Desafios para a construção de um corpus de aprendizes de Interpretação Simultânea

Challenges to build a learners' corpus of Simultaneous Interpreting

Luciana Latarini Ginezi*

Resumo: Os Estudos da Interpretação, vinculados aos Estudos da Tradução, são ainda novos no Brasil e, por conseguinte, poucas pesquisas experimentais são publicadas na área. Este estudo apresenta parte de uma pesquisa empírica que está sendo conduzida em âmbito de doutoramento na Universidade de São Paulo, cuja proposta alia os Estudos da Interpretação à metodologia da Linguística de Corpus. Nele, discutimos questões conceituais e metodológicas da Linguística de Corpus aplicadas à Interpretação. A descrição detalhada da construção de um corpus de aprendizes de interpretação simultânea, denominado CAIS – Corpus de Aprendizes de Interpretação Simultânea –, nos leva a refletir sobre os problemas que ainda permeiam a construção de corpora falados, bem como desafios para o futuro a serem enfrentados por pesquisadores. A proposta, assim, esclarece dúvidas quanto à pertinência em realizarmos pesquisas experimentais com a metodologia da Linguística de Corpus aplicada aos Estudos da Interpretação, como também engrandece ambas as áreas, no sentido de abrir possibilidades de novas pesquisas em Interpretação e oferecer um roteiro para a elaboração de corpus de aprendizes em língua falada.

* Doutoranda em Estudos Linguísticos e Literários em Inglês – USP – SP. Coordenadora acadêmica e professora do curso de Bacharelado em Tradutor e Intérprete – Universidade Nove de Julho – São Paulo – SP. luginezi@usp.br

Palavras-chave: Interpretação Simultânea; Linguística de Corpus; Corpus de aprendizes; Estudos da Interpretação.

Abstract: In Brazil, Interpreting Studies, a part of Translation Studies, are still new and a few experimental researches were conducted so far. This paper presents a branch of an ongoing empirical research for a doctoral degree at University of São Paulo. The proposal combines Interpreting Studies with Corpus Linguistics, considering conceptual and methodological aspects. The detailed description of CAIS - *Corpus de Aprendizes de Interpretação Simultânea* (Simultaneous Interpreting Learners Corpus) proposes a reflection on the problems still related to the building of spoken corpora. It also presents the future challenges researchers will face. Therefore, the paper reaffirms the advantages of conducting experimental researches using Corpus Linguistics and Interpreting Studies, at the same time that it contributes to these areas, helping new researches in Interpreting Studies and offering a guide to the construction of learners' corpus using spoken language.

Keywords: Simultaneous Interpreting; Corpus Linguistics; Learners' corpora; Interpreting Studies.

Introdução

Os programas de formação de intérpretes no Brasil ainda são jovens, datando de meados do século XX, mais especificamente nos anos 70 (PAGURA 2010A), sendo a PUC-RIO sua precursora no Brasil. Consequentemente, há poucas pesquisas conduzidas sobre o ensino de interpretação em território nacional se comparadas com áreas afins, como ensino de tradução, por razões diversas, dentre as quais citamos duas de grande impacto: o desinteresse de intérpretes profissionais pela área acadêmica, bem como a existência de poucos programas stricto Sensu em Estudos da Tradução, área da qual fazem parte os Estudos da Interpretação (WILLIAMS & CHESTERMAN 2002). Apesar de outros teóricos afirmarem

que os Estudos da Interpretação possam ser uma disciplina independente (MUNDAY 2012), ainda há controvérsias sobre tal divisão. Segundo ele:

> *In view of the very different requirements and activities associated with interpreting, and despite inevitable points of overlap, it would probably be best to consider interpreting as a parallel field, under the title of "interpreting studies"* (MUNDAY 2012: 20).

No Brasil, entretanto, ambas deveriam permanecer sob a mesma área, nesse caso, Estudos da Tradução, atribuindo aos Estudos da Interpretação linhas de pesquisa específicas, que consigam abranger suas peculiaridades, visto que uma ruptura causaria o enfraquecimento de ambas as áreas.

WILLIAMS & CHESTERMAN (2002) afirmam que os Estudos da Interpretação estão inseridos nos Estudos da Tradução, pois reúnem características semelhantes em termos de função social, mediação cultural, dentre outras. No entanto, por maior que seja a proximidade e entrelaçamento das duas disciplinas, outras áreas influenciaram os Estudos da Interpretação, tais como a Psicologia, Linguística, Sociologia, os Estudos Culturais e suas ramificações (PÖCHHACKER 2004).

Os autores agruparam os tópicos para a realização de pesquisas em Estudos da Interpretação, que consistem em: 1) Estudos cognitivos; 2) Estudos comportamentais; 3) Estudos linguísticos; 4) Estudos sociológicos, éticos e históricos; 5) Ensino de Interpretação; 6) Avaliação da qualidade; 7) Tipos especiais de interpretação (WILLIAMS & CHESTERMAN 2002). Para este trabalho, desenvolveremos o quinto tópico acima apresentado.

A relevância do tema ora apresentado, que parte do doutoramento em andamento da autora, está atrelada à importância do desenvolvimento dos Estudos da Interpretação no Brasil, bem como ao crescente número de cursos de formação de intérpretes nas Universidades (IES – Instituições de Ensino Superior)

brasileiras, que podem se beneficiar dos estudos aqui conduzidos, com a replicação do experimento.

