

FREQUÊNCIA DE USO DE ÍTEM E INTELIGIBILIDADE DO INGLÊS COMO LÍNGUA FRANCA

TOKEN FREQUENCY AND INTELLIGIBILITY OF ENGLISH AS A
LINGUA FRANCA

Márcia Regina Becker¹, Denise Cristina Kluge²

Resumo: *A expressão latina que tem sido usada e considerada como a melhor para nomear o inglês do novo milênio é “inglês como língua franca”, ELF, cujo objetivo principal é ser mutuamente inteligível entre seus falantes de diferentes nacionalidades. Este artigo apresenta resultados de estudos de percepção de inteligibilidade, em que falantes de inglês de quatro diferentes nacionalidades - alemães, americanos, chineses e japoneses - produziram os estímulos que foram percebidos por brasileiros, falantes de português e que também falam inglês. Concluiu-se que, para os ouvintes brasileiros, a inteligibilidade de alemães, americanos e chineses foi acima de 75% e é independente de sua língua materna, e o grupo de japoneses apresentou os índices mais baixos. Aspectos relacionados à frequência de uso de item, cuja análise foi feita à luz da Fonologia de Uso e da Teoria dos Exemplares, pareceram ter tido um papel determinante na inteligibilidade: palavras pouco frequentes foram também pouco inteligíveis.*

Palavras-chave: Inteligibilidade; Frequência de Uso de Item; ELF (Inglês como Língua Franca).

Abstract: *The Latin expression used and considered to be the best to describe the English of the new millennium is “English as a Lingua Franca”, ELF, whose main aim is to enable speakers of different nationalities to be mutually intelligible. This article shows results of research into perception of intelligibility between speakers of English of four different nationalities - German, American, Chinese and Japanese - who produced the stimuli heard and analyzed by native speakers of Brazilian Portuguese, who also speak English (the listeners). The results showed that, for the Brazilian listeners, the intelligibility of Germans, Americans and Chinese was above 75% and independent of the speakers’ mother tongue, whereas the Japanese group showed the lowest*

1 Doutora em Letras pela UFPR. Professora do Departamento Acadêmico de Línguas Estrangeiras Modernas da UTFPR.

2 Doutora em Letras pela UFSC. Professora adjunta da UFPR.

intelligibility scores. Aspects related to token frequency, analyzed against the backdrop of Usage-Based Phonology and Exemplar Theory, seemed to have played an important role: less frequent words were the least intelligible.

Keywords: Intelligibility; Token Frequency; ELF (English as a Lingua Franca).

Introdução

Considerando-se que o papel da língua inglesa no mundo globalizado é sem precedentes, coloca-se a situação de que, dentro de um mundo em que falantes que possuem o inglês como língua materna já são minoria³, a grande maioria trará, para o inglês que fala, características próprias de sua primeira língua. E esses diferentes “ingleses” podem conviver pacificamente, desde que sejam mutuamente inteligíveis. A questão primordial, então, quando se fala em inglês como *lingua franca*, a expressão latina que tem sido usada e é considerada a melhor (JENKINS 2000, 2007) para nomear esse inglês do novo milênio, é atingir a inteligibilidade.

A inteligibilidade não é um conceito novo, e pesquisas reportadas academicamente a esse respeito datam da década de 50 (CATFORD, 1950). No entanto, o que a princípio pode parecer um conceito bastante simples, na verdade é um construto complexo (BECKER, 2012), com variáveis que podem não ter sequer ainda sido divisadas, apesar de a inteligibilidade ser o objetivo dos falantes de qualquer língua, e também, conseqüentemente, dos usuários de inglês como *lingua franca* ao redor do mundo (JENKINS, 2000; SEIDLHOFER, 2011).

Com relação a estudos de inteligibilidade entre falantes dos diferentes círculos Kachruvianos⁴, de diversas L1s, Pickering (2006) deixa claro que a maioria das pesquisas tem priorizado falantes do círculo interno como ouvintes. Este artigo pretende mostrar como a questão da frequência de uso de item (*token frequency*) foi de relevância para a inteligibilidade de

3 Erling, já em 2005, apontava que a proporção entre o número de falantes de inglês como L2 e o número de falantes de inglês como L1 era de três para um.

4 A difusão da língua inglesa pelo mundo foi visualizada pelo linguista Braj Kachru (KACHRU, 1985) através de círculos concêntricos: círculo interno (países em que o inglês é L1), círculo externo ou estendido (países em que a língua inglesa desempenha um papel de segunda língua num ambiente multilíngue), e círculo em expansão (países que reconhecem o papel do inglês como língua internacional, porém dentro de suas fronteiras ele é ensinado como língua estrangeira, como ocorre com o Brasil).

enunciados - palavras pouco frequentes foram também pouco inteligíveis - em um experimento de inteligibilidade da língua inglesa num contexto de língua franca: brasileiros falantes de português brasileiro foram os ouvintes em testes de percepção de inteligibilidade, e os produtores dos estímulos foram falantes de inglês de quatro nacionalidades distintas: alemães, americanos, chineses e japoneses.

A inteligibilidade está relacionada ao reconhecimento de palavras, e “palavras frequentes são mais rapidamente reconhecidas do que palavras não frequentes”⁵ (BOD; HAY; JANNEDY, 2003, p. 1). Jurafsky (2003, p. 62) também aponta o papel da frequência como fundamental tanto na produção como na percepção, porém deixa claro que evidências sólidas existem apenas para itens lexicais:

palavras de alta frequência são reconhecidas mais rapidamente, com menor *input* sensorial, e com menor interferência de seus vizinhos do que palavras de baixa frequência.[...] Palavras de baixa frequência estão mais sujeitas a erros fonológicos na fala⁶.(Id.)

Por essa razão, a questão da relação entre inteligibilidade e frequência de uso será aqui analisada à luz da linguística probabilística, especificamente através de dois de seus modelos, a Teoria dos Exemplos (JOHNSON, 1997; PIERREHUMBERT, 2001, 2003) e a Fonologia de Uso (BYBEE, 2001, 2007, 2010), que serão detalhados a seguir.

