dessas médias, que foram subsequentemente tratadas com testes estatísticos paramétricos.

A Tabela 2 apresenta as médias e desvios-padrão dos julgamentos dos dois perfis de participantes, referidos como LL (formação na área de Letras e Linguística) e ≠LL (formação em área diferente de Letras e Linguística). Na Tabela 2 essas médias são apresentadas para as duas condições de latência temporal máxima e para os três tipos críticos de sentenças: sentenças com verbos de modo de movimento em construção transitiva (movimento induzido), falsas sentenças transitivas com verbos intransitivos inergativos (falsa transitiva), sentenças com violações de

concordância e de dependências de longa distância (violação sintática) e sentenças gramaticais.

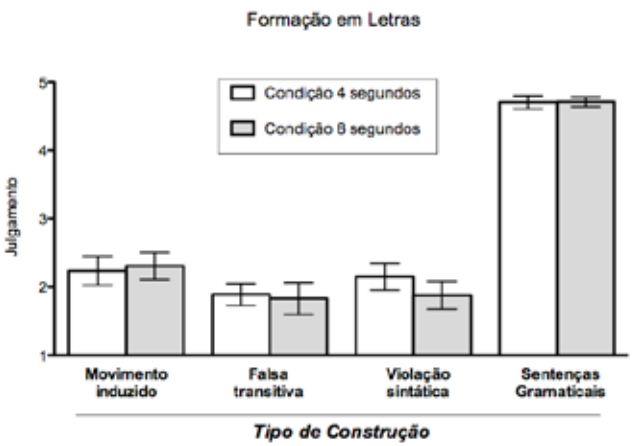
Tabela 2 - Médias e desvio-padrão de julgamentos de aceitabilidade por tipo de sentença, perfil dos participantes e condição de latência máxima para emissão de julgamentos.

TIPO	LL (4 segundos)	≠LL (4 segundos)	LL (8 segundos)	≠LL (8 segundos)
Movimento induzido	2,19 (0,48)	2,40 (0,91)	2,31 (0,60)	2,00 (0,58)
Falsa transitiva	1,88 (0,36)	2,46 (0,94)	1,83 (0,35)	1,82 (0,57)
Violação sintática	2,16 (0,66)	2,84 (0,93)	1,89 (0,49)	2,02 (0,74)
Sentenças gramaticais	4,68 (0,31)	4,28 (0,67)	4,71 (0,26)	4,45 (0,49)

O Gráfico 1 sintetiza a comparação entre as médias de julgamentos atribuídos aos quatro tipos de sentenças pelos participantes com formação na área de Letras e Linguística, nas condições de janela temporal de 4 segundos (Condição 1) e de janela temporal de 8 segundos (Condição 2) para a emissão do julgamento.

Gráfico 1 - Médias de julgamentos emitidos por participantes com formação na área de Letras e Linguística, por tipo de sentença e por condição.

Observamos um efeito principal do tipo de sentença na Condição 1 (4 segundos), ao considerarmos tanto os sujeitos ($F1(3,33)=123,59$, $p<0,001$) quanto os itens ($F2(3,18)=51,86$, $p<0,001$) como fatores



aleatórios. Houve um efeito principal da mesma natureza na Condição 2 (8 segundos), igualmente confirmado tanto na análise por sujeitos ($F1(3,33)=160,32$, $p<0,001$), quanto na análise por itens ($F2(3,18)=56,84$, $p<0,001$). Pós-testes pareados, ajustados pela correção de Bonferroni, revelaram que, na Condição 1, os julgamentos atribuídos às sentenças gramaticais foram significativamente diferentes dos julgamentos atribuídos às demais sentenças ($p<0,001$), tendo os julgamentos atribuídos às sentenças com verbos de modo de movimento em construção transitiva atingido diferença marginalmente significativa na comparação com as falsas sentenças transitivas com verbos intransitivos inergativos ($p=0,063$). Na condição 2, os mesmos pós-testes indicaram que os julgamentos atribuídos às sentenças com verbos de modo de movimento em construção transitiva diferenciaram-se significativamente tanto dos julgamentos das sentenças com falsas sentenças transitivas com verbos intransitivos inergativos, quanto daquelas com violações de concordância e de dependências de longa distância ($p=0,017$ e $p=0,007$, respectivamente). Não obstante tal diferenciação, é bastante claro que a média dos julgamentos atribuídos às sentenças com verbos de modo de movimento em construção transitiva revela que eles permaneceram no polo da percepção de agramaticalidade, uma vez que tal média (2,31) foi menos do que a metade, além de ter sido significativamente diferente ($p<0,001$) da média dos julgamentos atribuídos às sentenças gramaticais (4,71).