O contexto geral da pesquisa de doutoramento em desenvolvimento é avaliar as produções de alunos em interpretação simultânea, em diferentes momentos da aprendizagem, a fim de analisar se o ensino prévio da interpretação consecutiva poderá ser indicativo de que a consecutiva deva realmente ser pré-requisito para a simultânea. Apesar de haver recomendação da literatura para o uso da consecutiva como pré-requisito da simultânea, (ex.: SELESKOVITCH 1999, PAGURA 2010) não há pesquisas científicas que comprovem essa necessidade (GILE 2005A). Dentre os paradigmas da pesquisa em Interpretação (PÖCHHACKER 2004), o estudo somente do texto de chegada atrelado ao estudo das regularidades nele presentes é defendido por SHLESINGER (1989 apud PÖCHHACKER 2004) e PÖCHHACKER (2004), e é a abordagem selecionada para a continuidade da pesquisa de doutoramento. Não se trata, no entanto, de dizer que esse paradigma é melhor ou não que outros, mas de que é a opção para o estudo ora conduzido.

No entanto, para este artigo, ater-nos-emos ao levantamento e à discussão dos desafios para construir o CAIS - Corpus de Aprendizes de Interpretação Simultânea - a partir da produção de alunos de interpretação simultânea de graduação, utilizando a Linguística de Corpus como metodologia.

Tendo introduzido o tópico de discussão do artigo, passemos à divisão do artigo. No capítulo 1, abordamos a Linguística de Corpus (LC) e suas aplicações nos Estudos da Interpretação, visto que o CAIS (sua coleta, transcrição e compilação) é fundamentado na metodologia da LC. No capítulo 2, descrevemos os objetivos do estudo, diretamente relacionados à construção do CAIS. No terceiro capítulo, apresentamos a metodologia conduzida para a elaboração do corpus, detalhando os passos necessários para sua conclusão. Nosso objetivo é explicitar a composição do CAIS, a partir de sua coleta de dados. Na conclusão

final, refletimos sobre a pertinência da realização de pesquisas na área de interpretação com o uso de corpus.

1. Linguística de Corpus e os Estudos da Interpretação

Iniciaremos a discussão apresentando a metodologia da Linguística de Corpus (LC) aplicada aos Estudos da Tradução (e Interpretação), que, juntamente com os Estudos Descritivos da Tradução, viabilizaram a pesquisa empírica, não prescritiva na área.

> *The main reason behind the use of corpora in these two fields is identifying phenomena of translation and interpretation as a whole and on a wide scale, in a more or less overt attempt to confirm or disavow results from 50 years ago, resulting from the analysis of rather limited corpora or case studies* (STRANIERO SERGIO & FALBO, 2012: 10).

Não podemos deixar de mencionar que há, no entanto, críticas sobre a validade de uso da LC como ferramenta metodológica de pesquisa, ou sobre os limites de seu uso. O uso de dados quantitativos sem as devidas explicações, por exemplo, tornam o uso da LC sem propósito, como observou TYMOCZKO (1998: 657). Assim, é preciso atentar para que o uso da LC em uma pesquisa possa explicar os dados obtidos e também viabilizar a análise de grande volume de dados.

A aplicação de LC aos Estudos da Tradução é consideravelmente maior em relação aos Estudos da Interpretação, obviamente porque o discurso escrito

permite que textos sejam rapidamente coletados pela internet para a construção de corpora, enquanto o discurso falado exige gravação e transcrição, ou seja, demanda mais tempo.

Vários corpora de interpretação existem ou estão em desenvolvimento, permitindo que pesquisadores realizem investigações relacionadas à interpretação, dentre os quais podemos citar: EPIC – *The European Parliamentary Interpreting Corpus* (RUSSO et al. 2006); DIRSI-C – *Directionality in Simultaneous Interpreting Corpus* (BENDAZZOLI 2010); FOOTIE – *Football in Europe* (SANDRELLI 2012); CorIT – *Italian Television Interpreting Corpus* (FALBO 2012). No Brasil, duas dissertações de mestrado utilizam a metodologia da LC para a construção de seus corpora de interpretações: GINEZI (2007) apresenta um estudo sobre a socioterminologia na área do café, baseado em dados que incluem um corpus de interpretação consecutiva; NEJM (2011) discute a preparação do intérprete para a simultânea com o uso da LC no desenvolvimento do glossário.

Apesar das dificuldades de compilação de corpora orais, como coleta de dados, transcrição e características extralinguísticas que se perdem do texto original (SHLESINGER 1998), é válido que a LC seja utilizada para pesquisas em Interpretação, pois o uso de corpora permite que possamos ir além da observação de alguns poucos textos produzidos na Interpretação, seja por alunos ou profissionais, buscando regularidades na produção de intérpretes, tentando explicá-las, o que dificilmente é obtido através da observação natural. Há, certamente, várias considerações a esse respeito que merecem atenção. Em primeiro lugar, ainda segundo SHLESINGER (1998), temos de superar dois obstáculos principais para o sucesso dos Estudos da Interpretação com Linguística de Corpus: transcrição e dimensão extralinguística.

Os corpora de interpretação são representações ortográficas (ou linguísticas) transcritas a partir da produção oral dos sujeitos envolvidos no discurso.

> *In particular, whereas words in a written language corpus have an orthographic existence prior to the corpus, the words that appear in an orthographic transcription of a speech event only constitute a partial representation of the original speech event (*COOK 1995 apud STRANIERO SERGIO & FALBO 2012: 31).

Por essa razão, a transcrição de um evento de interpretação deve conter, além de representações ortográficas, as representações de pausa, hesitação, repetições, erros de pronúncia, truncamento de palavras e correções da fala como parte das normas de transcrição, considerando que detalhes dessa natureza interferem no processo e na produção final.