Fonologia de Uso

Bybee (2001, 2007, 2010) propõe um modelo para fonologia e morfologia que considera uma alternativa para a teoria gerativista, ao tratar de relações entre formas morfológicamente relacionadas, incorporar descobertas da categorização fonética e das relações lexicais e ser bastante influenciada pela natureza do *input* de língua recebido: a Fonologia de Uso. A ideia do modelo é a de que

5 A tradução das citações cujos originais estão em inglês é de responsabilidade das autoras deste artigo. Os originais aparecerão em itálico, em notas de rodapé.

Frequent words are recognized faster than infrequent words

6 *High-frequency words are recognized more quickly, with less sensory input, and with less interference by neighbors than low-frequency words. [...] Low-frequency words are more subject to phonological speech errors.*

...o uso que se faz da língua tem um papel importante na formação da forma e conteúdo dos sistemas linguísticos. Em particular a frequência com a qual palavras individuais ou sequências de palavras são usadas e a frequência com a qual certos padrões são recorrentes na língua afetam a natureza da representação mental e, em alguns casos, a real forma fonética de palavras⁷. (BYBEE, 2001, p. 1)

O armazenamento dos itens linguísticos para uso não ocorre na forma de uma listagem, mas de uma rede de conexões de itens relacionados: o esquema (*schema*). Esquemas são, portanto, padrões organizacionais do léxico e assim não têm existência independente das unidades lexicais de onde emergem, sendo altamente afetados pelo número de itens participantes e suas particularidades.

A unidade de armazenamento é a “palavra”, “uma unidade de uso que é tanto fonologicamente quanto pragmaticamente apropriada de forma isolada [...] [Palavras] são unidades de produção e percepção que podem estar sujeitas a categorização. [...] [morfemas] mostram consideravelmente menos autonomia”⁸. (BYBEE, 2001, p. 30). Ainda, “mais comumente uma palavra é representada por um conjunto de exemplares fonéticos com uma pequena faixa de variação associada diretamente com um conjunto de significados”⁹ (BYBEE, 2010, p. 19). Note-se que a noção de “palavra” aqui utilizada é baseada em uso, gradiente, e pode ser entendida como uma sequência do discurso, composta por diversas unidades (ou palavras, com o sentido que usualmente damos a um grupamento de letras), como, por exemplo, *going to* e *I don't know*.

Uma característica importante nesse modelo de categorização fonética é que o principal critério usado é o da similaridade, e unidades (*tokens*) fonéticas são classificadas como membros de uma mesma categoria se forem altamente similares em suas propriedades acústicas e articulatórias. Sendo um postulado linguístico básico o de que a mudança é inerente à natureza da língua, tem-se que as representações da memória são categorizações de uni-

7 *Language use plays a role in shaping the form and content of sound systems. In particular, the frequency with which individual words or sequences of words are used and the frequency with which certain patterns recur in a language affects the nature of mental representation and in some cases the actual phonetic shape of words.*

8 *...a unit of usage that is both phonologically and pragmatically appropriate in isolation. [...] they are units of production and perception that can undergo categorization. [...] they show considerably less autonomy.*

9 *More commonly a word is represented by a set of phonetic exemplars with a small range of variation associated directly with a set of meanings.*

dades de uso - e se elas começarem a mudar, o centro da categoria também vai gradualmente mudar. Isto é particularmente interessante se considerarmos pessoas expostas a diferentes dialetos ou variedades de uma língua:

[...] podemos perguntar por que as pessoas não alteram seu dialeto mais facilmente quando estão cercadas por outro dialeto. As unidades que são ouvidas diferem da armazenada e isso poderia fazer com que a imagem se alterasse. A resposta tem duas partes. Primeiro, quase todos realmente mudam de alguma forma com exposição prolongada a um dialeto diferente, mas eles frequentemente retêm algum traço saliente de seu dialeto original, o que faz com que pareça que não mudaram. Segundo, a forma fonética não é apenas determinada pela representação da memória e experiência – é também determinada por padrões neurais e motores que foram estabelecidos na infância e reforçados por constante repetição¹⁰. (BYBEE, 2001, p. 58).

Os padrões articulatórios compartilhados entre falantes de determinada língua são transferidos para a produção de novas palavras, em palavras emprestadas de outra língua e nas tentativas iniciais de produção numa língua estrangeira. Neste modelo de fonologia de uso, podem-se explicar as regularidades articulatórias com esquemas que se generalizam em relação a padrões aprendidos e que estão presentes na língua.

O fator frequência de ocorrência de item – *token frequency* – é de relevância tanto do ponto de vista de produção como de percepção, e é o que tem mais relação com o trabalho de pesquisa aqui discutido. Há também a frequência de tipo – *type frequency*, esta última sendo a frequência como mostrada em dicionário de um certo padrão (por exemplo, um padrão de acento tônico).

O modelo proposto por Bybee deixa claro que o entendimento tanto da estrutura quanto do uso da língua é realçado pelo fato de que a memória linguística é grandemente afetada pelo próprio uso que se faz da língua.

10 ...we might ask why people don't change their dialect more easily when they are surrounded by another dialect. The incoming tokens differ from the stored image and should cause that image to shift. The answer has two parts. First, almost everyone does shift somewhat with prolonged exposure to a different dialect, but they often retain salient features of their original dialect, which makes it appear that they have not shifted. Second, phonetic shape is not just determined by memory representation and experience – it is also determined by the neural and motor patterns that have been laid down in childhood and reinforced by constant repetition.

