

O uso da tarefa de *cloze* no delineamento experimental de estudos sobre o processamento linguístico antecipatório

The use of *cloze* task in the experimental design of studies on anticipatory linguistic processing

Cláudia Brandão Vieira¹

Resumo: Este trabalho enfatiza a necessidade de realização da tarefa de *cloze* (TAYLOR, 1953) em experimentos que investigam o processamento linguístico antecipatório. Estudos sobre a ocorrência de antecipação no processamento linguístico propõem que, durante a compreensão, o falante procura antecipar, com o máximo de certeza, os desdobramentos das sentenças ao longo do tempo. Na literatura psicolinguística, a tarefa de *cloze* é uma das formas utilizadas para se mensurar a previsibilidade de uma palavra em um determinado contexto. A tarefa apresentada neste artigo verificou como se caracteriza o comportamento de subseleção de argumentos internos por 520 verbos. O conhecimento desse comportamento pode ser utilizado em experimentos que investigam o processamento antecipatório, uma vez que argumentos frequentemente subselecionados por certos verbos – cujas probabilidades de *cloze* são muito altas – podem direcionar a antecipação de traços semânticos de determinados itens lexicais. Os resultados mostraram que os verbos em português brasileiro podem apresentar diferentes comportamentos de subseleção de argumentos, o que pode refletir na antecipação de traços semânticos de argumentos internos muito esperados.

Palavras-chave: *cloze*; antecipação; verbos; informações temáticas.

Abstract: This study highlights the relevance of the *cloze* task (TAYLOR, 1953) for experiments that investigate anticipatory linguistic processing. Studies on anticipation in linguistic processing propose that during comprehension the speaker seeks to anticipate, with maximum certainty, the unfolding of sentences over time. In the psycholinguistic literature, the *cloze* task is one of the means to measure the predictability of a word in a given context. The task presented in this article assessed the behavior of subselection of internal arguments by 520 verbs. The knowledge of this behavior can be used in experiments that investigate anticipatory processing, as arguments frequently subselected by certain verbs – those with very high *cloze* probabilities – can direct the anticipation of semantic features of certain lexical items. The results showed that verbs can feature different behavior of subselection of arguments, which can show in the anticipation of semantic traits of highly expected internal arguments.

Keywords: *cloze*; anticipation; verbs; thematic information.

¹ Universidade de Brasília, Departamento de Linguística, Português e Línguas Clássicas, Brasília, DF, Brasil. Endereço eletrônico: claubrandaov@gmail.com.

Introdução

O conceito de antecipação está relacionado à capacidade de o cérebro lidar com a velocidade e a complexidade do processamento linguístico por *pensar à frente*, gerando informações sobre prováveis estímulos e preparando-se antecipadamente para o processamento desses estímulos em múltiplos níveis (FEDERMEIER, 2007). Em outras palavras, a antecipação durante a compreensão da linguagem pode ser definida como o uso de representações internas de alto nível – conhecimentos do falante sobre o mundo real em interação com informações provenientes do contexto linguístico, discursivo/situacional – na pré-ativação de representações linguísticas de níveis que se encontram mais abaixo durante o desenvolvimento de um enunciado.

Conforme Kuperberg e Jaeger (2016), é possível que, durante a compreensão da linguagem, o falante procure antecipar, com o máximo de certeza, os desdobramentos das sentenças ao longo do tempo. Nessa perspectiva, a antecipação de unidades lexicais é considerada um fenômeno probabilístico, pois há diferentes expectativas paralelas de ocorrência de palavras a serem consideradas durante o processamento linguístico. Ou seja, diferentes possibilidades de desdobramentos das sentenças são consideradas ao mesmo tempo, e cada uma possui diferentes graus de probabilidade.

Em estudos psicolinguísticos que visam analisar o caráter probabilístico da antecipação da linguagem, é necessária a utilização de métodos cujos dados empíricos presentes nos experimentos sejam baseados em ocorrências que efetivamente estejam presentes no mundo real, isto é, não sejam apenas artefatos criados pelo pesquisador. Logo, para que os experimentos psicolinguísticos sejam realizados adequadamente, é necessário que os dados utilizados nos experimentos possuam a chamada *validade ecológica*.

A utilização do método de *cloze* é uma das formas de se garantir a validade ecológica dos dados em estudos que consideram o caráter antecipatório do processamento da linguagem. As probabilidades verificadas nessa tarefa são o método mais utilizado para quantificar a possibilidade de ocorrência de certas palavras em contextos específicos.

O presente artigo enfatiza o valor da realização da tarefa de *cloze* em estudos que visam verificar o caráter antecipatório do processamento linguístico. Especificamente, evidenciaremos como informações sobre o comportamento de subseleção dos verbos, em português brasileiro, podem ser obtidas por meio da tarefa de *cloze*. Pretende-se expor que a prática de procedimentos rigorosos é bastante importante para a criação de itens experimentais. Para tanto, será apresentada, brevemente, a perspectiva antecipatória do processamento da linguagem. Em seguida, serão elucidadas as características da tarefa de *cloze*. Por fim, será

exposto um exemplo de aplicação da metodologia de *cloze* para a criação de um *corpus* cujo objetivo é a investigação das propriedades de subseleção dos verbos.

A antecipação no processamento da linguagem

A discussão a respeito da atuação de mecanismos de antecipação na cognição humana não se restringe ao processamento linguístico. Em sentido amplo, o processamento antecipatório refere-se a qualquer tipo de processo que incorpore ou gere não apenas informações sobre o passado ou o presente, mas também sobre estados futuros do corpo ou do ambiente. Historicamente, a investigação sobre mecanismos de antecipação iniciou-se quase paralelamente aos estudos sobre processamento perceptual e motor. O interesse pela inter-relação entre percepção e ação nasceu da busca por explicações para uma questão fundamental em nossa cognição: como o nosso mundo visual permanece estável, apesar do constante deslocamento de imagem introduzido pelos movimentos dos olhos e da cabeça? Uma das primeiras propostas a relacionar expectativas e ações foi formulada no século XIX dentro do princípio ideomotor, que recentemente foi atualizado por teorias que sugerem a existência de códigos compartilhados entre percepção e ação (BUBIC; VON CRAMON; SCHUBOTZ, 2010; CLARK, 2013).