A transcrição tem sido vista como um problema para os pesquisadores de corpus na Interpretação por seu caráter lento e meticuloso, considerando também que a validade ecológica de um corpus só será efetiva se a transcrição seguir um conjunto de normas pré-estabelecidas pelo projeto, devidamente explicitadas no seu início. Obviamente, não haverá normas de transcrição idênticas para cada corpus desenvolvido, uma vez que elas devem respeitar e atender o objetivo de cada corpus. No entanto, uma vez que as convenções sejam claras, será mais fácil adaptá-las e utilizar esse corpus para outras pesquisas, confirmando sua validade ecológica. É preciso, porém, documentar todo o processo de transcrição e etiquetagem do corpus para as pesquisas posteriores. Segundo ZANETTIN (2012), "*in order to make these resources available to the wider research community, it is essential to provide adequate documentation for the corpus and its contents.*" Além disso, a transcrição depende, e muito, do software que será utilizado para a análise dos dados, pois cada software é programado de uma forma, e consequentemente exigirá formatação específica de representações ortográficas e de caracteres especiais. Outro detalhe importante é que as ferramentas de análise selecionadas estejam

disponíveis e que sejam de fácil manuseio, para que a pesquisa possa ser amplamente utilizada e replicada pela comunidade acadêmica, bem como tenha continuidade em plataformas de fácil acesso público. Sobre isso, afirma o autor,

> *Finally, translation driven-corpus resources should be designed and implemented with the provision of being distributed across different software platforms, and the data should be as far as possible accessible to the wider research community. (…) So, ideally, the data and the tools themselves should be openly available and accessible to researchers for cross-validation of results. To this end, the adoption as far as possible of common encoding standards seems an advisable choice to ensure access, reuse and exchange of corpus resources by the research community (*ZANETTIN 2012: 78).

Alguns corpora de Interpretação já existentes, como o projeto EPIC, por exemplo, vão além da transcrição, incluindo em sua organização os áudios referentes à transcrição. Esse tipo de corpus, chamado multimodal (STRANIERO SERGIO & FALBO 2012: 32), permite que as análises sejam feitas também a partir do áudio, em muitos casos ampliando o potencial da pesquisa, ou até mesmo utilizando apenas os áudios, de acordo com o objetivo do trabalho. Assim, vemos que há uma diferença entre os corpora de Interpretação: corpus falado, que contém transcrições das interpretações; corpus multimodal, que contém transcrições e áudios das interpretações.

O conceito de corpus que utilizaremos neste trabalho é de uma seleção organizada de dados linguísticos digitalizados, compilados de acordo com o objetivo do estudo, ou seja, deverá conter dados que permitam a análise e avaliação da produção de alunos de interpretação simultânea em duas fases distintas de aprendizagem, que são: 1) de calouros – aprendizes que nunca praticaram a simultânea, nem consecutiva, porém sabem a teoria básica dos modos de interpretação, e estão no início do 3º semestre do curso; 2) de veteranos - após terem aprendido a consecutiva, ou seja, no início do 5º

semestre do curso, sem terem, no entanto, praticado a simultânea. Os alunos selecionados pertencem à mesma IES, que oferece o curso de Tradutor e Intérprete em seis semestres. Como curso híbrido, há disciplinas voltadas à tradução e outras à interpretação. O ensino de teoria da Interpretação ocorre no 2º semestre, com início de prática de *sight translation*. A partir do 3º semestre, haverá prática de interpretação consecutiva curta, em seguida consecutiva com anotações. Nos 5º e 6º semestres os alunos aprendem a simultânea.

FALBO (2001 apud STRANIERO SERGIO & FALBO 2012: 12), relacionou os critérios que fazem parte da interpretação, e que devem ser considerados para a construção de um corpus, descritos de acordo com cinco macrofatores, a saber:

1. intérprete;
2. contexto situacional;
3. modo;
4. linguagem e direcionalidade;
5. tipo de interação.

Cada um desses critérios, sendo o primeiro superordenado do próximo, pode ser dividido em novas categorias, como, por exemplo, o intérprete (1): intérprete profissional, aluno de interpretação, intérprete ad hoc. Ainda, cada uma dessas subcategorias pode ser dividida em idade, gênero, anos de experiência profissional ou formação.

> *By selecting one combination of categories or sub-categories, it is possible to concentrate on a particular communicative situation, thus devising a corpus suited to the specific purpose of obtaining results related to interpretation in that particular communicative situation* (STRANIERO SERGIO & FALBO 2012: 13).

Dessa forma, o corpus será representativo do que se pretende observar, bastando para isso definir os parâmetros de seleção de dados, acima mencionados, relacionados à interpretação. Em relação à representatividade do corpus, conceito controverso na LC, ZANETTIN (2012: 46) se posiciona da seguinte forma:

> *…representativeness is a very elusive concept, often something to strive for rather than something which can reasonably be attained. Given the number of variables to be taken into consideration in the design of a corpus and the practical limitations often posed by funding, copyright restrictions, etc., it is usually unlikely that the ideal corpus can be constructed. Furthermore, representativeness can be seen not only as a descriptive concept, but also as a normative one. It can reasonably be asked 'of what' and 'according to whom' a corpus should be representative* (ZANETTIN 2012: 46).

A abordagem quantitativa nem sempre é suficiente quando se trata da investigação de língua falada. Alguns fenômenos não permitem a simples etiquetagem, mas requerem a observação natural, manual, como o caso de completude do texto de chegada, item que avaliamos na produção do aprendiz de simultânea. Nesses casos, o corpus de tamanho modesto é mais fácil para a manipulação.