A maioria das unidades de uso (*tokens*) da língua são eventos rotineiros que respondem ao meio-ambiente – tanto social quanto físico [...]. Essas respostas são parcialmente automáticas [...] como quaisquer comportamentos neuro-motores de sintonia fina. Como quaisquer outras habilidades neuro-motoras, a linguagem responde à prática. **Habilidades perceptuais também melhoram com repetição.** Então, temos toda a razão para acreditar que a repetição pode ser o fator principal da construção de estrutura da língua¹¹. (BYBEE, 2007, p. 332) (grifos nossos)

Como outra consequência desse modelo, ao considerarmos que cada unidade (*token*) de experiência tem impacto na memória, pois reforça um exemplar já existente ou adiciona-o a um grupo de palavras já existentes, tem-se que a pronúncia de um adulto pode mudar com o tempo, e que para uma criança ou aprendiz da língua em questão, cada nova unidade tem mais impacto do que para um adulto, pois esse já tem um repertório maior na memória. Essas mudanças em adultos são mais sutis, e provavelmente mais lentas. Mas o que se deduz é que, mesmo para um adulto, a adição de novos exemplares a seu repertório tem impacto na sua pronúncia (BYBEE, 2010, p. 21-22).

Teoria dos Exemplos

Johnson (1997) lançou as bases da abordagem de exemplares aplicada à percepção da fala, enquanto Pierrehumbert (2001, 2003) estendeu o modelo também à produção. Nele, as categorias são representadas mentalmente como se fossem nuvens de itens, com os itens muito similares sendo agrupados como um único exemplar, que terá sua “força” aumentada com a adição de mais itens ao grupo. A densidade é maior no centro - onde se encontra(m) o(s) item(ns) mais forte(s), ou o(s) mais procurado(s)/acessado(s). A escolha do exemplar para a produção de uma categoria é aleatória, mas há preferência pelos exemplares mais fortes. Uma nuvem de exemplares traz não apenas informações linguísticas – fonéticas, morfológicas, semânticas e pragmáticas, mas também não linguísticas – fatores sociais -, o que também é apontado por Bybee (2001, p. 52; 2010, p. 21).

¹¹ *Most tokens of language use are routine events that respond to the environment – both social and physical [...] These responses are partially automatic [...] as do other fine-tuned neuromotor behaviors. As with other neuromotor skills, language responds to practice. Perceptual skills also improve with repetition.*

Segundo Pierrehumbert (2003), o fato de um falante adquirir o sistema fonético de uma língua envolve a questão da aquisição de distribuições probabilísticas no espaço fonético: “por ‘espaço fonético’ quero dizer a parametrização acústica e articulatória da fala como evento físico. Por exemplo, [...], as vogais podem ser vistas como distribuições probabilísticas dentro do espaço [dos formantes] F1 e F2.”¹² (*Ibid*, p. 182). Isso corrobora o fato de que “não há caso conhecido de um fonema que tenha exatamente a mesma fonética em duas línguas diferentes”¹³ (*Ibid*, p. 184).

Como em Bybee, na Teoria de Exemplares o falante não armazena os itens (as unidades de armazenamento, sejam morfemas, palavras ou seqüências de palavras) numa lista estruturada, mas em uma rede, com tais itens conectados uns aos outros através de relações fonológicas e semânticas. Em adultos, durante a aquisição de uma língua, em que similaridades e contrastes entre palavras têm um papel importante no vocabulário que está se formando, a competitividade de cada palavra, tanto na percepção como na produção, é fortemente afetada pelo número e frequência de itens a ela similares.

O modelo de exemplares baseia-se na premissa de que “o sistema de uma língua requer uma robusta discriminação de categorias [as nuvens] para garantir a integridade da comunicação.”¹⁴ (PIERREHUMBERT, 2003, p. 209). A classificação dos estímulos na percepção então fornece os dados para as distribuições probabilísticas que controlam a produção, e é mais provável que um indivíduo entenda – e aprenda – uma palavra frequente que uma com baixa frequência.

Uso de Corpora

Quando se fala em estudos que envolvem frequência lexical, não se pode deixar de considerar os *corpora* de que se dispõe como fontes de dados, alguns dos quais foram aqui utilizados. Isso não os exime de problemas. Jurafsky (2003) comenta que desde o início dos estudos relacionados a estatísticas de frequência de palavras, *corpora* foram considerados como eventualmente problemáticos. O pesquisador indica como primeira fonte

12 By “phonetic space,” I mean the acoustic and articulatory parameterization of speech as a physical event. For example, [...], vowels can be viewed as probability distributions over F1-F2 space.

13 ...there is no known case of a phoneme that has exactly the same phonetics in two different languages.

14 ...the language system requires robust discrimination of categories in order to guarantee the integrity of communication.

desses problemas o fato de que um *corpus* é um exemplo de produção de língua, mas as frequências que deles derivam são usadas para experimentos de percepção, e “conquanto frequências de percepção e produção sejam presumivelmente altamente correlacionadas, não há razão para esperar que sejam idênticas.”¹⁵ (JURAFSKY, 2003, p. 43). Outros problemas apontados dizem respeito à qualidade e variedade de fontes dos *corpora*, assim como sua data de coleta dos dados. Jurafsky (2003), no entanto, acredita que esses problemas são mostrados como preconceituosos com relação aos efeitos de frequência apontados por experimentos, e que resultados robustos têm sido encontrados, mostrando que frequências de diferentes *corpora* são altamente correlacionadas.

Apesar de problemas não solucionados com relação ao uso de *corpora*, eles ainda são a ferramenta utilizada para estudos de frequência, e com a disponibilidade de uso de diversos *corpora*, alguns de uso restrito, mas outros gratuitos e *online*, a possibilidade de estudos neles baseados revela padrões quantitativos que antes não eram possíveis, mas que hoje são muito importantes para se entender como os falantes armazenam e processam as unidades da língua (BYBEE, 2007).

No experimento aqui relatado, para fins de verificação da frequência de uso de item, utilizaram-se dois *corpora*: NBC (*British National Corpus*), com 100 milhões de palavras (1980s – 1994), 90% das quais de registro escrito, e o COCA (*Corpus of Contemporary American English*), de 450 milhões de palavras, sendo mais recente (1990-2012) e com fontes tanto orais quanto escritas, de revistas populares a textos acadêmicos e seus correspondentes gêneros. A utilização desses *corpora* das duas principais variedades da língua inglesa tem em vista que o banco de dados utilizado para a produção dos estímulos era de uma universidade americana, mas os ouvintes, na sua grande maioria, utilizaram para formação em língua inglesa material didático cuja variedade britânica foi a utilizada.