Passamos, agora, ao exame da questão do impacto do parâmetro temporal da tarefa de julgamento de aceitabilidade sobre as médias dos níveis de julgamento atribuídos pelos participantes com formação na área de Letras e Linguística na condição 1 (janela temporal de 4 segundos) e na condição 2 (janela temporal de 8 segundos). Procedemos tal exame através da condução de testes t de Student para a comparação das médias de cada um dos quatro tipos críticos de sentença, nas duas condições do parâmetro temporal. Os resultados encontram-se sintetizados na Tabela 3.

Tabela 3 - Comparação entre condições de latência máxima para emissão de julgmentos. Participantes com formação em Letras e Linguística

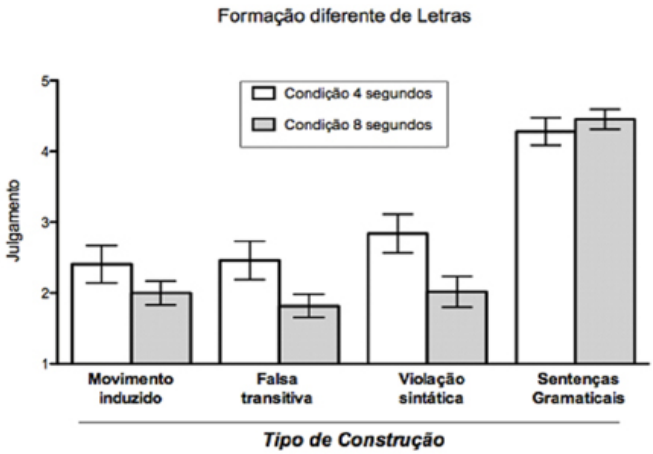
TIPO	Estatística	Valor-p
Movimento induzido	t = -0,29 (GL=11)	0,78
Falsa transitiva	t = 0,18 (GL=11)	0,86
Violação sintática	t = 1,1 (GL=11)	0,29
Sentenças gramaticais	t = -0,56 (GL=11)	0,96

Verifica-se, portanto, que entre os participantes com formação em Letras e Linguística não houve diferenças significativas nas médias de julgamentos observadas nas duas condições de temporalização da tarefa. Portanto, a partir das informações obtidas, através de análise de variância dos julgamentos emitidos em cada condição de latência da tarefa de julgamento e através da comparação das médias de julgamento para cada tipo crítico de sentença em cada uma das duas condições temporais, emerge um quadro sobre esse grupo de participantes. Esse quadro sugere que esses participantes são capazes de perceber diferenças de gramaticalidade e emitir julgamentos que refletem tal percepção dentro de uma janela temporal mínima, sendo que tais julgamentos mantêm-se razoavelmente estáveis com a ampliação do tempo alocado para sua emissão.

Passamos, então, à análise dos julgamentos emitidos pelo outro perfil amostral enfocado neste estudo. No Gráfico 2, apresentamos as médias totais dos julgamentos atribuídos por participantes com formação em áreas diferentes de Letras e Linguística. Mais uma vez, são apresentadas médias dos julgamentos eliciados pelos quatro tipos críticos de sentenças, nas condições de janela temporal de 4 segundos (Condição 1) e de janela temporal de 8 segundos (Condição 2).