Estudos em Psicologia Cognitiva e em Neurociência indicam que a capacidade de antecipar informações pode influenciar diversas funções cognitivas, como o processamento atencional e o processamento de informações para tomada de decisões. Tal direcionamento em relação ao futuro tem sido reconhecido como relevante e benéfico para diferentes aspectos do processamento de informações, tais como percepção, aprendizado implícito, teoria da mente e outros processos cognitivos em humanos e, de uma forma mais rudimentar, em animais².

Sob a perspectiva linguística, o conceito de antecipação está relacionado à capacidade de o cérebro lidar com a velocidade e a complexidade do processamento linguístico por *pensar à frente*, gerando informações sobre prováveis próximos estímulos e preparando-se antecipadamente para o processamento desses estímulos, em múltiplos níveis (FEDERMEIER, 2007). Nesta pesquisa, assim como Kuperberg e Jaeger (2016), consideramos que a capacidade de previsão do falante está apoiada em uma representação interna do contexto. Essa representação interna é composta por informações linguísticas e não linguísticas – que podem incluir desde representações parciais inferidas de entradas contextuais previamente processadas, até representações de nível superior, como representações de eventos. Sob essa

² Cf. BUTZ, 2004; CLARK, 2013 para revisões sobre o assunto.

perspectiva, a representação interna do contexto do falante altera o estado do sistema de processamento linguístico antes do surgimento de novos estímulos que estejam em níveis inferiores (*e.g.*, representações mentais de níveis ortográfico, fonético, fonológico).

Diferentes estudos experimentais sugerem que previsões de informações de nível semântico ocorrem durante processamento linguístico. Um exemplo é o estudo seminal de DeLong, Urbach e Kutas (2005), que encontrou evidências de processos de antecipação durante a compreensão da leitura. No experimento, os pesquisadores usaram manipulações que exploravam o conhecimento de mundo de leitores e a variação entre os usos dos artigos indefinidos do inglês *a* e *an*. Embora possuam significados idênticos, a ocorrência de cada um desses artigos é direcionada pelo fonema inicial da palavra seguinte. Os pesquisadores criaram frases que poderiam ser continuadas por palavras esperadas ou palavras inesperadas, porém plausíveis. O artigo adequado para a palavra esperada era sempre diferente daquele adequado para a palavra inesperada, conforme exemplificado a seguir:

(1) The day was breezy so the boy went outside to fly a kite in the park.

* Ventava naquele dia, então o menino saiu para empinar uma pipa no parque.

Continuação inesperada

(2) The day was breezy so the boy went outside to fly an airplane in the park.

* Ventava naquele dia, então o menino saiu para empinar um avião no parque.

Os resultados mostraram que o processamento *on-line* dos artigos *a* e *an* era influenciado pela expectativa por certas palavras³. Artigos adequados para palavras esperadas (*e.g.* *a* para *kite*) apresentavam custos processuais significativamente menores do que os custos produzidos quando os artigos eram inadequados para palavras esperadas, mas adequados para palavras inesperadas (*e.g.* *an* é inadequado para a palavra esperada *kite*, mas adequado para a palavra inesperada *airplane*). Uma questão importante é o fato de que os participantes previram os sintagmas nominais esperados, antes mesmo de visualizá-los. Os resultados mostram que os participantes utilizaram a representação contextual e conhecimento sobre eventos do mundo real para prever que, em um dia de muito vento, é mais provável que um menino empine uma pipa (*kite*) do que um avião (*airplane*). Como as duas formas do artigo possuem significados

³ Os artigos indefinidos *a* e *an* do inglês possuem um comportamento distinto dos artigos indefinidos do português, pois não realizam concordância de gênero gramatical. No caso dos artigos em inglês, se trata de um efeito puramente fonológico, motivado pela cliticização do artigo, levando um sintagma nominal como *an airplane* a ser realizado como uma única palavra fonológica. Não há informação morfológica em ação, contrariamente ao que ocorre em português.

idênticos, esses resultados não podem ser relacionados à facilidade/dificuldade de integrar informações.

O estudo de DeLong *et al.* (2005) é bastante relevante para as propostas que relacionam a antecipação ao processamento da linguagem, pois mostra que os participantes foram capazes de prever a forma ortográfica dos sintagmas nominais mais esperados, antes mesmo de visualizá-los. Os resultados sugerem que os participantes utilizaram a representação contextual e conhecimento mais refinados sobre eventos do mundo real para preverem ativamente que, na sentença em que se descreve um dia de muito vento, seria mais provável que o verbo *fly* fosse seguido por *kite* (pipa) do que por *airplane* (avião). Como as duas formas do artigo possuem significados idênticos, esses resultados não podem ser relacionados à facilidade/dificuldade de integrar informações conceituais, mas indicam que os participantes anteciparam a forma ortográfica da palavra a ser apresentada.

Outros estudos experimentais foram propostos com o objetivo de averiguar a existência e a natureza de processos antecipatórios. Pesquisas em espanhol, holandês e polonês (WICHA *et al.*, 2004; VAN BERKUM *et al.*, 2005) também investigaram tais questões, por meio da apresentação de artigos ou adjetivos que concordassem ou não com as expectativas dos leitores/ouvintes (*e.g.*, “Caperucita Roja cargaba la comida para su abuela en una canasta (complemento esperado)/un sesto (complemento inesperado)”). De modo geral, foi encontrado aumento de custos de processamento relacionados à leitura/audição de artigos ou adjetivos quando há violação de expectativas, indicando que a antecipação é parte dos mecanismos envolvidos no processamento linguístico.

Apesar do grande número de evidências experimentais que se acumularam ao longo das últimas décadas⁴, muitas questões a respeito dos processos preditivos durante o processamento linguístico ainda se encontram em aberto. Um número relevante de investigações procura responder a questões específicas sobre o processamento antecipatório – por exemplo, quais níveis representacionais são pré-ativados durante o processamento linguístico. Huettig (2015) afirma que, embora haja progressos nas pesquisas acerca do tipo de informações utilizadas, ainda existem muitas questões a serem investigadas a respeito dos processos de previsão. Pouco se conhece sobre os níveis de representações ativadas e sobre o quanto essas ativações são dependentes do contexto. Huettig e Mani (2016) questionam a obrigatoriedade da antecipação para a compreensão da linguagem e criticam a metodologia empregada em diversos estudos. Os pesquisadores propõem que os mecanismos de antecipação podem auxiliar na compreensão

⁴ Para revisões sobre o assunto, confira Van Pette e Luka (2012) e Kuperberg e Jaeger (2016).

em situações específicas, mas o processamento linguístico não necessita frequentemente da previsão para ser realizado com sucesso.