Metodologia de Pesquisa

Os testes de inteligibilidade propostos por Isaacs (2008a, 2008b), assim como os de Derwing e Munro (1997), Munro e Derwing (1995a, 1995b, 1999) e Munro, Derwing e Morton (2006) serviram de base para as avaliações feitas neste experimento.

¹⁵ *While comprehension and production frequencies are presumably highly correlated, there is no reason to expect them to be identical.*

Produção dos estímulos de percepção

Neste estudo, os estímulos orais dos testes de percepção foram retirados do *Speech Accent Archive*, da George Mason University (WEINBERGER, 2013), um *corpus* de língua inglesa tomada sob a perspectiva de língua franca. Ele dá uma grande possibilidade de escolha entre falantes de 323 línguas¹⁶, e no caso de muitas delas – e também no caso das línguas maternas dos falantes de inglês escolhidas para esta pesquisa – o alemão, o inglês, o mandarim e o japonês – tem-se uma variedade bastante grande de falantes. A escolha destas quatro línguas foi baseada num aspecto prático: elas são as línguas de países com os quais o Brasil mantém o maior volume de relações comerciais (respectivamente a União Europeia – da qual foi escolhida a Alemanha por ser a economia mais robusta do bloco –, os Estados Unidos, a China e o Japão), e cuja língua de contato nessas interações será, com grande probabilidade, a língua inglesa.

Os falantes do *Speech Accent Archive* escolhidos, que produziram os estímulos de percepção em inglês, foram dois para cada uma das quatro nacionalidades/línguas maternas, um homem e uma mulher. Buscaram-se falantes jovens (média de idade de 21,2 anos), cuja idade era bastante similar à dos seus respectivos ouvintes (média de idade de 22,3 anos), e que eram oriundos de regiões próximas em seus países de origem. Além disso, deu-se preferência aos falantes que não conhecessem outras línguas estrangeiras além da língua inglesa, e que, à exceção do grupo de americanos, houvessem aprendido a língua inglesa em ambiente escolar, tivessem idades próximas, e que morassem por um menor tempo possível em países onde o inglês fosse a língua materna.

A metodologia adotada pelo *Speech Accent Archive* parte da leitura de um parágrafo em inglês, constituído de palavras simples, que, no entanto, contém quase todos os fonemas da língua, sequências de sons e grupamentos de sons que são considerados possíveis causadores de dificuldades. Esse parágrafo, de 69 palavras, foi dividido em 11 excertos (os excertos estão separados um do outro por um “/”) e é o seguinte:

16 Número de entradas quando se pede a pesquisa por línguas. É necessário observar que, na entrada “Chinese”, por exemplo, têm-se 81 falantes de oito línguas maternas diversas, sendo que a maioria são falantes de mandarim (48) e cantonês (21). As oito línguas aparecem ainda na listagem geral das línguas, isto é, aparecem também especificadas isoladamente. Por exemplo, pode ser feita a pesquisa diretamente na entrada “Mandarin”.

Please call Stella¹⁷./ Ask her to bring these things/ with her from the store:/ Six spoons of fresh snow peas,/ five thick slabs of blue cheese,/ and maybe a snack for her brother Bob./ We also need a small plastic snake/ and a big toy frog for the kids./ She can scoop these things into/ three red bags, and we will go/ meet her Wednesday at the train station¹⁸.

Os ouvintes brasileiros

Os ouvintes, dez para cada falante (20 para cada uma das L1s analisadas; 80 no total), eram alunos do curso de Letras Português/ Inglês de uma universidade pública. A escolha de alunos do curso de Letras, além de ser de ordem prática, por se ter mais contato com esses usuários da língua, levou em consideração o fato de que não são ouvintes leigos, e podem, portanto, prestar mais atenção a detalhes fonéticos que serão úteis quando da análise da razão da inteligibilidade ou falta dela - entendimento ou não¹⁹ - de algumas porções da fala. Todos eles tinham tido, no mínimo, 450 horas de instrução em língua inglesa, e precisavam ter tido, nos seus testes de inteligibilidade, um desempenho de, no mínimo, 50% do desempenho de um grupo de ouvintes americanos (um para cada falante) usados para validar o teste. A maioria dos ouvintes era do sexo feminino (73,5%), uma característica do curso de Letras onde estavam matriculados, em que 70% dos alunos eram desse gênero quando os dados foram coletados – 2º semestre de 2012.

17As palavras que estão marcadas em negrito no parágrafo são as *content words* – palavras que carregam conteúdo e que, a princípio, darão conta da mensagem do falante. São as que “em termos de conteúdo, relacionam-se a coisas, ações, e estados no mundo” (McARTHUR, 1992, p.1120). Elas são em número de 41, das originais 69, e são as que foram analisadas sob o ponto de vista de como a sua frequência de uso poderia intervir na inteligibilidade. As demais 28 são palavras funcionais.

...in terms of content relate to things, actions and states in the world.

18 Por favor, chame (telefone para) Stella. Peça a ela para trazer estas coisas da loja: seis colheres de vagens de ervilhas (ervilhas tortas) frescas, cinco fatias grossas de queijo azul, e talvez um lanche para o irmão dela, Bob. Nós também precisamos de uma pequena cobra de plástico e um grande sapo de brinquedo para as crianças. Ela pode colocar essas coisas em três sacolas vermelhas, e nós a encontraremos na quarta-feira, na estação de trem.

19 Apesar de diversos deles já estarem familiarizados com a ideia do construto ‘inteligibilidade’, evitou-se usar o termo, como propõe Isaacs (2008b, p.29), para evitar ser causa de confusão ou eventuais julgamentos tendenciosos.