Gráfico 2 - Médias de julgamentos emitidos por participantes com formação

em áreas diferentes de Letras e Linguística, por tipo de sentença e por condição



Um efeito principal do tipo de sentença foi observado na condição 1 (4 segundos), na análise tanto das médias por sujeitos ($F1(3,33)=18,18$, $p<0,001$) quanto das médias por itens ($F2(3,18)=31,64$, $p<0,001$). Mais uma vez, um efeito principal semelhante foi observado na condição 2 (8 segundos), na análise tanto por sujeitos ($F1(3,33)=90,65$, $p<0,001$), quanto por itens ($F2(3,18)=10,83$, $p<0,001$). Através de pós-testes pareados ajustados pela correção de Bonferroni, observamos que, na condição 1, os julgamentos atribuídos às sentenças gramaticais foram significativamente diferentes dos julgamentos atribuídos aos outros três tipos de sentenças contendo violações ($p<0,01$). Por outro lado, as sentenças com violações não produziram médias de julgamentos estatisticamente diferentes entre si. Esse padrão foi integralmente preservado através da aplicação dos mesmos pós-testes aos julgamentos dos participantes com formação em áreas diferentes de Letras e Linguística na condição 2. Ou seja, a média dos julgamentos atribuídos às sentenças gramaticais é significativamente diferente da média dos julgamentos atribuídos aos três grupos de sentenças com violações ($p<0,001$), ao passo que as médias dos julgamentos atribuídos a cada grupo de sentenças com violações não apresenta diferença estatística.

A Tabela 4 sintetiza a comparação dos dois parâmetros de configuração

temporal da tarefa de julgamento nos quais realizamos observações com participantes de áreas diferentes de Letras e Linguística, ou seja, latências máximas de 4 ou 8 segundos para a emissão dos julgamentos. Mais uma vez, a tabela apresenta a comparação, através do teste t, das médias para os quatro tipos críticos de sentenças.

Tabela 4 - Comparação entre condições de latência máxima para emissão de julgamentos. Participantes com formação em áreas diferentes de Letras e Linguística

TIPO	Estatística	Valor-p
Movimento induzido	t = 1,24 (GL=11)	0,24
Falsa transitiva	t = 2,22 (GL=11)	<0,05*
Violação sintática	t = 2,40 (GL=11)	<0,05*
Sentenças gramaticais	t = -0,72 (GL=11)	0,48

* Estatisticamente significativo

Assim, para os participantes com formação em áreas diferentes de Letras e Linguística, chegamos a um quadro de informações, obtidas através da análise de variância das médias de julgamentos em cada uma das duas condições de latência para emissão de julgamentos e da comparação entre médias por tipo crítico de sentença entre cada uma das duas condições. Verifica-se, nesse quadro, que esses participantes compartilharam com os participantes com formação na área de Letras e Linguística a capacidade de emitir julgamentos que refletem uma diferenciação clara do estatuto de aceitabilidade de sentenças em sua língua materna. Porém, os participantes com formação em áreas diferentes de Letras e Linguística parecem ter obtido alguma vantagem da latência maior (8 segundos) para a execução da tarefa. Tal como foi verificado na Tabela 2, suas médias de julgamentos das sentenças agramaticais nessa condição refletem uma rejeição levemente mais acentuada, diferença que atinge significância estatística para as sentenças falsas transitivas com verbos inergativos e sentenças com violações de concordância e de dependências de longa distância.

Tal padrão de diferenciação motivou um exame comparativo mais detalhado do desempenho dos dois perfis de participantes, já que eles constituíam o parâmetro amostral cuja comparabilidade era alvo de escrutínio do presente estudo. Assim, prosseguimos a análise com a comparação das médias de julgamentos obtidas em cada um dos tipos críticos de sentenças,

para cada um dos dois grupos de participantes, em cada uma das condições de latência para emissão dos julgamentos. Os resultados dessa comparação entre médias encontra-se sintetizado na Tabela 5.

Tabela 5 - Comparação entre médias de julgamento dos participantes com formação na áreas de Letras e Linguística e com formação em áreas diferentes de Letras e Linguística, por tipo de sentença e por condição de latência

TIPO/CONDIÇAO TEMPORAL	Estatística	Valor-p
Movimento induzido (4 segundos)	t = -0,52 (GL=11)	0,61
Falsa transitiva (4 segundos)	t = -1,58 (GL=11)	0,14
Violação sintática (4 segundos)	t = -1,89 (GL=11)	0,09
Sentenças gramaticais (4 segundos)	t = 2,20 (GL=11)	0,05*
Movimento induzido (8 segundos)	t = 1,25 (GL=11)	0,24
Falsa transitiva (8 segundos)	t = 0,05 (GL=11)	0,96
Violação sintática (8 segundos)	t = 0,47 (GL=11)	0,65
Sentenças gramaticais (8 segundos)	t = 1,69 (GL=11)	0,12

Na comparação das médias de julgamentos dos dois perfis de participantes que compõem o parâmetro amostral almejado neste estudo, para cada tipo de sentença examinado e em ambas as condições que compõem o parâmetro de tarefa, verificamos que as diferenças entre essas médias não atingiram genericamente significância estatística. A única exceção foram as médias atribuídas às sentenças gramaticais na condição de latência de 4 segundos, na qual a superioridade das médias dos julgamentos atribuídos pelos participantes com formação na área de Letras e Linguística (4,68 contra 4,28) foi estatisticamente confiável. É plenamente cabível observar que, não obstante a significância estatística dessa diferenciação, ambos os grupos julgaram essas sentenças em consonância com sua aceitabilidade, atingindo médias na metade superior da escala, na qual se encontrava representada a aceitabilidade.