Tal fato revela a necessidade de delineamentos experimentais cuidadosos em investigações específicas sobre os efeitos da antecipação no processamento linguístico. Na próxima seção, abordaremos as características da tarefa de *cloze* e as aplicações em estudos psicolinguísticos.

O teste de *cloze*

Na literatura psicolinguística, a tarefa de *cloze* (TAYLOR, 1953) é uma das formas utilizadas para se mensurar a previsibilidade de uma palavra em determinado contexto. Embora tenha sido criada para avaliar e comparar a legibilidade de pequenas passagens de texto, a tarefa é frequentemente utilizada para quantificar previsibilidade de palavras após contextos sentenciais e discursivos. Segundo Taylor (1953), o termo *cloze* se relaciona ao conceito de *closure*, proveniente da psicologia Gestalt. Esse conceito se refere à tendência dos seres humanos em completar padrões familiares, mas não finalizados.

Na tarefa de *cloze*, participantes devem criar continuações para fragmentos de sentenças isoladas, ou diversos fragmentos de textos mais extensos. A probabilidade de uma palavra ser produzida nessas tarefas é chamada de probabilidade de *cloze*. Em geral, são investigadas as probabilidades de ocorrência de palavras após contextos bastante informativos e contextos neutros. Quando a maior parte dos participantes da tarefa utiliza uma única palavra para completar um mesmo fragmento, infere-se que, naquele contexto, tal palavra é bastante provável, possui uma alta probabilidade de *cloze*.

Para Federmeier *et al.* (2007), os efeitos de facilitação de palavras em contextos moderada e altamente previsíveis são mediados pelo grau de ajuste entre uma palavra específica e determinado contexto. Tal ajuste é frequentemente definido empiricamente usando um procedimento de *cloze*. Kuperberg (2016) sugere que a interpretação dos dados de *cloze* podem refletir a força e as probabilidades de expectativas paralelas que o cérebro de um indivíduo calcula durante determinada tarefa.

Estudos eletrofisiológicos (KUTAS; HILLYARD, 1984; KUTAS; VAN PETTEN, 1994; VAN PETTEN; KUTAS, 1990) indicam que a probabilidade de *cloze* de uma palavra está relacionada à amplitude do N400, que consiste em uma onda cerebral de polaridade negativa cujo pico acontece a partir de 400ms da apresentação de um estímulo. Na década de 1980, o N400 tornou-se conhecido na literatura psicolinguística como um índice de utilização de informações de significado. Em geral, todas as palavras com conteúdo semântico produzem

o N400. A amplitude desse componente é sensível a fatores relacionados à facilidade de acesso a informações na memória – como frequência da palavra, repetição e vizinhança ortográfica (KUTAS; VAN PETTEN; KLUENDER, 2006).

Há evidências (KUTAS; HILLYARD, 1984) de que a amplitude do componente também é influenciada – de modo inversamente proporcional – pela expectativa de ocorrência de uma palavra em determinado contexto discursivo. Kutas e Hillyard (1980) mostraram que, durante a leitura de frases como “Eu prefiro café com creme e ...”, palavras conceitualmente esperadas, como a palavra *açúcar*, provocavam a diminuição da amplitude do N400 se comparadas a palavras conceitualmente inesperadas, como a palavra *meias*.

Palavras muito previsíveis, que possuem alta probabilidade de ocorrência nas tarefas de cloze, produzem ondas N400 com amplitudes reduzidas. Contudo, palavras com baixa probabilidade de *cloze* também são capazes de gerar redução do N400 caso compartilhem traços semânticos com palavras de alta probabilidade (FEDERMEIER; KUTAS, 1999).

Existem discussões sobre o tipo de conhecimento aferido em uma tarefa de *cloze*. Usualmente, as probabilidades de *cloze* são consideradas estimativas de previsibilidade de palavras, considerando-se o uso real da língua. Contudo, a conexão entre as informações fornecidas na tarefa e o conhecimento linguístico implícito dos falantes ainda é pouco explorada.

Para Staub *et al.* (2015), duas perspectivas teóricas são usadas para explicar como as informações expressas durante a tarefa de *cloze* estão associadas aos conhecimentos probabilísticos (linguísticos e não linguísticos) dos falantes. A primeira perspectiva propõe que os resultados da tarefa expressam a previsibilidade de uma palavra para uma comunidade de falantes, mas não para os falantes individualmente. Por exemplo, em uma tarefa de *cloze* com o fragmento (3), foram aferidas as seguintes probabilidades: 0.90 para a continuação *bolo* e 0.10 para a continuação *doces*.

(3) Carla confeitou _____.

Sob tal perspectiva teórica, existem dois grupos de participantes: um grupo composto por 90% dos participantes, para os quais a palavra *bolo* é a melhor continuação; um segundo grupo composto por 10% dos participantes, para os quais a palavra *doces* é a melhor continuação. Tal diferença ocorre porque as experiências linguísticas e os conhecimentos de mundo de cada indivíduo dos grupos são diferentes. Assim, cada indivíduo desses grupos considera apenas uma palavra como a melhor continuação para o trecho.

A segunda perspectiva teórica assumida neste artigo propõe que, embora cada participante produza uma única resposta durante a tarefa, outras respostas também podem fazer parte do conhecimento probabilístico do falante (ROLAND *et al.*, 2012). Em outras palavras, se considerarmos o exemplo (3), a divisão das probabilidades em 90% e 10% não indica que cada grupo de participantes preveja apenas uma palavra como melhor continuação para o trecho. É possível que a divisão esteja presente, na verdade, na mente dos falantes. Dessa forma, *bolo* seria a palavra mais provável, mesmo para os participantes que não a tenham expressado durante a tarefa. Aspectos probabilísticos do processo de resposta à tarefa seriam os responsáveis pela variabilidade das respostas, e não as diferenças nas experiências linguísticas individuais (STAUB *et al.*, 2015).