Apesar dos obstáculos colocados, os corpora permitem analisar maior volume de dados, explorando as várias regularidades lexicais, discursivas e extralinguísticas, e também podem contribuir para a comunidade acadêmica em geral, se disponibilizados para que outras pesquisas sejam realizadas.

No entanto, cabe frisar que os corpora nos ajudam a compreender como o intérprete realiza seu trabalho, mas não por quê. Para isso, seria preciso investigar a produção dos aprendizes em seus aspectos cognitivos, éticos, sociais, culturais e ideológicos, que vão além do texto traduzido (STRANIERO SERGIO & FALBO

2011), o que não é escopo deste trabalho, portanto não desenvolveremos esse assunto.

2. Descrição geral do CAIS – Corpus de Aprendizes de Interpretação Simultânea

Como dito anteriormente, os critérios para a composição do CAIS foram selecionados em vista da posterior análise e avaliação da produção de alunos de interpretação, em dois momentos distintos de sua aprendizagem.

Veja abaixo, detalhadamente, a composição do CAIS feita a partir dos macrofatores citados anteriormente por STRANIERO SERGIO & FALBO (2012), porém expandidos devido às especificidades de nossa pesquisa:

a) Intérprete - alunos da Graduação, curso de Bacharelado em Tradutor e Intérprete, sem experiência profissional como intérpretes, divididos em dois grupos, que representam as duas fases distintas de aprendizagem:

- CAL (calouros de interpretação): vinte alunos cursando o início do 3º semestre, sem experiência prática em consecutiva ou simultânea, mas com conhecimento teórico dos modos de interpretação;
- VET (veteranos em consecutiva): vinte alunos no início do ensino de interpretação simultânea, cursando o 5º semestre, sem experiência prática em simultânea.

b) Contexto situacional - simulação de prática de interpretação simultânea, em cabine, com uso de recursos audiovisuais (vídeo transmitido em notebook).

c) Modo - interpretação simultânea.

d) Linguagem e direcionalidade - inglês (língua de partida e língua B dos alunos) para português (língua de chegada e língua A dos alunos).

e) Competência linguística – dois tipos de informantes:

- Aprendizes (intérpretes): classificados como B2, C1 ou C2, de acordo com o índice CEFR – *Common European Framework Reference for Languages*[1] - em teste de nivelamento realizado na própria Universidade. Como a variação de nível linguístico é um fator a ser considerado nos cursos de interpretação, optamos por selecionar um número equiparado de alunos em cada nível, a fim de obter resultados descritivos da realidade. A distância linguística entre o informante B2 e o C2 é equilibrada pela presença do informante C1;
- Informantes dos textos de partida (palestrantes): fluentes em inglês, convidados para palestras pelo TED TALKS[2].

f) Tipo de interação - apresentação de uma palestra em vídeo, selecionada por sua linguagem do cotidiano, simulando a realidade.

[1] O curso analisado aplica provas de nivelamento em todo início de semestre, para que professores tenham um parâmetro do nível de língua inglesa dos alunos inscritos em suas aulas. Aplica-se um teste de nivelamento classificatório, quer dizer, todos os alunos são classificados em determinados níveis, os mesmos utilizados pela CEFR – A1 a C2. Assim, alunos e professores possuem um parâmetro da evolução de cada aluno em termos linguísticos, visto que o teste é aplicado semestralmente. O teste utilizado é realizado por professores de Língua Inglesa, considerando as orientações da CEFR. Para seleção de alunos nesta pesquisa, consideraremos os níveis B2, C1 e C2.

[2] Disponível em: http://www.ted.com/talks?lang=pt-br. Acesso em: 06/10/2013.

g) **Texto de partida -** os textos de partida selecionados aleatoriamente apresentam as seguintes características:

- duração: cerca de 5 min.;
- tópicos: gerais, linguagem cotidiana.

h) **Preparação prévia**: inexistente, para que não fosse uma variável na produção individual de cada aluno.

i) **Treinamento dos alunos anterior ao experimento**: o mesmo para todos os alunos, visto que os informantes pertencem à mesma Universidade.

j) **Velocidade da fala nos textos de partida**: cerca de 170 palavras por minuto.

Em relação ao número de alunos selecionados como informantes, observamos os valores absolutos de alunos ingressantes no curso (média de 100 alunos por semestre), número de egressos (média de 30 alunos por semestre), número de alunos com competência linguística para interpretação (média de 50% dos alunos por semestre) e tempo disponível para pesquisa, e determinamos que a coleta fosse de três produções diferentes (três textos de partida por aluno) para 20 aprendizes de cada grupo, atingindo 60 produções distintas em cada grupo, em um curso que comporta cerca de 500 alunos distribuídos em seis semestres.

O curso normalmente atende a média de 100 novos alunos por semestre, divididos em várias turmas iniciantes. No entanto, somente 30% desses alunos conseguem finalizar o curso. Várias razões são apontadas como fatores de evasão pelos próprios desistentes, sendo a mais comum a dificuldade em língua inglesa. Outro fator importante é que dos 100 alunos ingressantes, nem todos possuem nível de proficiência adequado para a prática de interpretação e, dessa forma, não podem pertencer à população pesquisada. Em 2013, uma pesquisa interna nessa mesma IES com os alunos do curso de Tradutor e Intérprete demonstrou

que 60% dos alunos de 5º semestre estavam divididos entre os níveis C1 e C2. No entanto, quando os alunos de todos os semestres foram mensurados, apenas 39% correspondiam a C1 e C2. Dados os números acima, selecionar 20 alunos para cada grupo indica que estamos selecionando uma amostra de cerca de 60% dos alunos com competência linguística suficiente para a pesquisa de cada semestre.