Procedimento

O experimento constituiu-se de 3 tarefas. Na Tarefa 1, os informantes ouviam todo o texto (os 11 excertos sem pausa, tal como foram gravados e se encontram disponíveis no site) e faziam uma análise impressionística²⁰ do quanto haviam entendido, em percentuais. Na Tarefa 2 era feita a audição excerto a excerto, quando os ouvintes faziam a transcrição ortográfica do que haviam entendido, usando caneta e formulário próprio. Na Tarefa 3, após lerem no formulário cinco itens que poderiam eventualmente vir a ter gerado problemas de inteligibilidade (qualidade da voz, velocidade da fala, ritmo da sentença, sílaba tônica, sons individuais de vogais e consoantes), ouviam pela terceira vez o texto, sem pausas, indicando quais foram os itens causadores de perda de inteligibilidade, e a língua materna do falante, caso tivessem-na reconhecido. O objetivo da Tarefa 3 foi substituir uma espécie de entrevista que poderia ter sido realizada no sentido de analisar o que os havia levado ao não-entendimento (exatamente essas foram as palavras utilizadas) de partes do texto em questão (os itens foram adaptados de Isaacs, 2008b).

Todas as tarefas estão ligadas ao construto inteligibilidade, no entanto a Tarefa 2 está diretamente ligada à definição que orientou este trabalho de pesquisa, que é a de Munro e Derwing (1995b, p. 291),

*A inteligibilidade refere-se à extensão na qual uma produção é entendida de fato. A inteligibilidade, tanto de fala normal quanto patológica, pode ser avaliada apresentando aos ouvintes palavras, sentenças ou unidades mais longas, e pedindo para eles escreverem, em ortografia padrão, o que eles ouviram*²¹.

Procedeu-se à contagem de palavras corretas, isto é, palavras iguais às do texto ouvido. Os mesmos autores, em outro experimento (DERWING; MUNRO, 1997), fizeram a distinção entre erros, categorizando-os como triviais (como por exemplo, correção de pequenos erros gramaticais, plurais) e não triviais (como por exemplo, a transcrição de uma palavra ou grupo de palavras, quando outro foi produzido: *chicken*, ao invés de *she can*, estas últimas as palavras que haviam sido efetivamente produzidas), o que foi também feito nesta pesquisa. Os erros cometidos na transcrição ortográfica e considerados triviais não foram computados.

20 Anotavam o percentual que acreditavam haverem compreendido numa escala que ia de 0 a 100%, organizada em múltiplos de dez.

21 *Intelligibility refers to the extent to which an utterance is actually understood. The intelligibility of both normal and pathological speech may be assessed by presenting listeners with words, sentences, or longer units, and asking them to write, in standard orthography, what they have heard.*

Frequência de uso das palavras de conteúdo

A frequência das palavras de conteúdo utilizadas no experimento foi verificada através dos dois *corpora* já mencionados, o *British National Corpus* – BNC – e o *Corpus of Contemporary American English* – COCA, conforme a Figura 1. Na lista consultada do BNC constam as 6.318 palavras com mais de 800 ocorrências, e não aparecem nomes próprios (*Stella, Bob, Wednesday*), nem numerais (*six, five, three*). Na versão gratuita utilizada do COCA, são listadas as 5.000 palavras mais frequentes, e também não aparecem nomes próprios (*Stella, Bob, Wednesday*).

Excerto	Palavra	BNC		COCA	
		Posição	Frequência	Posição	Frequência
1	<i>please</i>	790	12862	1171	34709
	<i>call</i>	175	53396	122	308050
	<i>Stella</i>	----	----	----	----
2	<i>ask</i>	154	60879	131	284632
	<i>bring</i>	211	43894	216	174366
	<i>things</i>	115	77612	97	400724
3	<i>store</i>	1645	5800	701	56147
4	<i>six</i>	----	----	426	90571
	<i>spoons</i>	5906	890	4384	6194
	<i>fresh</i>	1390	6954	1109	35974
	<i>snow</i>	2627	3112	1795	21011
	<i>peas</i>	Não aparece		Não aparece	
5	<i>five</i>	----	----	300	125571
	<i>thick</i>	2052	4379	1734	21932
	<i>slabs</i>	5985	872	Não aparece	
	<i>blue</i>	1109	9089	845	47622
	<i>cheese</i>	2783	2864	2122	17416

6	<i>maybe</i>	965	10472	384	108421
	<i>snack</i>	Não aparece		Não aparece	
	<i>brother</i>	864	11757	615	63406
	<i>Bob</i>	----	----	----	----
7	<i>also</i>	81	124884	87	464606
	<i>need</i>	147	62201	132	276744
	<i>small</i>	183	51626	203	185463
	<i>plastic</i>	1999	4511	1532	24563
	<i>snake</i>	5513	1005	3512	8523
8	<i>big</i>	282	33300	162	227169
	<i>toy</i>	3697	1848	2441	13935
	<i>frog</i>	5688	956	Não aparece	
	<i>kids</i>	1627	5860	313	126428
9	<i>can</i>	37	266116	37	1022775
	<i>scoop</i>	Não aparece		Não aparece	
	<i>things</i>	115	77612	97	400724
10	<i>three</i>	----	----	135	266744
	<i>red</i>	791	12857	598	66217
	<i>bags</i>	1389	6955	1011	40007
	<i>go</i>	40	249540	35	1151045
11	<i>meet</i>	267	34970	289	128737
	<i>Wednesday</i>	---	---	---	---
	<i>train</i>	1220	8220	1701	21766
	<i>station</i>	829	12328	844	46299

* A coluna “frequência” indica o número de ocorrências no corpus em questão. Por exemplo, no caso do verbo *can*, a posição que ocupa tanto no BNC quanto no COCA é a mesma, 37ª, mas o número de ocorrências é bastante diferente.

** O “não aparece”, presente em espaços de informações de algumas palavras, significa que sua frequência está acima da 6318ª posição para o BNC e acima da 5000ª posição para o COCA. São palavras de uso pouco frequente, portanto.

Figura 1: frequência de uso das palavras de conteúdo. Fonte: BNC, 2012; COCA, 2012.

Observando-se que quanto mais baixa a posição em cada um dos *corpora*, mais frequente é a palavra, e tomando-se por base as palavras de conteúdo utilizadas neste trabalho e a frequência de uso segundo esses dois

corpora, temos a distribuição dessas palavras por faixa de frequência, conforme a Figura 2.