Shaoul, Baayen e Westbury (2014) realizaram um estudo acerca da relação entre informações probabilísticas derivadas de *corpus* e medidas de *cloze* verificadas após microcontextos de três palavras (*3-grams cloze task*). As probabilidades de *cloze* refletiram as distribuições de frequências observadas no *corpus*. A probabilidade condicional de uma palavra ocorrer em determinado contexto foi associada à frequência de respostas e à ordem de múltiplas respostas fornecidas pelos participantes na tarefa. Os autores sugerem que a ligação entre as medidas verificadas em *corpus* e as probabilidades de *cloze* indicam o uso de um mesmo tipo de informação probabilística pelo sistema linguístico durante a tarefa. Tais informações são derivadas do conhecimento implícito do falante, que foi adquirido por meio do uso das palavras em diversos contextos.

Staub *et al.* (2015) verificaram a latência de respostas durante uma tarefa de *cloze*. Por meio de uma versão adaptada da tarefa, na qual os estímulos eram apresentados visualmente e as respostas deveriam ser pronunciadas, os pesquisadores encontraram evidências de que a latência das respostas é influenciada pelas probabilidades de *cloze*. Respostas mais esperadas eram produzidas em menor tempo do que respostas menos esperadas. O nível de restrição dos contextos também influenciou o período de latência. Contextos muito informativos e restritivos possibilitavam respostas mais rápidas do que contextos pouco informativos.

Estudos que utilizam rastreamento ocular comprovam a relação entre medidas de *cloze* e atividades cognitivas relacionadas ao processamento linguístico. Um considerável número de pesquisas revela que, durante a leitura, palavras com alta probabilidade de *cloze*⁵ são fixadas por menos tempo e saltadas com maior frequência do que palavras com baixa probabilidade de

⁵ Para Frisson, Rayner e Pickering (2005), palavras com alta probabilidade de *cloze* possuem percentuais de ocorrência entre 0.70 a 0.90. Já as palavras com baixa probabilidade possuem percentuais abaixo de 0.10.

cloze (ASHBY; RAYNER; CLIFTON, 2005; BALOTA; POLLATSEK; RAYNER, 1985; EHRLICH; RAYNER, 1981; RAYNER; WELL, 1996). Contudo, para que sejam encontradas diferenças significativas em medidas de rastreamento ocular, uma forte diferenciação entre os índices de previsibilidade das palavras parece ser necessária.

As investigações supracitadas evidenciam a relevância das tarefas de *cloze* para os estudos que utilizam a previsibilidade de palavras como variável independente. Sua ampla utilização em estudos sobre o impacto da previsibilidade na compreensão da língua (BICKNELL *et al.*, 2010; FEDERMEIER; KUTAS, 1999; MATSUKI *et al.*, 2011; WARREN; MCCONNELL, 2007, dentre outros) é um indicativo de sua confiabilidade. Contudo, poucos estudos possuem propostas de compreenderem as propriedades das informações ativadas durante a tarefa de *cloze*.

A existência de um conjunto de estímulos adequadamente normatizados é de suma importância para estudos que manipulam a previsibilidade de palavras. A aplicação de tarefas de *cloze* é um dos procedimentos necessários para o manejo seguro dessa propriedade e constitui um dos métodos mais adequados para se confirmar a intuição de pesquisadores sobre a previsibilidade de palavras em dado contexto.

A tarefa de *cloze* (TAYLOR, 1953) apresentada neste artigo tem o intento de verificar como se caracteriza o comportamento de subseleção de argumentos internos por alguns verbos do português brasileiro. O conhecimento sobre propriedades de subseleção dos verbos pode ser utilizado em experimentos que investigam o processamento antecipatório, uma vez que argumentos frequentemente subselecionados por certos verbos – cujas probabilidades de *cloze* são muito altas – podem direcionar a antecipação de traços semânticos de determinados itens lexicais.

Há possibilidade de que, quanto menor o número de entidades citadas como previsíveis diante de determinado verbo flexionado, mais forte é a capacidade de subseleção de argumentos internos por parte desse verbo, devido ao número limitado de itens lexicais capazes de ocupar a posição. Ou seja, quando um número restrito de entidades é citado na tarefa, possivelmente as informações temáticas relacionadas ao verbo testado são mais bem delimitadas e mais restritas. Com isso, os traços semânticos das palavras passíveis de ocupar a posição de argumento desses verbos seriam mais bem definidos, facilitando o processamento antecipatório.

Por meio dos resultados das tarefas de *cloze*, foi criado um *corpus* com os dados de subseleção de 520 verbos. O *corpus* contém dados que mostram o comportamento temático de verbos classificados como portadores de forte e fraca capacidade de subseleção de argumentos.

A realização dessa parametrização possibilita uma melhor compreensão do comportamento temático dessa classe e, por consequência, um maior controle dos itens utilizados em experimentos psicolinguísticos. Na seção a seguir, serão apresentados o método de criação e a análise dos dados.

Materiais

Os estímulos das tarefas de *cloze* consistiam em 17 listas com fragmentos de sentenças em que nomes próprios eram colocados na posição de argumento externo e combinados aleatoriamente com verbos flexionados no pretérito perfeito do indicativo. Uma lacuna foi posicionada após cada um dos fragmentos para que os participantes completassem as frases. A seguir, estão expostos dois exemplos de estímulos empregados na tarefa:

(4) Maria derramou_____.

(5) João chutou _____.

A restrição dos conteúdos semânticos dos fragmentos de sentenças possibilitou a verificação das propriedades de antecipação de argumentos por parte dos verbos. A opção por se utilizar nomes próprios foi motivada pelo fato de esse tipo de nome possuir uma carga semântica mínima. A simplicidade dos contextos permitiu que os eventos denotados pelos verbos fossem salientados.