Sendo assim, 20 alunos foram considerados capacitados como informantes, para cada fase do aprendizado, representando apenas 20% dos calouros, mas 70% dos veteranos.

Para a análise do CAIS, utilizaremos o software *WordSmith Tools*[3], versão 6, desenvolvido por Mike Scott (2013). A escolha está relacionada à facilidade de uso da ferramenta, seu valor acessível, bem como às possibilidades de análise oferecidas. Como dito anteriormente, saber qual será o software utilizado antes da transcrição é fundamental para evitar o retrabalho. Como software para o alinhamento, utilizamos o *Abbyy Aligner Online*[4], cuja interface e resultado mostraram-se mais vantajosos para a pesquisa.

O CAIS é constituído de dois subcorpora, que podem ser utilizados como paralelos (bilíngues) ou comparáveis (monolíngues), bem como bilíngues ou multilíngues. Alguns dados são pesquisados apenas no subcorpus comparável monolíngue, como as hesitações, por exemplo, que veremos logo abaixo. No CAIS, os subcorpora bilíngues são alinhados, para que os textos de partida possam ser comparados aos textos de chegada (esta fase ainda está em andamento na pesquisa). Essa é, inclusive, a definição de corpus paralelo que utilizamos: coleção de textos na língua de partida e sua respectiva interpretação para a língua de chegada (BAKER 1996). Para a análise de texto de partida e de chegada, recorremos aos subcorpora paralelos. Já os subcorpora comparáveis serão

[3] Disponível em: www.lexicallynet.com. Acesso em: 06/10/2013.

[4] Disponível em http://www.abbyy.com/aligner/. Acesso em 30/04/2014.

utilizados para a análise contrastiva entre os dois grupos, ou seja, individualmente e em cada língua.

Vejamos, no diagrama abaixo, como o CAIS está subdividido, e sua explicação detalhada em seguida:

Figura 1 - Representação do CAIS

VET_par: corpus paralelo de textos produzidos por alunos com experiência em consecutiva, mas sem experiência em simultânea, do grupo VET, que corresponde ao início do 5º semestre do curso. O corpus é formado por três textos de partida alinhados aos três textos de chegada de cada aluno. O corpus possui 60 textos de chegada e três de partida, devidamente alinhados.

CAL_par: corpus paralelo de textos produzidos por alunos sem experiência em consecutiva, do grupo CAL (calouros), que corresponde ao início do 3º semestre do curso. O corpus é formado por três textos de partida alinhados aos três textos de chegada de cada aluno. O corpus possui 60 textos de chegada e três de partida, devidamente alinhados.

VET_com: corpus comparável, que contém as 60 produções na língua de chegada, sem os textos de partida ou alinhamento, porém etiquetados com representações extralinguísticas.

CAL_com: corpus comparável, que contém as 60 produções na língua de chegada, sem os textos de partida ou alinhamento, porém etiquetados com representações extralinguísticas.

Portanto, o CAIS é constituído de 120 produções em língua de chegada, sendo 60 do grupo VET e 60 do grupo CAL.

Os textos selecionados para a simulação são de língua geral, com vocabulário que envolve conhecimento extralinguístico. Selecionamos língua geral para facilitar a compreensão dos alunos, pois se fosse uma linguagem de especialidade, e o assunto anunciado anteriormente, alguns alunos poderiam pesquisá-lo enquanto outros não, criando uma nova variável. Os textos foram retirados do website TED TALKS, amplamente empregado durante as aulas de interpretação, por seu conteúdo didático, gravações claras, cujos assuntos são interessantes para os alunos. São conferências reais, cuja velocidade de fala é comparada às conferências presenciais, sem o recurso de TP (*tele-prompter*). BENDAZZOLI (2010) afirma que a velocidade de fala de um corpus deve ser comparada à velocidade do tipo de material que estamos trabalhando, ou seja, os textos de partida do CAIS são simulações de conferências. Tais conferências possibilitam nosso trabalho com o conceito de linguagem autêntica, não produzida artificialmente ou para fins específicos (BERBER SARDINHA 2004).

A Interpretação apresenta grande complexidade em seu processamento, portanto quanto mais controladas foram as variáveis da amostra da pesquisa, mais confiáveis serão os resultados finais (SHLESINGER 1998). Porém, segundo GILE

(2005b), devemos replicar a pesquisa a outras combinações de variáveis, para que os resultados obtidos pelas variáveis controláveis (ou independentes) comprovem-se significativos em outras situações. Vejamos por ele mesmo:

> *More generally, it is not only recognized in the literature, but also increasingly stressed in publications on research methodology, that if the effect of an independent variable on a dependent variable is found significant, replications should cover a sufficient number of combinations of values of controlled variables and generate consistent findings to give confidence in the actual significance of this effect* (GILE 2005b: 149-171).

Passemos agora às fases de compilação do corpus.

3. Coleta, transcrição, alinhamento e etiquetagem do corpus

Para coletar os dados do CAIS, vimos que uma série de critérios foi estabelecida, organizando os dados em dois grupos. A coleta foi realizada da mesma maneira para todos os alunos. Enviamos um convite via e-mail aos alunos interessados em participar do experimento, que preenchessem os requisitos já mencionados. Após o aceite, aplicávamos um questionário escrito, a fim de nos certificarmos que os critérios da pesquisa estavam sendo seguidos a rigor. Antes de iniciarmos as gravações, fazíamos um aquecimento com os alunos, em inglês, seguindo um roteiro pré-estabelecido para todos os informantes:

a) Explicação sobre o modo de Interpretação Simultânea e estratégias – revisão de aulas dadas;

b) Possibilidade de errar, corrigir, usar estratégias já aprendidas, ressaltando que o texto de partida não poderia ser interrompido, ou seja, seria uma simulação de simultânea;
c) Orientação sobre o contexto da palestra, informando o vocabulário de língua geral.