Faixa de frequência (posição)	BNC		COCA	
	Palavras	% de palavras do total ^a	Palavras	% de palavras do total
Até 1000 ^a	<i>please, call, ask, bring, things, maybe, brother, also, need, small, big, can, red, go, meet, station</i> (16 palavras)	47	<i>call, ask, bring, things, store, blue, maybe, brother, also, need, small, big, kids, can, red, go, meet, station</i> (18 palavras)	53
1000 ^a a 2000 ^a	<i>store, fresh, blue, plastic, kids, bags, train</i> (7 palavras)	20,5	<i>please, fresh, snow, thick, plastic, bags, train</i> (7 palavras)	20,5
2000 ^a a 3000 ^a	<i>snow, thick, cheese</i> (3 palavras)	9	<i>toy, cheese</i> (2 palavras)	6
3000 ^a a 4000 ^a	<i>toy</i> (1 palavra)	3	<i>snake</i> (1 palavra)	3
4000 ^a a 5000 ^a	-	0	<i>spoons</i> (1 palavra)	3
Acima de 5000 ^a	<i>spoons, peas, slabs, snack, snake, frog, scoop</i> (7 palavras)	20,5	<i>peas, slabs, snack, frog, scoop</i> (5 palavras)	14,5

^a O total de palavras de conteúdo aqui considerado para cálculo do percentual foi de 34 palavras, pois a palavra *things* aparece duas vezes; os números não aparecem no BNC (são três números: *six, five, three*), então não foram aqui considerados (apesar da alta frequência), e os nomes não aparecem em nenhum dos dois corpora (são três: *Stella, Bob e Wednesday*).

Figura 2 - Palavras de conteúdo por faixa de frequência

Verifica-se que praticamente metade das palavras de conteúdo analisadas encontra-se numa faixa de posição que vai até a 1000^a, isto é, são palavras muito frequentes. Apenas sete palavras do BNC e cinco do COCA são muito pouco frequentes, ocupando uma posição acima da 5000^a.

Resultados e discussão

A variável ‘número total de palavras corretas’ (69), obtida através da Tarefa 2, foi a utilizada para o teste estatístico de Kruskal-Wallis (intersujeitos, mais de três níveis), para averiguar se fazia alguma diferença para os brasileiros ouvirem a estes diferentes grupos de produção²². Considerando-se os 80 ouvintes, percebeu-se que isso ocorreu, pois o resultado do teste estatístico foi significativo ($\Pi^2=20,863$; $p<0,001$). Procedeu-se, então, à análise estatística de par a par dos grupos de produção²³, para se verificar onde residia a diferença. Utilizou-se, para isso, o teste de Mann-Whitney, par a par. Verificou-se que os resultados foram significativos para todos os pares em que o grupo de produção de japoneses se encontrava, o que o destacou em relação aos outros três grupos.

Os resultados são mostrados na Figura 3, que aponta também o percentual de acertos, a inteligibilidade (de acordo com os pressupostos adotados para a sua avaliação neste trabalho de pesquisa), portanto, de todas as palavras do texto, que são 69, apenas das palavras de conteúdo, que são 41, e das palavras funcionais, 28.

	% INTELIGIBILIDADE TOTAL DE PALAVRAS	% INTELIGIBILIDADE PALAVRAS DE CONTEÚDO	% INTELIGIBILIDADE PALAVRAS FUNCIONAIS
Alemães	77,2	75,1	80,4
Americanos	77,0	78,8	74,3
Chineses	80,1	76,3	85,7
Japoneses	61,3	56,0	69,3

Figura 3 - Resultados dos testes de percepção da inteligibilidade dos 4 grupos de produção

Para os ouvintes brasileiros, portanto, os grupos de falantes alemães, americanos e chineses apresentam a mesma inteligibilidade, superior ao grupo de japoneses, que se constituiu um grupo distinto. Com relação às palavras consideradas na sua totalidade, o grupo de falantes de japonês foi o que os brasileiros entenderam menos (61,3%), o mesmo ocorrendo com relação às palavras de conteúdo (56,0% de inteligibilidade para os japoneses). Para os brasileiros, não há diferenças significativas na inteligibilidade

22 Tendo sido verificados os valores de curtose e simetria, a distribuição dos dados mostrou-se não-paramétrica, por isso a escolha dos testes mencionados.

23 Os pares testados foram: alemães e americanos, alemães e chineses, alemães e japoneses, americanos e chineses, americanos e japoneses, chineses e japoneses.

de para o total de palavras entre os grupos de alemães (77,2%), chineses (80,1%) e americanos (77,0%). Este fato se repete para as palavras de conteúdo, isto é, os brasileiros não distinguem esses grupos como diferentes entre si (alemães: 75,1%; americanos: 78,8%; chineses: 76,3%) em termos de inteligibilidade. Para as palavras funcionais, no entanto, os brasileiros, além do contraste com o grupo de japoneses, cujo valor de inteligibilidade é o menor dos quatro grupos de produção (69,3%), com diferença significativa em relação a todos os demais, distinguem os grupos de americanos e chineses (cujos valores de inteligibilidade são de 74,3% e 85,7% respectivamente), com os chineses apresentando melhores resultados de inteligibilidade. Pode-se supor que isto se deve ao fato de que os americanos reduzem praticamente todas as vogais das palavras funcionais, monossilábicas em sua maioria, o que dificultaria o entendimento dos brasileiros, acostumados a produzirem (WATKINS, 2001) – e provavelmente perceberem – vogais não reduzidas em língua inglesa, o que poderia vir a comprometer a inteligibilidade das palavras em que elas se encontram.

Tomando apenas as palavras de conteúdo, que foram aquelas utilizadas para análise mais detalhada no que diz respeito ao papel da frequência como um dos fatores determinantes de seu nível de inteligibilidade, teremos, por faixa de inteligibilidade, considerando todos os falantes de todos os grupos de produção envolvidos, os resultados mostrados na Figura 4. Nas duas últimas colunas desta Figura são apresentadas as posições das palavras de acordo com os *corpora* utilizados (posição mais baixa, correspondendo a palavras mais frequentes).