O Dicionário gramatical de verbos (BORBA, 1990) foi usado para selecionar os verbos presentes nos fragmentos de sentença. Foram escolhidos 520 verbos que possuem tendência a apresentar regência transitiva direta. Ao final desse procedimento, criaram-se 17 diferentes listas que continham entre 30 e 33 itens. A variação entre o número de itens nas listas ocorreu devido à impossibilidade de todas as listas possuírem os mesmos números de itens. Cada fragmento ocorria em apenas uma lista.

Participantes

Cada uma das 17 listas com tarefas de *cloze* foi respondida por 52 estudantes de graduação e pós-graduação da Universidade Federal de Minas Gerais. Todos os participantes eram falantes nativos de português brasileiro, com idades entre 18 e 30 anos (média de 22 anos). Dos 884 respondentes, 646 (73%) eram mulheres.

Procedimentos

As tarefas foram feitas *on-line*, por meio das plataformas *Sona Systems* e *Google Forms*, em uma única sessão de 15 minutos, em média. Os participantes realizavam a tarefa no lugar que julgassem mais adequado, sem a presença do pesquisador. Solicitou-se aos participantes que realizassem a tarefa em ritmo natural e que outras tarefas não fossem efetuadas simultaneamente. Os participantes deveriam completar os fragmentos de sentenças com “os primeiros finais que lhes viessem à cabeça”. As orientações da tarefa explicavam que respostas muito complexas não eram demandadas e que, caso fossem necessárias, algumas respostas poderiam ser repetidas.

Resultados e discussão

As tarefas de 26 participantes (3% do total) foram excluídas por não possuírem no mínimo 70% das respostas ou por apresentarem, em mais de 70% dos itens, respostas inadequadas ao propósito deste estudo – com continuações sem sentido (*e.g.*, continuações implausíveis) ou com repetições injustificadas (*e.g.*, repetições da mesma palavra em mais de 30% das respostas). Após a exclusão, as 17 diferentes listas mantiveram-se com pelo menos 50 respondentes.

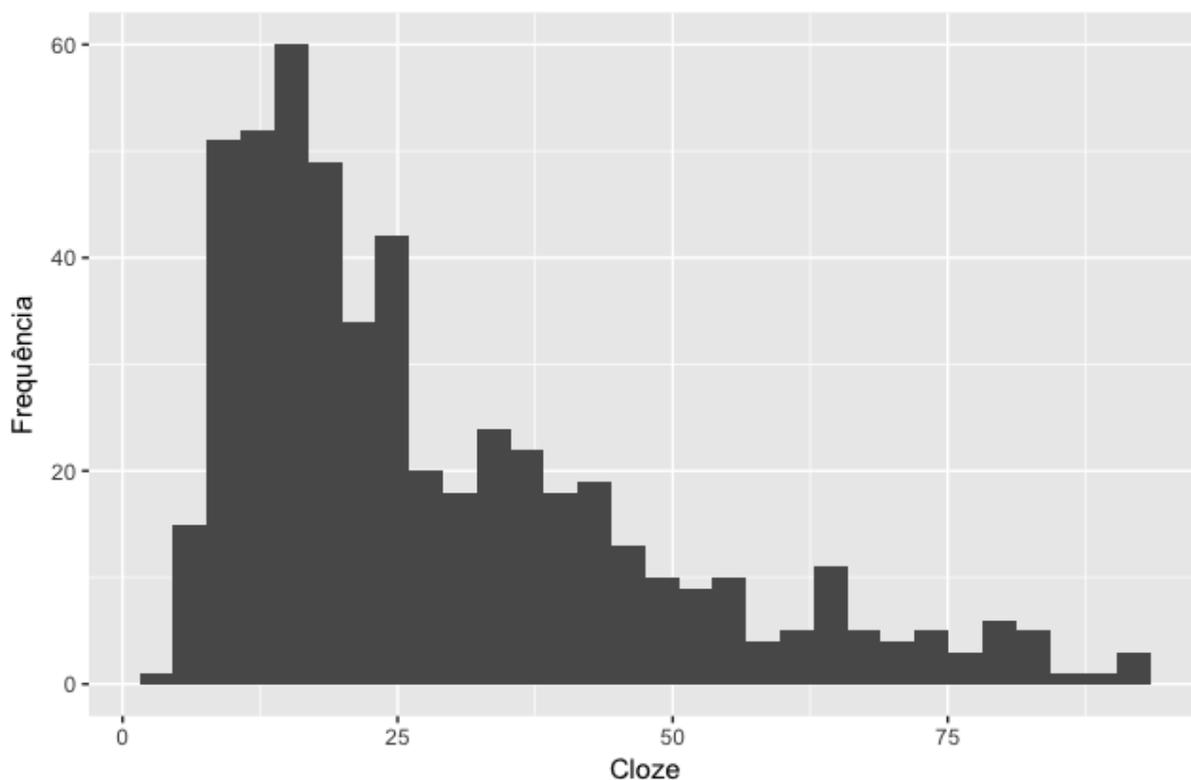
As respostas elaboradas para cada fragmento foram analisadas e codificadas. Passagens respondidas com continuações sem sentido ou em branco eram codificadas como SR (sem resposta). As respostas que possuíam apenas adjuntos foram transcritas da forma como foram respondidas, ou seja, sem qualquer complemento. (*e.g.*, “Joana arrasou na festa”). Apenas o núcleo do sintagma nominal foi considerado quando as continuações eram longas, como “derramou o suco que estava em cima da mesa”. Por meio de um *script*⁶ na linguagem Ruby, os dados das tarefas foram contabilizados, e o percentual de ocorrências de todas as respostas fornecidas pelos participantes em cada passagem foi calculado. Após o cálculo automático dos percentuais, efetuamos, por meio do programa Excel, o somatório de respostas que possuíam uma mesma palavra em sua forma singular e plural.

Nesta pesquisa, o interesse direcionou-se para a probabilidade da resposta mais citada em cada item. A constituição do *corpus* de dados de *cloze* permitiu a análise da força de subseleção de argumentos internos por parte dos verbos. A distribuição das probabilidades de *cloze* das respostas mais citadas em cada fragmento está no histograma da Figura 1. A maior

⁶ *Scripts* são programas escritos para um sistema de tempo de execução especial que automatiza a execução de tarefas que poderiam alternativamente ser executadas uma por vez por um operador humano.

parte dos verbos testados possui probabilidades de subselecionar argumentos menores do que 0.50, como é possível observar.