Portanto, a comparação dos julgamentos entre os perfis amostrais, nas duas condições de execução da tarefa de julgamento, claramente apontam para o fato de que os dois grupos de participantes são capazes de distinguir o estatuto de aceitabilidade das sentenças precocemente. Resta-nos, entretanto, examinar se há diferenças na natureza da execução da tarefa na condição de latência reduzida (4 segundos), para os dois grupos, uma vez que foi nessa condição que as diferenças apontadas foram observadas.

Como foi mencionado anteriormente, foram observados efeitos

principais do tipo de sentença em ambas as condições de latência para emissão de julgamentos, para os dois perfis de participantes. Entretanto, a inspeção do desvio-padrão das médias de julgamentos observadas revela a ocorrência de maior variabilidade nos julgamentos dos participantes com formação em áreas diferentes de Letras e Linguística na condição de latência máxima de 4 segundos. A maior variabilidade é paralela a uma diferença verificada entre o tamanho do efeito observado, para esse grupo de participantes, nessa condição específica, e o tamanho do efeito observado nas demais combinações de perfil de participantes e condições de latência.

Os tamanhos do efeito foram estimados através do cálculo de eta parcial ao quadrado (η_p^2) O η_p^2 é um estimador do grau de associação entre variável independente (tipo de sentença, no caso do presente estudo) e variável dependente (julgamento, no caso do presente estudo) (RICHARDSON, 2011). Os tamanhos de efeito observados encontram-se sintetizados na Tabela 6.

Tabela 6 - Comparação do tamanho do efeito observado por grupo de participantes em cada uma das duas condições temporais

CONDIÇÃO TEMPORAL	LL	≠LL
4 segundos	$\eta_{p_2}^2 = 0,92$	$\eta_{p_2}^2 = 0,62$
8 segundos	$\eta_{p_2}^2 = 0,92$	$\eta_{p_2}^2 = 0,89$

Nota-se, através da inspeção da Tabela 6, que tamanhos de efeito robustos foram observados nas duas condições de temporalização da tarefa e com os dois grupos amostrais. Entretanto, nota-se igualmente que o efeito verificado nas duas condições de temporalização para a amostra composta por participantes com formação na área de Letras e Linguística manteve-se constante, ao passo que foi quase 50% maior na condição de 8 segundos do que na condição de 4 segundos para o grupo amostral formado por participantes com formação em áreas diferentes.

Além do tamanho do efeito, consideramos a frequência de ocorrência de julgamentos não emitidos por ultrapassagem do tempo limite, nas duas condições de temporalização. Tais ocorrências foram interpretadas por nós como um índice da dificuldade da tarefa em cada uma das condições temporais, uma vez que equivalem a erros na ação de emissão de uma resposta aos estímulos. A frequência total dessas

ocorrências foi negligenciável para os dois grupos amostrais na condição de 8 segundos (uma única ocorrência entre os participantes com formação na área de Letras e Linguística; e três ocorrências para os participantes com formação em outras áreas, em um total de 336 observações). Na condição de 4 segundos, a frequência de tais ocorrências foi maior entre os participantes com formação em outras áreas (43 ocorrências em 336 observações) do que entre os participantes com formação em Letras e Linguística (28 ocorrências em 336 observações). Entretanto, essa diferença de frequência não atingiu significância estatística (χ^2=3,17; GL=1, p>0,05).