A seguir, na cabine de simultânea, alunos realizam a atividade com um notebook interno, fones de ouvido e microfone. O notebook realiza uma cópia das gravações, utilizando o software *Audacity*, versão 2.0.5, produzido pela *SourceForge*[5] e, ao mesmo tempo, há a gravação de backup, utilizando um gravador móvel, que, no caso, foi um *iPhone*[6], distribuído pela Apple (2013) versão 4, pela facilidade de manuseio, disponibilidade e qualidade de gravação. As conversões dos áudios para mp3 foram realizadas em ambos os casos, para que a transcrição pudesse ser testada em softwares de reconhecimento de voz como *Dragon Naturally Speaking*[7], distribuído pela Nuance. No entanto, o software não reconhece português do Brasil fora de aplicativos da empresa Apple. Já o *Online Dictation*[8] possui reconhecimento para o português brasileiro. Mesmo assim, ele não aceita arquivos para transcrição, ou seja, é preciso ouvir a gravação de alunos por fone de ouvido, enquanto dita o que dizem ao software. Assim, as transcrições foram feitas manualmente, pois mesmo utilizando o *Online Dictation* tivemos que fazer correções, incluir pontuação de acordo com a entonação, etc. Os textos de partida, por sua vez, possuem sua transcrição no website TED TALKS, o que facilitou nosso trabalho.

5 Disponível em http://audacity.sourceforge.net/?lang=pt-BR, acesso em 30/04/2014.
6 Disponível em http://www.apple.com/br/iphone/, acesso em 30/04/2014.
7 Disponível em http://www.nuance.com/dragon/index.htm, acesso em 30/04/2014.
8 Disponível em https://dictation.io/, acesso em 30/04/2014.

A interpretação é uma atividade de comunicação social, portanto os aspectos como pausa, respiração ofegante, marcadores conversacionais em geral, silêncio prolongado, dentre outros, são indicadores do comportamento do falante, bem como da mensagem que pretende passar. Alguns podem ser representados na transcrição, como a pausa e os marcadores conversacionais de reformulação e de formulação (prolongamento de vogais; uso de "ãh", "eee...", "então", etc.); no entanto, o silêncio exagerado, gestos, ruídos ao falar e/ou respiros durante afala, não são representados em transcrições, a não ser que o objetivo da pesquisa esteja atrelado a eles; nesse caso poderíamos criar um símbolo ortográfico equivalente ou etiquetar o material manualmente. Cabe observar que o aprendiz de interpretação muitas vezes transforma uma interrogação do texto de partida em uma afirmação no texto de chegada, ou vice-versa, o que também é registrado na anotação.

A convenção utilizada pelo CAIS (vide Tabela 1) considera vários fatores, dentre eles o software linguístico de análise, que, nesse caso, foi o *WordSmith Tools (WST)*, versão 6 (SCOTT 2013) e o *Abbyy Aligner Online* (2008) para o alinhamento. Isso significa que toda e qualquer pesquisa que queira utilizar o CAIS para análises linguísticas deverá utilizar o WST ou similar, devido às características da transcrição aceitas pelo software.

Assim, as convenções pré-estabelecidas para a transcrição do CAIS estão diretamente relacionadas ao objetivo do corpus. Todos os critérios envolvidos na avaliação da produção dos aprendizes estão, de alguma forma, relacionados com etiquetas ou convenções de transcrição, conforme veremos nas tabelas de 01 a 03 abaixo.

Convenções para transcrição
- utilizar pontuação (vírgula, ponto final, ponto de interrogação, ponto de exclamação) marcando a entonação do intérprete;

- utilizar letras maiúsculas de acordo com a norma padrão; - as palavras truncadas serão representadas com barra, próxima à palavra, e serão consideradas erros de pronúncia. Ex: pro/profissional; - risos serão etiquetados, conforme convenção da linguagem digital: <rs>; - erros de pronúncia serão reproduzidos, bem como gramaticais, porém haverá etiqueta para marcá-los <sic>; - os números serão escritos por extenso; - pausas serão representadas por ZZZ, sendo cada Z equivalente a cerca de 1s; - hesitações serão representadas como se fala. Ex: ahn. Para fins de análise, cada hesitação terá uma etiqueta manual logo após sua representação. <hes>; - prolongamento das vogais ou consoantes (três vezes) e serão consideradas hesitações. Ex: eee, aaa, maaaas. Também serão representadas com a etiqueta <hes>; - palavras ou frases inaudíveis: XXX. Cada palavra será representada por um grupo (XXX).

Tabela 1. Convenções de transcrição.

Além da transcrição ortográfica, o texto de chegada é etiquetado com outras informações das estratégias e ações utilizadas pelos aprendizes, que serão úteis para a avaliação de sua produção. Assim, o corpus é etiquetado manualmente, além das convenções acima mencionadas, com outras informações, vide Tabela 2:

	Estratégia ou ação	**Etiqueta**
a	Autocorreção: utilizada durante a simultânea, em que o intérprete, muitas vezes, antecipa a produção final, porém percebe que se equivocou, e volta atrás, corrigindo sua frase. Pode ocorrer em vários outros momentos, até mesmo quando há ajuda do colega de	<cor>

	cabine.	
b	Hesitação: etiquetada para que as palavras que contêm hesitações possam ser facilmente lidas pelo software.	<hes>
c	Neologismo: muitas vezes, ao desconhecer palavras na língua estrangeira, o aprendiz cria sua própria versão na língua materna. Os neologismos podem ser indicadores de erros de tradução e são identificados manualmente.	<neo>
d	Repetição: ocorre na interpretação como forma de elaboração para a próxima frase. Não representa erro, mas pode representar a omissão de uma parte do texto de partida.	<rep>
e	Sentido incompleto ou sem sentido: é indicação de erro na produção final do aprendiz.	<sinc>

Tabela 2: Grupo 1 de etiquetas manuais inseridas nos corpora.