Figura 1 – Histograma de distribuição das probabilidades de *cloze* no *corpus*



Na literatura sobre previsibilidade, existe certa dificuldade para se estabelecer os parâmetros em que as diferenças de probabilidades de *cloze* podem exibir efeitos processuais. Diversos estudos utilizam respostas congruentes – respostas passíveis de ocorrer – em dado contexto com objetivo de verificar a existência de diferenças entre respostas de alta e baixa probabilidades. Van Petten e Luka (2012) revisaram 18 estudos que manipulavam as probabilidades de palavras diante de contextos congruentes: nas condições mais previsíveis, as probabilidades de *cloze* variavam entre 0.27 e 0.92; naquelas menos previsíveis, entre 0.01 e <0.50.

Considerando-se a diversidade dos parâmetros que diferenciam a alta e a baixa probabilidade de *cloze*, optou-se por adotar um critério semelhante ao de DeLong *et al.* (2011), em que as palavras de alta probabilidade de *cloze* possuem um valor maior do que 0.50 e as palavras de baixa probabilidade possuem um valor menor do que 0.50. Especificamente, a escolha de tais parâmetros ocorreu devido ao estudo eletrofisiológico de DeLong *et al.* (2011), cujos resultados são consistentes ao comparar manipulações de alta e baixa probabilidades de *cloze*. Ademais, DeLong *et al.* (2011) é uma extensão de DeLong, Urbach e Kutas (2005), um

dos estudos de maior impacto para a literatura sobre antecipação no processamento da linguagem⁷.

Logo, em virtude da pequena restrição contextual dos itens do *corpus*, foi feita uma divisão em que verbos de forte tendência de subseleção possuem argumentos internos mais previsíveis, com probabilidades de *cloze* maiores do que 0.50; e verbos de fraca tendência de subseleção possuem argumentos internos mais previsíveis, com probabilidades de *cloze* menores do que 0.50. O Quadro 1 exemplifica alguns verbos, argumentos subselecionados e probabilidades de *cloze* verificadas para cada argumento.

Quadro 1 – Tabela de probabilidades de cloze: verbos e argumentos subselecionados

| Verbo | Argumento | Probabilidade | Verbo | Argumento | Probabilidade |
|---------------|-------------|---------------|-------------|------------|---------------|
| Embrulhou | Presente(s) | 74.55 | Manobrou | Carro | 85.45 |
| Untou | Forma | 74.55 | Assoou | Nariz | 87.50 |
| Lixou | Unha(s) | 76.36 | Sonegou | Imposto(s) | 90.91 |
| Alisou | Cabelo(s) | 76.36 | Afivelou | Cinto | 90.91 |
| Reformou | Casa | 76.47 | Esubalhou | Olhos | 92.86 |
| Comercializou | Roupas | 8.47 | Boicotou | Evento | 11.54 |
| Apelidou | Amiga | 8.62 | Cadastrou | Se | 11.54 |
| Escoltou | Presidente | 8.62 | Carregou | Celular | 11.54 |
| Eliminou | Sujeira | 4.00 | Impulsionou | Vendas | 9.62 |
| Categorizou | Livros | 5.26 | Inflamou | Causa | 9.62 |
| Invalidou | Documentos | 5.56 | Acusou | Irmão | 9.80 |

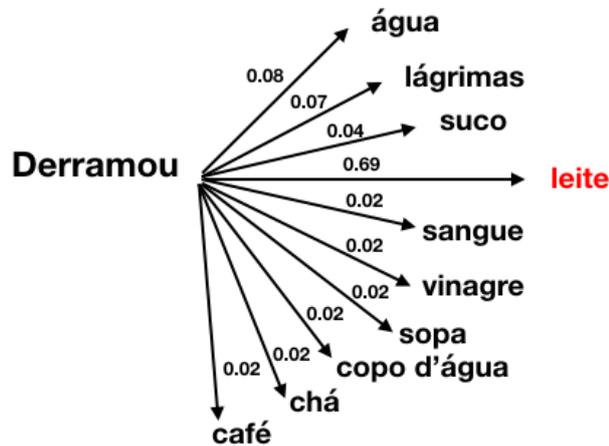
 Verbos de forte subseleção

 Verbos de fraca subseleção

Dessa forma, verbos com *cloze* acima de 0.50 são considerados de forte subseleção, devido ao pequeno número de entidades passíveis de ocupar a posição de argumento interno; e os verbos de *cloze* menor do que 0.50 são considerados verbos de fraca subseleção, devido ao grande número de entidades passíveis de ocupar a posição de argumento interno. Esse comportamento dos verbos considerados de forte e fraca subseleção é ilustrado nas Figuras 2 e 3.

⁷ Embora Nieuwland e Politzer-Ahles *et al.* (2018) não tenham conseguido replicar o efeito de antecipação de formas fonológicas de artigos encontrado em DeLong, Urbach e Kutas (2005), os efeitos relacionados à antecipação de nomes foram devidamente confirmados. Ademais, existem questionamentos acerca da metodologia e da análise inferencial utilizadas por Nieuwland e Politzer-Ahles *et al.* (2018). Para maiores detalhes, veja as respostas de Yan, Kuperberg e Jaeger (2017) e de DeLong, Urbach e Kutas (2017).