É possível afirmar, portanto, que tanto os participantes com formação em Letras e Linguística quanto os participantes com formação em áreas diferentes demonstraram igualmente a capacidade de emitir julgamentos de aceitabilidade coerentes com expectativas advindas da teoria linguística em tarefas temporalizadas. Cabe ressaltar que essa capacidade foi demonstrada por ambos os grupos amostrais tanto em desempenho em janela temporal bastante restrita e com pouco treinamento prévio na tarefa, quanto com janela temporal um pouco maior e maior treinamento na tarefa. Na circunstância de menor janela temporal, foram observadas mais ocorrências de ausência de respostas por ultrapassagem do teto temporal entre os participantes com formação em áreas diferentes de Letras e Linguística, mas, ainda assim, a proporção dessas respostas falhas foi baixa e não difere, com significância estatística, de falhas cometidas pelos participantes com formação na área de Letras e Linguística, na mesma circunstância. Logo, não é possível afirmar confiavelmente que um dos grupos tem maior probabilidade do que o outro de enfrentar dificuldades na emissão de julgamentos na janela temporal maximamente restrita. A única diferença palpável entre os dois grupos de participantes por nós investigados é que, para aqueles com formação em áreas diferentes de Letras e Linguística, o tamanho do efeito produzido pela diferenciação entre sentenças gramaticais e agramaticais é menor na execução da tarefa de julgamento com menor teto de tempo, o que reflete uma menor sistematicidade nos julgamentos observados nessa condição.

Passemos, a seguir, aos comentários de conclusão do presente estudo.

4 Conclusão

Este estudo foi norteado por dois objetivos. Primeiramente, buscamos delimitar a janela temporal mínima para que sujeitos monolíngues fossem capazes de emitir julgamentos de aceitabilidade com acurácia em relação às previsões sobre o estatuto de gramaticalidade de sentenças advindas da teoria da gramática. Igualmente, buscamos comparar o desempenho de sujeitos monolíngues leigos (com formação superior em áreas diferentes dos estudos em Letras e Linguística) com sujeitos não leigos (com formação na área de Letras e Linguística) em tarefas de julgamento de aceitabilidade temporalizadas, com vistas a avaliar o impacto do recrutamento de amostras de conveniência por pesquisadores da área de Linguística.

Ambos os objetivos foram plenamente alcançados. Estimamos o tempo de 4 segundos como uma janela temporal dentro da qual sujeitos monolíngues, com nível superior de instrução, são capazes de emitir julgamentos de aceitabilidade. Observamos, ainda, que após o recebimento de instruções que orientaram a atenção para a ordem e a escolha de palavras e uma sessão de prática com 8 itens de treinamento, tanto os participantes não leigos quanto os leigos de nosso estudo foram capazes de executar a tarefa de julgamento de aceitabilidade.

Nossos resultados, além disso, foram genericamente convergentes com as hipóteses da teoria da gramática sobre o estatuto de gramaticalidade dos tipos de construções empregadas nos dois experimentos. Assim, especificamente em relação às construções por nós analisadas, não foram replicadas as observações de Myers (2009a) e Gibson e Fedorenko (2013), que indicaram o falseamento de hipóteses sobre a gramaticalidade da teoria da gramática.

Nossas observações indicaram claramente efeitos do estatuto de gramaticalidade dos tipos de sentenças por nós escolhidos, efeitos esses constatados tanto com o parâmetro temporal mínimo quanto com seu dobro e nos dois perfis de participantes que compunham nosso parâmetro de amostragem. Desse modo, a partir de nossos resultados com julgamentos de aceitabilidade temporalizados, podemos afirmar que a percepção como inaceitáveis de sentenças, cuja agramaticalidade em uma dada língua é real, é um efeito robusto, de rápida detecção e não estritamente condicionado pela formação dos juízes nos estudos sobre a linguagem.

Por outro lado, nossos resultados convergem com a proposta de Culbertson e Gross (2009), no que diz respeito ao ponto de vista defendido por esses autores de que não há diferenças relevantes entre participantes com conhecimento de linguística ou não, mas sim diferenças relacionadas à familiaridade com a tarefa de julgamento de aceitabilidade. A única diferença por nós observada entre os sujeitos leigos e não leigos de nosso estudo foi o tamanho do efeito da agramaticalidade, menor na observação dos julgamentos emitidos pelos sujeitos leigos dentro da janela temporal mínima estabelecida (4 segundos). Não obstante, tal diferenciação de tamanho do efeito desapareceu quando os sujeitos leigos executaram a tarefa com 8 segundos para a emissão de seus julgamentos e após mais treinamento na mesma.

