ÉTUDES ROMANES DE BRNO 44/2023/1
HTTPS://DOI.ORG/10.5817/ERB2023-1-14

Construção do *corpus* "Produção Oral em Provas de Português L2" (POPL2)

Building the *Corpus* "Oral Production in Portuguese L2 Assessment-task types" (POPL2)

Tânia Ferreira [tania.ferreira@fl.uc.pt] / Isabel Santos [imas@fl.uc.pt]
Conceição Carapinha [mccarapinha@fl.uc.pt] / Cristina Martins [crismar@fl.uc.pt]
Isabel Pereira [mipp@fl.uc.pt] / Graça Rio-Torto [gracart@gmail.com]
Liliana Inverno [lcinverno@uc.pt] / Rui Pereira [rui.pereira@uc.pt]
Carla Ferreira [uc41719@uc.pt] / Sara Sousa [sarasousa@uc.pt]
Sandra Chapouto [schapouto@uc.pt]
Universidade de Coimbra, CELGA-ILTEC, Portugal

Resumo
Neste trabalho, apresentam-se os procedimentos adotados para a constituição do *corpus Produção Oral em Provas de Português L2* (POPL2). Pretende-se, com este projeto, obter dados de natureza oral produzidos por aprendentes tardios de Português L2 (PL2) em contexto instrucional e em momento de avaliação. Convocam-se, neste artigo, as questões associadas à conceção e disponibilização de *corpora* de produções orais de aprendentes tardios (Granger 2002; Adolphs & Knight 2010; Adolphs & Carter 2013; Ballier & Martin 2015; Santos *et al.* 2016; Bell & Payant 2021), com especial relevância para os constrangimentos que, neste âmbito, emergem do contexto e condições de recolha, do uso dos instrumentos técnicos para a captação de som e da posterior transcrição de dados orais. Descrevem-se, ainda, as recolhas experimentais que foram realizadas com vista à validação de opções metodológicas.

Palavras-chave
Português L2 (PL2); linguística de *corpus*; produções orais; *corpus* de aprendentes

Abstract
This paper presents the procedures adopted in creating the *Oral Production in Portuguese L2 assessment-task types* (POPL2) corpus. The purpose of this project is to obtain oral data produced by late learners of Portuguese L2 (PL2) in an instructional setting and during assessment tasks. Issues regarding the design and availability of late learners' oral production corpora are discussed (Granger 2002; Adolphs & Knight 2010; Adolphs & Carter 2013; Ballier & Martin 2015; Santos *et al.* 2016; Bell & Payant 2021), especially the constraints related to setting and data collection conditions, the use of technical instruments for sound recordings and the subsequent transcription of spoken data. The experimental data collections that have been carried out to validate methodological options are also described.

Keywords
Portuguese L2; Corpus Linguistics; Spoken data; Learner Corpus

RECEBIDO 2022-10-11; **ACEITE** 2022-12-01

Este trabalho foi financiado pelo CELGA-ILTEC, ao abrigo do respetivo Programa de Financiamento FCT (Fundação para a Ciência e a Tecnologia): UIDB/04887/2020 e UIDP/04887/2020.

1. Introdução

No âmbito do estudo da aprendizagem/aquisição de uma língua não materna, reconhece-se atualmente a importância dos dados de produção de não nativos, organizados em *learner corpora*, uma vez que tanto o trabalho teórico como a descrição das interlínguas encontram nesses materiais sustentação empírica (Myles 2015; Bell & Payant 2021; Meunier 2021). Ao mesmo tempo, reconhece-se que a observação das produções dos não nativos pode ser também muito relevante na definição de práticas pedagógicas e na elaboração de materiais instrucionais adequados (Granger 2002, 2009).

Passam então a integrar-se, nesta área de trabalho, os contributos da Linguística de Corpus e desenvolvem-se projetos de criação e disponibilização de acervos de dados produzidos por aprendentes. Se os dados escritos se configuram de mais fácil recolha e tratamento, é inquestionável a importância que, complementarmente, assumem os registos de produção oral.

Assim, neste trabalho pretende dar-se a conhecer um projeto (em curso) que visa a criação de uma base de dados orais produzidos por aprendentes tardios de Português L2 (PL2)[1], o *corpus Produção Oral em Provas de Português L2* (POPL2), bem como equacionar as questões teóricas e metodológicas tidas em conta no processo de constituição deste acervo.