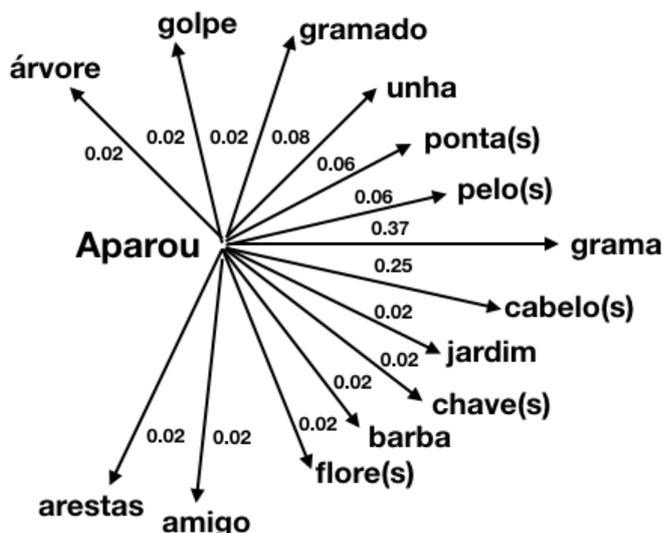
Figura 2 – Comportamento de um verbo de forte subseleção



As propriedades de subseleção do verbo *derramar* demonstram uma alta probabilidade de a palavra *leite* ser subselecionada em casos de contextos mínimos. Isso pode levar à ativação dos traços semânticos presentes na representação dessa palavra. Contudo, é possível observar a presença de outros múltiplos candidatos ativados em paralelo, os quais, por qualquer motivo, poderiam ocupar a posição de argumento interno. Embora esses candidatos possuam traços semânticos comuns à palavra mais esperada, existe uma forte preferência probabilística por uma palavra específica. Isso faz com que uma readaptação do sistema de processamento antecipatório possa ser necessária ao encontrar uma palavra de probabilidade menor.

Essa questão da alta probabilidade de subseleção não ocorre da mesma maneira com o verbo *aparar*. As probabilidades de subseleção desse verbo ativadas paralelamente são mais subdivididas, ocorrendo, inclusive, duas probabilidades muito próximas entre os argumentos com maiores probabilidades, *grama* e *cabelo*. Dessa forma, embora haja uma leve preferência por um argumento específico, essa preferência não parece ser suficiente para que, em contextos mínimos, sejam antecipados os traços semânticos de nenhuma palavra específica. É possível que a ausência de um papel temático bem delimitado possivelmente tornará menos impactante, caso seja necessária, a readaptação do sistema de processamento antecipatório.

Figura 3 – Comportamento de um verbo de fraca subseleção



Portanto, devido às propriedades probabilísticas de subseleção de argumentos internos, é possível que as informações inerentes a cada grupo de verbo influenciem, de maneira distinta, a antecipação de informações de nível semântico durante o processamento de palavras em contextos reduzidos. Essa hipótese, desenvolvida a partir da tarefa de *cloze*, é passível de ser confirmada em outros tipos de técnicas experimentais, com o objetivo de garantir maior compreensão do fenômeno.

Conclusão

Nesta pesquisa, o efeito de previsibilidade lexical, parametrizado pela tarefa de *cloze*, foi associado à capacidade de subseleção de argumentos por parte dos verbos. Verbos cujos argumentos internos possuem probabilidades de *cloze* maior do que 0.50 foram considerados de forte subseleção. Nos casos em que as probabilidades de *cloze* foram menores do que 0.50, consideraram-se os verbos de fraca subseleção.

Existe a possibilidade de que um menor número de entidades citadas como previsíveis diante de determinado contexto demonstre uma maior força de subseleção de argumentos internos por parte do verbo, devido ao número limitado de itens lexicais capazes de ocupar a posição de argumento interno durante a tarefa. Com isso, é possível que, quando um menor número de entidades é citado na tarefa, as informações temáticas relacionadas ao verbo sejam mais restritas.

Sob essa perspectiva, os papéis temáticos de um verbo seriam conceitos formados a partir de experiências cotidianas, durante as quais as pessoas aprendem sobre quem ou o que desempenha determinados papéis em situações específicas. Assim, seria esperado que verbos com capacidade de subselecionar um menor número de entidades do mundo real possuiriam papéis temáticos mais bem delimitados, o que facilitaria a antecipação de traços semânticos de argumentos internos muito esperados.

Vale observar que os resultados produzidos nas tarefas de *cloze* estão sujeitos a experimentos complementares com o objetivo de ratificar, por meio de diferentes técnicas, o comportamento aferido na tarefa. A utilização de técnicas diferentes poderá proporcionar maior conhecimento sobre a organização do conhecimento verbal.

Por fim, ao enfatizar a importância do uso de procedimentos mais cautelosos para a criação de itens experimentais, este artigo expõe a relevância da tarefa de *cloze* para estudos psicolinguísticos que investigam o caráter preditivo do processamento linguístico. Especificamente, evidenciamos como informações sobre o comportamento de subseleção dos verbos, em português brasileiro, podem ser obtidas por meio da tarefa de *cloze*. Por meio de um exemplo prático – a criação de um *corpus* com dada subseleção verbal – evidenciou-se a aplicabilidade da tarefa em contextos diferentes da tradicional legibilidade de passagens de texto.

Referências

ASHBY, J.; RAYNER, K.; CLIFTON, C. Eye movements of highly skilled and average readers: Differential effects of frequency and predictability. **The Quarterly Journal of Experimental Psychology Section A**, v. 58, n. 6, p. 1065–1086, 2005.

BALOTA, D. A.; POLLATSEK, A.; RAYNER, K. The interaction of contextual constraints and parafoveal visual information in reading. **Cognitive psychology**, v. 17, n. 3, p. 364–390, 1985.

BORBA, F. S. (Coord.) **Dicionário gramatical de verbos do português contemporâneo**. São Paulo: Unesp, 1990.

BICKNELL, K.; ELMAN, J. L.; HARE, M.; MCRAE, K.; KUTAS, M. Effects of event knowledge in processing verbal arguments. **Journal of Memory and Language**, v. 63, n. 4, p. 489–505, nov. 2010.

BUBIC, A.; VON CRAMON, D. Y.; SCHUBOTZ, R. I. Prediction, Cognition and the Brain. **Frontiers in Human Neuroscience**, v. 4, 22 mar. 2010. Disponível em: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2904053/>. Acesso em: 17 fev. 2022.

BUTZ, M. V. Anticipation for learning, cognition and education. **On the Horizon**, v. 12, n. 3, p. 111–116, set. 2004.

CLARK, A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. **Behavioral and brain sciences**, v. 36, n. 3, p. 181–204, 2013.

DELONG, K. A.; URBACH, T. P.; KUTAS, M. Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. **Nature Neuroscience**, v. 8, n. 8, p. 1117–1121, ago. 2005.

DELONG, K. A.; URBACH, T. P.; GROPE, D. M.; KUTAS, M. Overlapping dual ERP responses to low cloze probability sentence continuations: Dual ERPs to low probability sentence continuations. **Psychophysiology**, v. 48, n. 9, p. 1203–1207, set. 2011.