Essa etiquetagem ocorre sem a verificação do texto de partida, utilizando os subcorpora monolíngues e comparáveis.

Outro momento de etiquetagem dependerá do alinhamento do texto de partida ao texto de chegada. Nesse caso, será possível observarmos outras ações ou estratégias dos intérpretes, que ajudam na avaliação e análise final do produto da interpretação, porém são etiquetas manuais. São elas:

	Estratégia ou ação	**Etiqueta**
a	Omissão: omitir informação por perda do original, seja pela rapidez na fala do palestrante ou por falta de compreensão do significado.	<omis>
b	Erro: errar ao traduzir números ou fatos do texto de partida.	<err>
c	Cognatos: usar cognatos em demasia será evidência de erro na produção geral do aprendiz.	<cog>
d	Adição: inserir informações que não estão presentes no texto de partida.	<add>
e	Produção incompleta ou parcial: ao analisar o texto de partida e de chegada, o pesquisar observa que a	<pinc>

	produção está incompleta ou parcial.	

Tabela 3: Grupo 2 de etiquetas manuais inseridas nos corpora.

Para que o corpus fique bem documentado quanto aos detalhes de sua estrutura, após a transcrição, identificamos cada produção com um cabeçalho, que deverá conter os seguintes metadados, digitados em letras minúsculas, em formato *txt*, em formas de etiquetas. Veja o exemplo abaixo, na figura 2:

```
<header>
<recording date> 08.10.2010</recording date>
<interpreting speech number>01</interpreting speech number>
<source language>English</source language>
<target language>Brazilian Portuguese</target language>
<source video length>04:01</source video length>
<speaker student>Elaine</speaker student>
<gender>f</gender>
<course>interpreting practice iii</course>
<semester>5</semester>
<native language>Brazilian Portuguese</native language>
<english proficiency level>c1</english proficiency level>
</header>
```

Figura 2. Exemplo de cabeçalho para identificação dos textos.

No total, os subcorpora do CAIS, VET_COM e CAL_COM, possuem respectivamente 17.556 palavras e 16.289 palavras. São pequenos sob o ponto de

vista da LC em geral; porém, para os fins que servem, contêm dados suficientes para a realização das análises.

Conclusão

Várias questões são colocadas para o desenvolvimento da LC na Interpretação. Vimos que a dimensão dos corpora é uma característica especial para a língua falada, ou seja, os conceitos da LC à dimensão de corpus devem ser cuidadosamente utilizados em se tratando de corpus falado. Quando comparados, corpora escritos são maiores que corpora falados, devido às restrições de tempo e de preparação. Algumas críticas em relação à dimensão do corpus falado sugerem que o tempo e recursos investidos não trazem os resultados esperados. No entanto, dados os corpora de Interpretação já existentes e as possibilidades de pesquisas oferecidas por eles, é imprescindível que novos corpora falados ou multimodais sejam criados em várias línguas e direcionalidades, disponibilizados em plataformas acessíveis, para que outras investigações sejam possíveis. O projeto EPIC, por exemplo, possibilitou que fossem realizadas pesquisas sobre densidade lexical dos textos, contrastes entre textos falados e textos escritos, direcionalidade, estratégias de interpretação, etc.

Considerando a Interpretação como ato de comunicação social, o alinhamento dos áudios às transcrições amplia as aplicações do corpus como instrumento de análise de interações sociais. Se o objetivo da pesquisa está atrelado às relações face a face, o corpus multimodal é o mais apropriado. No entanto, há muitas restrições tecnológicas que impedem a construção de corpus com esse alcance, outro entrave para a realização de pesquisas na área.

Também presente sobre a reusabilidade do corpus, há a questão de que, quando construído especificamente para um objetivo, como o CAIS, como poderá servir para outras pesquisas? Como o CAIS possui um histórico de suas gravações, informantes, textos utilizados, etc. (vide Cap. 3), o investigador poderá recorrer aos dados e adaptar suas variáveis às já existentes. Novamente, definir os objetivos a que se propõe o corpus e enfatizar as variáveis presentes são exigências para garantir a cientificidade do trabalho. Nesse mesmo ponto, divulgar e explicar os códigos ou convenções de transcrição do corpus são também tarefas fundamentais, para que pesquisadores não caiam em armadilhas das nomenclaturas de etiquetas, que não possuem norma padrão. SETTON (2011: 68) afirma que é preciso haver consenso entre as normas básicas para a compilação, etiquetagem e disponibilização dos corpora, a fim de serem multiplicadores de pesquisas.

O CAIS é um corpus aberto, que poderá ser incrementado com o alinhamento de áudio, ser disponibilizado em plataforma acessível para consultas sobre a produção de aprendizes, e dessa forma auxiliar professores de interpretação em suas aulas, para a escolha de técnicas de ensino, bem como ser consulta da atuação de aprendizes em relação a escolhas lexicais, semânticas, comparações com textos escritos, dentre outras funções.