O presente texto apresenta a seguinte estrutura: na secção 2., analisam-se as questões que se prendem com a constituição dos *corpora* de produções de aprendentes. Em 3., apresentam-se os projetos de elaboração de *corpora* de produções de aprendentes de PL2 existentes no Centro de Estudos de Linguística Geral e Aplicada (CELGA-ILTEC) da Faculdade de Letras da Universidade de Coimbra (FLUC), de modo a contextualizar este novo projeto. A secção 4. é dedicada à descrição do *corpus* POPL2, nomeadamente no que respeita ao processo de recolha de dados (4.1.1), às tipologias de atividades de produção oral (4.1.2), à compilação/disponibilização dos metadados (4.1.3), e à transcrição dos dados orais (4.1.4). Em 4.2., descrevem-se os procedimentos adotados nas gravações-piloto já realizadas. Por fim, na secção 5., apresentam-se as considerações finais.

1 Em função dos objetivos do presente trabalho, e atendendo aos perfis diversificados dos sujeitos informantes que fornecem dados para este *corpus,* optou-se pela utilização da expressão *Português L2* (PL2), tomada em sentido lato, para nos referirmos às situações em que o português é assimilado como língua não materna (LNM), i.e., em fases mais tardias do desenvolvimento dos aprendentes (cf. Flores 2013: 35; Madeira 2017: 306).

2. *Learner corpora*: dados de produções escritas e dados de produções orais

Emergem, na literatura, diferentes propostas com vista à definição do que se considera um *corpus* de produções de aprendentes (veja-se, por exemplo, Granger 2002: 4; Granger, Guilquin & Meunier 2015: 1). Assim, segundo a proposta de Bell & Payant (2021: 54), um *corpus* de produções de aprendentes corresponde a:

> [A]n electronic collection of learner produced data formatted for automatic analyses, elicited from L2 or L3/x learners or users that provides essential metadata, details critical information on elicitation tasks, and is built around explicit and published design criteria".

Não obstante poder defender-se que só contextos autênticos de uso proporcionarão a recolha de linguagem natural (nesse caso, os conteúdos são selecionados com base na sua função comunicativa; cf. Bell & Payant 2021: 56), dificilmente um *corpus* de produções de aprendentes se constitui sem qualquer tipo de constrição e, portanto, sem que se coloque a questão da sua "autenticidade", qualidade cujos limites, aliás, são difíceis de estabelecer[2] (Granger 2002; Mauranen 2004; Bell & Payant 2021: 54)[3].

Com efeito, para o estudo da aquisição e aprendizagem de línguas não maternas, recorre-se a dados empíricos recolhidos, geralmente, a partir de estímulos. O objetivo é aferir o desempenho linguístico dos informantes relativamente a determinadas áreas da língua-alvo de aprendizagem. Deste modo, as produções obtidas poderão estar relativamente condicionadas pelo próprio processo de recolha. Todavia, quando se opta por recolher produções de aprendentes que resultam da aplicação de tarefas não controladas, corre-se o risco de não obter dados suficientemente comparáveis entre si[4] para aferir o domínio de determinadas estruturas linguísticas ou a competência necessária à verbalização de determinadas intenções comunicativas por grupos específicos de aprendentes[5]. A elicitação das produções tem, assim, a vantagem de garantir que, efetivamente, as estruturas em análise são usadas pelos inquiridos/estudantes[6].

2 Como observa Granger (2002: 5), "learner data is […] rarely *fully* natural" (expressão destacada no original) e esta é, efetivamente, uma particularidade inerente aos acervos de produções de aprendentes, tanto na modalidade oral como na modalidade escrita (Myles 2015: 313; Bell & Payant 2021: 54).

3 Como é amplamente difundido na literatura especializada, o contexto em que as produções ocorrem e o próprio processo de recolha de dados orais de aprendentes tardios de uma L2 criam condições que potencialmente limitam a sua "autenticidade". No entanto, e na linha do que defende Mauranen (2004: 92), um *corpus* oral "can therefore achieve a high level of authenticity in the similarity sense by representing crucial aspects of used language, such as the actual wordings and sequences used, with repetitions, overlaps, hesitations and misunderstandings code".

4 "Moreover, unlike role play data which can be replicated using similar social variables in comparable contexts (e.g. age, gender, education, social class, social distance, social power), ordinary natural conversation is not easily comparable and cannot be replicated" (Félix Brasdefer 2007: 179).