EHRlich, S. F.; RAYNER, K. Contextual effects on word perception and eye movements during reading. **Journal of verbal learning and verbal behavior**, v. 20, n. 6, p. 641–655, 1981.

FEDERMEIER, K. D. Thinking ahead: The role and roots of prediction in language comprehension. **Psychophysiology**, v. 44, n. 4, p. 491–505, jul. 2007.

FEDERMEIER, K. D. et al. Multiple effects of sentential constraint on word processing. **Brain research**, v. 1146, p. 75-84, 2007.

FEDERMEIER, K. D.; KUTAS, M. A Rose by Any Other Name: Long-Term Memory Structure and Sentence Processing. **Journal of Memory and Language**, v. 41, n. 4, p. 469–495, nov. 1999.

KUPERBERG, G. R. Separate streams or probabilistic inference? What the N400 can tell us about the comprehension of events. **Language, Cognition and Neuroscience**, p. 1–15, 2016.

KUPERBERG, G. R.; JAEGER, T. F. What do we mean by prediction in language comprehension? **Language, Cognition and Neuroscience**, v. 31, n. 1, p. 32–59, 2 jan. 2016.

KUTAS, M.; HILLYARD, S. A. Reading senseless sentences: Brain potentials reflect semantic incongruity. **Science**, v. 207, n. 4427, p. 203–205, 1980.

KUTAS, M.; HILLYARD, S. A. Brain potentials during reading reflect word expectancy and semantic association. **Nature**, v. 307, n. 5947, p. 161, 1984.

KUTAS, M.; VAN PETTEN, C. K. Psycholinguistics electrified: Event-related potential investigations. In: GERNSBACHER, M. A. (Org.). **Handbook of psycholinguistics**. Cambridge: Academic Press, 1994. p. 83–143.

KUTAS, M.; VAN PETTEN, C. K.; KLUENDER, R. Psycholinguistics electrified II (1994–2005). In: TRAXLER, M.; GERNSBACHER, M. A. (Org.). **Handbook of Psycholinguistics**. 2 ed. Cambridge: Elsevier, 2006. p. 659–724.

MATSUKI, K.; CHOW, T.; HARE, M.; ELMAN, J. L.; SCHEEPERS, C.; MCRAE, K. Event-based plausibility immediately influences on-line language comprehension. **Journal of Experimental Psychology: Learning, Memory, and Cognition**, v. 37, n. 4, p. 913–934, 2011.

MCRAE, K.; FERRETTI, T. R.; AMYOTE, L. Thematic Roles as Verb-specific Concepts. **Language and Cognitive Processes**, v. 12, n. 2–3, p. 137–176, 1 mar. 1997.

NIEUWLAND, M. S.; POLITZER-AHLES, S.; HEYSELAAR, E.; SEGAERT, K.; DARLEY, E.; KAZANINA, N.; WOLFSTHURN, S. V. G. Z.; BARTOLOZZI, F.; KOGAN, V.; ITO, A.; MÉZIÈRE, D.; BARR, D. J.; ROUSSELET, G.; FERGUSON, H. J.; BUSCH-MORENO, S.; FU, X.; TUOMAINEN, J.; KULAKOVA, E.; HUSBAND, E. M.; DONALDSON, D. I.; KOHÚT, Z.; RUESCHEMEYER, S.-A.; HUETTIG, F.. Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. **eLife**, v. 7, p. 1-24, 3 abr. 2018.

RAYNER, K.; WELL, A. D. Effects of contextual constraint on eye movements in reading: **A further examination. Psychonomic Bulletin & Review**, v. 3, n. 4, p. 504–509, 1996.

ROLAND, D.; YUN, H.; KOENIG, J. P.; MAUNER, G. Semantic similarity, predictability, and models of sentence processing. **Cognition**, v. 122, n. 3, p. 267–279, mar. 2012.

SHAOUL, C.; BAAYEN, H.; WESTBURY, C. **N-gram probability effects in a cloze task.** *The Mental Lexicon*, v. 9, p. 1-43, 2014.

STAUB, A. The effect of lexical predictability on eye movements in reading: Critical review and theoretical interpretation. **Language and Linguistics Compass**, v. 9, n. 8, p. 311–327, 2015.

TAYLOR, W. L. “Cloze procedure”: A new tool for measuring readability. **Journalism quarterly**, v. 30, n. 4, p. 415-433, 1953.

VAN BERKUM, J. J. A.; BROWN, C. M.; ZWITSERLOOD, P.; KOOIJMAN, V.; HAGOORT, P. Anticipating upcoming words in discourse: evidence from ERPS and reading times. **Journal of Experimental Psychology: Learning, Memory, and Cognition**, v. 31, n. 3, p. 443–467, 2005.

VAN PETTEN, C.; KUTAS, M. Interactions between sentence context and word frequency in event-related brain potentials. **Memory & cognition**, v. 18, n. 4, p. 380–393, 1990.

VAN PETTEN, C.; LUKA, B. J. Prediction during language comprehension: Benefits, costs, and ERP components. **International Journal of Psychophysiology**, v. 83, n. 2, p. 176–190, fev. 2012.

WARREN, T.; MCCONNELL, K. Investigating effects of selectional restriction violations and plausibility violation severity on eye-movements in reading. **Psychonomic bulletin & review**, v. 14, n. 4, p. 770–775, 2007.

WICHA, N. Y. Y.; MORENO, E. M.; KUTAS, M. Anticipating words and their gender: an event-related brain potential study of semantic integration, gender expectancy, and gender agreement in Spanish sentence reading. **Journal of Cognitive Neuroscience**, v. 16, n. 7, p. 1272–1288, set. 2004.

Sobre a autora

Cláudia Brandão Vieira <https://orcid.org/0000-0003-3298-3366>

Mestre em Estudos Linguísticos pela Universidade Federal de Minas Gerais. Doutora em Estudos Linguísticos pela Universidade Federal de Minas Gerais. Professora substituta no Departamento de Linguística, Português e Línguas Clássicas da Universidade de Brasília.

Recebido em fevereiro de 2022.

Aprovado em setembro de 2022.