Pretendemos, como projeto futuro, incluir o CAIS, em sua forma completa, multimodal, no escopo do projeto COMET[9] – Corpora Multilíngue para Ensino e Tradução, a fim de ampliar as possibilidades de pesquisa na direcionalidade português – inglês do ensino de Interpretação no Brasil.

[9] Disponível em http://www.fflch.usp.br/dlm/comet/, acesso em 30/04/2014.

Bibliografia

BAKER, M. Corpora in Translation Studies: the Challenges that Lie Ahead. In: SOMERS, H (Org.) *Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager.* Amsterdam/Philadelphia: John Benjamins, 1996: 175-186

BENDAZZOLI, C.; SANDRELLI, A.; RUSSO, M. European Parliament Interpreting Corpus (EPIC): methodological issues and preliminary results on lexical patterns in simultaneous interpreting. In: *IJT – International Journal of Translation,* n. 22/1-2, 2010: 165-203.

BERBER SARDINHA, T. *Linguística de Corpus.* Barueri, SP: Manole, 2004.

COMMON EUROPEAN FRAMEWORK REFERENCE FOR LANGUAGES. *CEFR.* Disponível em: http://www.coe.int/t/dg4/education/elp/elp-reg/cefr_EN.asp. Acesso em: 06/10/2013.

FALBO, C. CorIT (Italian Television Interpreting Corpus): classification criteria. In: STRANIERO SERGIO, F.; FALBO, C (Eds.). *Breaking Ground in Corpus-Based Interpreting Studies.* Linguistic Insights: studies in language and communication; v. 147. Berna: Peter Lang, 2012.

GILE, D. Teaching conference interpreting: a contribution. In: TENNENT, M. (Ed.). *Training for the new millennium***.** Amsterdam/Philadelphia: John Benjamins Publishing Company, 2005a.

_____. Empirical Research into the Role of Knowledge in Interpreting Studies: Methodological Aspects. In: DAM, Helle V.; ENGBERG, J. & GERZYMISCH-ARBOGAST, H. (Eds). *Knowledge Systems and Translation.* Berlin & New York: Mouton de Gruyter, 2005b: 149-171.

GINEZI, L. L. Cafés do Brasil: Estudo de Variantes em Português e Inglês na Língua Falada. Dissertação de Mestrado. Departamento de Letras Modernas. Área de Concentração: Estudos Linguísticos e Literários em Inglês. São Paulo: USP, 2007.

MUNDAY, J. *Introducing Translation Studies.* Theories and Applications. New York: Routledge, 2012.

NEJM, C. C. L. Interpretação simultânea: atividade voltada à comunicação e a linguística de corpus como ferramenta na preparação do intérprete. Dissertação de Mestrado. Departamento de Letras Modernas. Área de Concentração: Estudos Linguísticos e Literários em Inglês. São Paulo: USP, 2011.

PAGURA, R. A interpretação de conferências no Brasil: História de sua prática profissional e a formação de intérpretes brasileiros. Tese de Doutoramento. Departamento de Letras Modernas. Área de Concentração: Estudos Linguísticos e Literários em Inglês. São Paulo: USP, 2010a.

_____. O consenso internacional sobre a formação de intérpretes de conferência. In: *Tradução & Comunicação.* Revista Brasileira de Tradutores. nº 21. São Paulo: Anhanguera Educacional Ltda., 2010b.

PÖCHHACKER, F. *Introducing Interpreting Studies.* London and New York: Routledge, 2004.

RUSSO, M.; BENDAZZOLI, C.; SANDRELLI, A. Looking for lexical patterns in a trilingual corpus of source and interpreted speeches: extended analysis of EPIC (European Parliament Interpreting Corpus). IN: Forum 4/1, 2006: 221-254.

SANDRELLI, A. Introducing FOOTIE (Football in Europe): simultaneous interpreting in football press conferences In: STRANIERO SERGIO, F.; FALBO, C. (Eds.) *Breaking Ground in Corpus-Based Interpreting Studies.* Linguistic Insights: studies in language and communication; v. 147. Berna: Peter Lang, 2012.

SCOTT, M. *WordSmith Tools.* Oxford: Oxford University Press, 1996/2013. Disponível em: www.lexicallynet.com. Acesso em: 06/10/2013.

SELESKOVITCH, D. The Teaching of Conference Interpretation in the Course of the Last 50 Years. In: *Interpreting,* vol. 4(1). London: John Benjamin Publishing Co., 1999: 55-66.

SETTON, R. Corpus-based interpreting studies (CIS): reflections and prospects. In: KRUGER, A.; WALMACH, K.; MUNDAY, J. (Ed.) *Corpus-based Translation Studies: research and applications.* London and New York: Continuum, 2011: 33-75.

SHLESINGER, M. *Corpus-based Interpreting Studies as an Offshoot of Corpus-Based Translation Studies.* Meta, XLIII, 4, 1998: 486-493.

STRANIERO SERGIO, F.; FALBO, C. Studying interpreting through corpora. An introduction. In: STRANIERO SERGIO, F.; FALBO, C. (Eds.) *Breaking Ground in Corpus-Based Interpreting Studies.* Linguistic Insights: studies in language and communication; v. 147. Berna: Peter Lang, 2012: 09-52.

TYMOCZKO, M. Computerized corpora and the future of Translation Studies. In: *Meta* n. 43/4, 1998: 652-660.

WILLIAMS, J.; CHESTERMAN, A. *The Map. A Beginner's Guide to Doing Research in Translation Studies.* Manchester: St. Jerome Publishing, 2002.

ZANETTIN, F. *Translation-driven corpora.* Manchester & Kinderhook: St. Jerome Publishing, 2012.