Nossos resultados são díspares do mencionado relato de Barile e Maia (2008) e Maia (2013) sobre a comparação entre julgamentos de sujeitos recém-egressos de um curso de teoria sintática e julgamentos de sujeitos ingênuos. Entretanto, cabe a ressalva de que a questão abordada no relato dos autores era o efeito de saturação produzido pela exposição ostensiva a um tipo de construção agramatical, o que pode levar sujeitos que recebem esse tipo de exposição a tornarem-se menos sensíveis à agramaticalidade em questão. Até onde vai nosso conhecimento, nenhum dos participantes dos nossos dois grupos havia recebido qualquer espécie de treinamento específico nos tipos de sentenças que compuseram os estímulos de nossos experimentos. Portanto, não é plausível supor-se que eles demonstrariam o efeito de saturação à exposição ostensiva em questão.

Diante de tais resultados, julgamo-nos capazes de opinar que estudos experimentais baseados em julgamento de aceitabilidade temporalizados não são afetados de modo significativo quando são conduzidos com amostragem de conveniência, em que ocorre recrutamento de participantes com formação na área de Letras e Linguística. É seguro afirmar que, no caso de hipótese sobre o estatuto de gramaticalidade serem confirmáveis empiricamente, efeitos de gramaticalidade serão experimentalmente produzidos tanto nesse segmento populacional específico, quanto na população em geral de mesma faixa etária e nível instrucional.

A associação mais forte entre a gramaticalidade e os julgamentos observados dos participantes da área de Letras e Linguística, mensurada pelo tamanho do efeito maior, sugere que esses participantes têm alguma vantagem apenas na emissão de julgamentos em janela temporal

restrita. Entendemos que essa vantagem é irrelevante para estudos cujos objetivos contemplam tão somente a confirmação da gramaticalidade de construções, através da observação experimental de sua aceitabilidade. Entretanto, certamente é possível conceber situações nas quais a aparente maior destreza de um tipo de população possa porventura ser uma variável a ser controlada, como, por exemplo, em estudos planejados com vistas à inspeção de mecanismos psicolinguísticos subjacentes (ou anteriores) à emissão de julgamentos. Nessa situação, em casos onde houvesse recrutamento de participantes sem formação específica em estudos de linguagem, recomendaríamos uma ampliação do teto temporal mínimo por nós adotado em julgamentos temporalizados, acoplada a sessão de treinamento mais prolongada antes do início das observações críticas.

Referências

BARD, E. G.; ROBERTSON, D.; SORACE, A. Magnitude Estimation of Linguistic Acceptability. *Language*, v. 72, n. 1, p. 32-68, 1996.

BARILE, W.; MAIA, M. Aspectos Prosódicos do Qu in situ no Português Brasileiro. Revista Virtual de Estudos da Linguagem, v. 6, no. 11, p.1-21, 2008.

BOWLES, M. Measuring Implicit and Explicit Linguistic Knowledge – What can Heritage Language Learners Contribute? *Studies in Second Language Acquisition,* 33, p. 247-271, 2011.

CAMBRUSSI, M. F. *A alternância causativa de verbos inergativos no português brasileiro*. Tese (Doutorado em Linguística). Florianópolis: Universidade Federal de Santa Catarina, 2009.

CLIFTON, C; FANSELOW, G; FRAZIER, L. Amnestying Superiority Violations: Processing Multiple Questions. *Linguistic Inquiry,* v. 37, p. 51-68, 2006.

COOK, V. Timed Grammaticality Judgments of the Head Parameter in L2 Learning. In: Bartelt, G. (Org.). *The Dynamics of Language Processes* – Essays in Honour of Hans W. Dechert. Tübingen: Gunter Narr, 1994. p. 155-180.

COSBY, P. *Métodos de Pesquisa em Ciências do Comportamento.* São Paulo: Atlas, 2009.

COWART, W. *Experimental Syntax:* Applying Objective Methods to

Sentence Judgments. Thousand Oaks: Sage Publications, 1997.

CULBERTSON, J.; GROSS, S. Are Linguists Better Subjects? *British Journal for the Philosophy of Science,* v. 60, p. 721-736, 2009.

DANCEY, C.; REIDY, J. *Estatística sem Matemática para a Psicologia* – Usando SPSS para Windows. 3. ed. Porto Alegre: Artmed, 2006.

DEVITT, M. Intuitions in Linguistics. *British Journal for the Philosophy of Science*, v. 57, p. 481-513, 2006.