Faixa de inteligibilidade	Palavra	% de acertos	Posição segundo BNC	Posição segundo COCA
>90%	<i>need</i>	97,5	147°	132°
	<i>station</i>	97,5	829°	844°
	<i>maybe</i>	95,0	965°	384°
	<i>also</i>	95,0	81°	87°
	<i>train</i>	92,5	1220°	1701°
	<i>kids</i>	93,7	1627°	313°
	<i>snack</i>	92,5	Acima de 6318°	Acima de 5000°
	<i>fresh</i>	91,2	1390°	1109°
	<i>ask</i>	90,0	154°	131°
	<i>bring</i>	90,0	211°	216°
	<i>big</i>	90,0	282°	162°

de 80 a 89%	<i>six</i>	88,7	----	426°
	<i>plastic</i>	88,7	1999°	1532°
	<i>store</i>	86,2	1645°	701°
	<i>cheese</i>	86,2	2783°	2122°
	<i>please</i>	83,7	790°	1171°
	<i>brother</i>	83,7	864°	615°
	<i>blue</i>	82,5	1109°	845°
	<i>things</i>	81,8	115°	97°
	<i>three</i>	81,2	----	135°
	<i>bags</i>	81,2	1389°	1011°
de 70 a 79%	<i>five</i>	80,0	----	300°
	<i>meet</i>	75,0	267°	289°
	<i>Bob</i>	71,2	----	-----
	<i>small</i>	70,0	183°	203°
de 60 a 69%	<i>go</i>	70,0	40°	35°
	<i>can</i>	68,7	37°	37°
	<i>call</i>	67,5	175°	122°
	<i>Wednesday</i>	66,2	----	-----
de 50 a 59%	<i>toy</i>	61,2	3697°	2441°
	<i>frog</i>	58,7	5688°	Acima de 5000°
	<i>Stella</i>	57,5	----	----
de 40 a 49%	<i>spoons</i>	51,2	5906°	4384°
	<i>red</i>	45,0	791°	598°
de 30 a 39%	<i>snake</i>	43,7	5513°	3512°
	<i>peas</i>	30,0	Acima de 6318°	Acima de 5000°
de 20 a 29%	<i>snow</i>	20,0	2627°	1795°
de 10 a 19%	<i>thick</i>	17,5	2052°	1734°
	<i>scoop</i>	17,5	Acima de 6318°	Acima de 5000°
de 0 a 9%	<i>slabs</i>	8,7	5985°	Acima de 5000°

* Valor médio entre os valores das duas ocorrências da palavra things: 85,0 % no excerto 2 e 78,7% no excerto 9

Figura 4 - Faixas de inteligibilidade x frequência de uso de item

Nas faixas abaixo de 39% de inteligibilidade (poder-se-ia dizer abaixo de 30%, considerando-se que a palavra *peas* teve 30%, isto é, foi inteligível para 30% dos ouvintes), as palavras *snow peas*, *thick slabs* e *scoop* são comuns a todos os grupos de produção, isto é, para cada um dos grupos individualmente, estas palavras figuraram em faixas de inteligibilidades mais baixas (não são o resultado apenas de um cálculo de médias). As palavras nos pares *snow peas* e *thick slabs* eram contíguas e, assim como a palavra *scoop*, desconhecidas dos ouvintes. Esse é um fator bastante significativo

e apontado até por alguns ouvintes explicitamente como causa maior de falta de inteligibilidade (apesar de não aparecer como opção na Tarefa 3). Nestes pares há casos de diferenças de duração de vogais (*peas*), que geraram interpretações diversas (como os homófonos *peace* e *piece*), mas não se pode assegurar se o problema ocorreu por este fator ou pelo desconhecimento das palavras em questão. O mesmo aconteceu com *thick*, em que a fricativa dental foi produzida como alveolar pelos japoneses. A contiguidade com a palavra *slabs* pode ter sido um fator preponderante na sua falta de inteligibilidade, pois nos grupos de alemães e americanos, em que a pronúncia de *thick* era com a fricativa dental, para ela também – assim como para *slabs* – houve muitos casos de falta de inteligibilidade. O padrão observado para cada um dos grupos de produção analisados individualmente se repete, e também fica claro, no que diz respeito à maioria das palavras, o efeito da frequência nos resultados da inteligibilidade: menos frequente, menos inteligível. Reforça-se a ideia da Fonologia de Uso de que “habilidades perceptuais também melhoram com a repetição”²⁴ (BYBEE, 2007, p. 332).

Considerações finais

Para as palavras de conteúdo analisadas neste trabalho de pesquisa, de inteligibilidade mais baixa, a baixa frequência de seu uso pode ser um dos fatores para essa sua baixa inteligibilidade. Os dois *corpora* utilizados neste estudo, o BNC e o COCA, mostram a frequência nos locais de origem (Grã Bretanha e Estados Unidos, respectivamente), sendo que os ouvintes, estudantes da língua inglesa, utilizaram (e têm utilizado) na sua formação material didático que são destas fontes. Consequentemente, as palavras pouco comuns nas regiões de origem são as que pouco aparecem nos livros. Este parece ter sido um fator determinante para a falta de inteligibilidade de algumas palavras, o que corrobora os pressupostos da Teoria dos Exemplares e da Fonologia de Uso: “palavras de alta frequência são reconhecidas mais rapidamente, com menor *input* sensorial, e com menor interferência de seus vizinhos do que palavras de baixa frequência.”²⁵ (JURAFSKI, 2003, p. 62).

O primeiro pressuposto da Fonologia de Uso (BYBEE, 2001) foca exatamente na importância da frequência de uso, quando coloca que a experiên-

24 *Perceptual skills also improve with repetition.*

25 *High-frequency words are recognized more quickly, with less sensory input, and with less interference by neighbors than low-frequency words.*

cia afeta a representação, tanto na percepção quanto na produção. De acordo com o modelo proposto por Bybee, o entendimento tanto da estrutura quanto do uso da língua é realçado pelo fato de que a memória linguística é grandemente afetada pelo próprio uso que se faz da língua, a frequência. A autora deixa claro que as habilidades perceptuais são também melhoradas pela repetição (BYBEE, 2007). Pierrehumbert (2003), na Teoria de Exemplos, também coloca que há mais “força” num exemplar mais utilizado do que num de baixa frequência, e este exemplar mais frequente é o mais acessado, mais produzido e mais facilmente percebido. Os resultados, consideradas as limitações desta pesquisa em termos de abrangência, parecem confirmar que adquirir o sistema fonético de uma língua envolve a questão da aquisição de suas distribuições probabilísticas no espaço fonético.

BIBLIOGRAFIA

- BECKER, Marcia Regina. O Construto “inteligibilidade” da língua inglesa sob o paradigma de língua franca. *Anais do X Encontro do CELSUL – Círculo de Estudos Linguísticos do Sul*. Unioeste, Cascavel, 2012.
- BECKER, Márcia Regina. *Inteligibilidade da Língua Inglesa sob o Paradigma de Língua Franca: Percepção de Discursos de Diferentes L1s por Brasileiros*. 257f. Tese (Doutorado em Letras) – Universidade Federal do Paraná, Curitiba, 2013.
- BOD, Rens; HAY, Jennifer; JANNEDY, Stefanie. Introduction. In: BOD, Rens; HAY, Jennifer; JANNEDY, Stefanie. (Ed.) *Probabilistic Linguistics*. Cambridge, USA: MIT Press, 2003. p. 1-10.
- BNC – British National Corpus. Disponível em <<http://www.natcorp.ox.ac.uk/>> Acesso em 10 de setembro de 2012.
- BYBEE, Joan. *Phonology and Language Use*. Cambridge: CUP, 2001.
- BYBEE, Joan. *Frequency of Use and the Organization of Language*. Oxford: OUP, 2007.
- BYBEE, Joan. *Language, Usage and Cognition*. Cambridge: CUP, 2010.
- CATFORD, John. *Intelligibility*. *English Language Teaching Journal*, v. 1, n.1, p. 7-15, 1950.
- COCA – The Corpus of Contemporary American English, Brigham Young University. Disponível em <<http://www.wordfrequency.info/>>. Acesso em 10 de setembro de 2012.
- CRYSTAL. David. *English as a Global Language*. Cambridge: CUP, 2003.

- DERWING, Tracey M.; MUNRO, Murray J. Accent, Intelligibility, and Comprehensibility – Evidence from four L1s. *SSLA - Studies in Second Language Acquisition*, 19, p. 1 – 16, 1997.
- ERLING, E. J. The many names of English. *English Today* 81, v. 21, 40-44, 2005.
- ISAACS, Talia. Towards defining a valid assessment criterion of pronunciation proficiency in non-native English-Speaking graduate students. *The Canadian Modern Language Review*, v. 64, n. 4, p. 555-580, June 2008a.
- ISAACS, Talia. *Assessing Second Language Pronunciation: A Mixed Methods Study*. Saarbrücken: VDM Verlag Dr. Müller e.k., 2008b.
- JENKINS, Jennifer. *The Phonology of English as an International Language*, Oxford: OUP, 2000.
- JENKINS, Jennifer. *English as a Lingua Franca: Attitude and Identity*. Oxford: OUP, 2007.
- JOHNSON, Keith. Speech perception without speaker normalization: an exemplar model. In: JOHNSON, Keith; MULLENIX, John W. (Ed.) *Talker Variability in Speech Processing*. San Diego: Academic Press, 1997. p. 145-166.
- JURAFSKY, Dan. Probabilistic Modeling in Psycholinguistics: Linguistic Comprehension and Production. In: BOD, Rens; HAY, Jennifer; JANNEDY, Stefanie. (Ed.) *Probabilistic Linguistics*. Cambridge, USA: MIT Press, 2003. p.39-95.
- KACHRU, Braj. B. Standards, codification and sociolinguistic realism: the English language in the outer circle. In: QUIRK, Randolph; WIDDOWSON, Henry G. (Ed.) *English in the World: Teaching and learning the language and literatures*. Cambridge: CUP, 1985. p. 11-30.
- McARTHUR, T. (Ed.) *The Oxford Companion to the English Language*. Oxford: OUP, 1992.
- MUNRO, Murray J.; DERWING, Tracey M. Foreign Accent, Comprehensibility, and Intelligibility in the Speech of Second Language Learners. *Language Learning*, 45:1, p. 73-97, 1995a.
- MUNRO, Murray J.; DERWING, Tracey M. Processing Time, Accent, and Comprehensibility in the Perception of Native and Foreign-Accented Speech. *Language and Speech*, v. 38, n. 3, p. 289-309, 1995b.
- MUNRO, Murray J.; DERWING, Tracey M. Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, 49, p. 285–310, 1999.
- MUNRO, Murray J.; DERWING, Tracey M.; MORTON, Susan L. The Mutual Intelligibility of L2 Speech. *SSLA - Studies in Second Language Acquisition*, 28, p. 111 – 131, 2006.

- PICKERING, Lucy. Current research on intelligibility in English as a lingua franca. *Annual Review of Applied Linguistics*, 26, p. 219-233, 2006.
- PIERREHUMBERT, Janet B. Stochastic phonology. *Glott International*, v. 5, n. 6, p. 195-207, 2001.
- PIERREHUMBERT, Janet B. Probabilistic Phonology: Discrimination and Robustness. In: BOD, Rens; HAY, Jennifer; JANNEDY, Stefanie (Ed.) *Probabilistic Linguistics*. Cambridge, USA: MIT Press, 2003. p. 177-228.
- SEIDLHOFER, Barbara. *Understanding English as a Lingua Franca*. Oxford: OUP, 2011.
- WATKINS, M. A. *Variability in vowel reduction by Brazilian speakers of English*. 159 f. Tese (Doutorado em Letras/Inglês e Literatura Correspondente), Universidade Federal de Santa Catarina, Florianópolis, 2001.
- WEINBERGER, Steven. *Speech Accent Archive*. George Mason University. Disponível em: <<http://accent.gmu.edu>>. Acesso em: 28 de Fevereiro de 2013.

Recebido em: 30/12/2014. Aceito em: 19/03/2015.