ELLIS, R. Measuring Explicit and Implicit Knowledge of a Second Language – A Psychometric Study. *Studies in Second Language Acquisition,* v. 27, p. 141-172, 2005.

FERREIRA, F. Psycholinguistics, Formal Grammars, and Cognitive Science. *The Linguistic Review*, v. 22, p. 365-380, 2005.

GIBSON, E.; FEDORENKO, E. The Need for Quantitative Methods in Syntax and Semantics Research. *Language and Cognitive Processes,* v. 28, n. 1/2, p. 88-124, 2013.

GUTIÉRREZ, X. The Construct Validity of Grammaticality Judgment Tests as Measures of Implicit and Explicit Knowledge. *Studies in Second Language Acquisition,* v. 35, p. 423-449, 2013.

HAEGEMAN, L. *Introduction to Government and Binding Theory.* 2. ed. Oxford, UK; Malden, MA: Blackwell Publishers, 1994.

HAEGEMAN, L.; GUÉRON, J. *English Grammar* – A Generative Perspective. Oxford, UK; Malden, MA: Blackwell Publishers, 1999.

HARRIS, R. A. *The Linguistics Wars*. Oxford; New York: Oxford University Press, 1995.

JIANG, N. *Conducting Reaction Time Research in Second Language Studies*. New York: Routledge, 2012.

KELLER, F. Grammaticality Judgments and Linguistic Methodology. Centre for Cognitive Science – University of Edinburgh. *Research Paper.* EUCCS-RP, 1998.

LEVIN, B. *English Verb Classes and Alternations*. A Preliminary Investigation. Cambridge, MA: The MIT Press, 1993.

LEVIN, B.; RAPPAPORT HOVAV, M. *Unaccusativity* – At the Syntax-Lexical Semantics Interface. Cambridge, MA: The MIT Press, 1995.

MAIA, M. Sintaxe experimental: uma entrevista com Marcus Maia. *Revista Virtual de Estudos da Linguagem*, Belo Horizonte, v. 10, n. 18, p. 184-193, 2012.

MAIA, M. Linguística experimental: aferindo o curso temporal e a profundidade do processamento. *Revista de Estudos da Linguagem*, Belo Horizonte, v. 21, n. 1, p. 9-42, 2013.

MYERS, J. Syntactic Judgment Experiments. *Language and Linguistics Compass,* v. 3, n. 1, p. 406-423, 2009a.

MYERS, J. The Design and Analysis of Small-scale Syntactic Judgment Experiments. *Lingua*, n. 119, p. 425-444, 2009b.

NORMAN, G. Likert Scales, Levels of Measurement, and the "Laws" of Statistics. *Advances in Health Science Education*, v. 15, n. 5, p. 625-632, 2010.

PHILIPS, C. Should we Impeach Armchair Linguists? In: IWASAKI, H.; CLANCY S.; SOHN, O. (Eds.). *Japanese-Korean Linguistics,* Stanford, CA: CSLI Publications, 2009. v. 17. p. 116-145.

PHILIPS, C.; WAGERS, M. Relating Structure and Time in Linguistics and Psycholinguistics. *Handbook of psycholinguistics,* ed. by Gareth Gaskell. Oxford, UK: Oxford University Press, 2007. p. 739-756.

RICHARDSON, J. T. E. Eta Squared and Partial Eta Squared as Measures of Effect Size in Educational Research. *Educational Research Review*, v. 6, n. 2, p 135-147, 2011.

RITTER, E.; ROSEN, S. Event Structure and Ergativity. In: TENNY, C.; PUSTEJOVSKY, J. (Orgs.). *Events as Grammatical Objects* – The Converging Perspectives of Lexical Semantics and Syntax. Stanford, CA: CSLI Publications, 2000.

SCHÜTZE, C. T. *The Empirical Base of Linguistics* – Grammaticality Judgments and Linguistic Methodology. Chicago: University of Chicago Press, 1996.

SOUZA, R. A. L2 Argument Structure in L2 Acquisition: Language Transfer Revisited in a Semantics and Syntax Perspective. *Ilha do Desterro*, n. 60, p. 153-187, 2011.

WASOW, T.; ARNOLD, J. Intuitions in Linguistic Argumentation. *Lingua*, n. 115, p. 1481-1496, 2005.