5 Segundo Tracy-Ventura & Myles (2017: 60), "collecting more open-ended samples of learner language can be a gamble, potentially leading to somewhat limited productions, with learners 'playing safe' in order to avoid making errors and not fully demonstrating how much they know".

6 Ressalve-se que a produção elicitada pode ser condicionada por graus muito variados de controlo temático e formal, dependendo do protocolo de recolha de dados. Por elicitação podem obter-se, assim, dos informantes, enunciados temática e formalmente muito próximos, apenas diferentes num ou noutro pormenor de interesse para

São claramente maioritários os acervos de produções escritas, não obstante a importância que se atribui aos dados orais. A assunção de que partimos é, precisamente, a de que a descrição dos conhecimentos linguísticos dos aprendentes implica que se tenha acesso também ao seu desempenho na oralidade; enquanto evidências do comportamento linguístico do aprendente, os dados orais e os dados escritos serão, então, complementares (Adolphs & Carter 2013: 5). Relativamente aos dados escritos, os dados orais envolvem, tipicamente, um menor grau de automonitorização, sendo o falante instado a produzir discurso sem o mesmo tempo para processar e organizar ideias e para refletir sobre a língua, acionando conhecimento metalinguístico (Bell & Payant 2021: 57). Já os dados escritos, e porque resultam, tipicamente, de uma situação em que o aprendente tem oportunidade de reestruturar o texto, introduzindo correções ou reformulações, permitem aferir estratégias relacionadas com a mobilização do conhecimento explícito sobre a língua-alvo de aprendizagem (Myles 2015: 314)[7].

As razões que explicam a assimetria, no que diz respeito ao volume de dados escritos e orais disponíveis, são várias (Bell & Payant 2021; O'Keeffe & McCarthy 2010; Ballier & Martin 2015, entre outros). Basicamente, os *corpora* de produções orais são mais onerosos, uma vez que o tempo necessário para a recolha é substancialmente superior ao permitido pelos *corpora* de produções escritas[8], exigindo ainda muitos recursos humanos e técnicos para as sucessivas fases da sua elaboração (Ballier & Martin 2015: 107). Antes de mais, há que assegurar a qualidade das gravações que dependerá, por um lado, das condições do espaço onde as sessões decorrem e, por outro, dos recursos técnicos de captação e reprodução de som. Depois, há que, assegurando os requisitos técnicos adequados, minorar o seu impacto sobre o desempenho do informante. Finalmente, para que a sua análise seja viável, os *corpora* orais têm de ser transcritos, isto é, tem de ser fornecida uma representação textual da fala. A partir do momento em que a produção oral é, por natureza, multimodal, isto é, a partir do momento em que, na oralidade, a construção do significado resulta não só da interação entre os elementos textuais e prosódicos, mas envolve também, por exemplo, os elementos gestuais, a sua transcrição exige ainda que se façam várias escolhas, em função do nível de detalhe pretendido, e implica, em certa medida, uma atividade interpretativa (Adolphs & Knight 2010: 44; Ballier & Martin 2015: 108-109).

uma dada investigação, ou, pelo contrário, enunciados bastante diferentes entre si, obtidos através de tarefas de produção temática, mas não formalmente condicionadas, por exemplo.

7 A este propósito, afirmam Bell & Payant (2021: 57) que "[u]nder time pressure, typically representative of oral tasks, learners may exhibit performance errors that may not be representative of their complete knowledge about language. In the written medium, with additional time to process ideas and reflect on the language, fewer performance errors may occur. Although both oral and written texts provide a window into learners' interlanguage, unplanned oral data may be more representative of internalized rules compared to planned written data which provides evidence of a learner's system and ability to apply learned rules (when given enough time to plan and revise)."

8 Tipicamente nos *corpora* de produções escritas é possível a recolha simultânea de um grande volume de dados; já nos *corpora* de produções orais, procede-se habitualmente à recolha de dados de forma individualizada ou em pequenos grupos.

3. *Corpora* de produções de aprendentes: recursos do CELGA-ILTEC

É num contexto de desenvolvimento de recursos para a investigação que, no grupo de investigação *Português em Contacto*, do Centro de Estudos de Linguística Geral e Aplicada (CELGA-ILTEC), sediado na Faculdade de Letras da Universidade de Coimbra (FLUC), têm vindo a ser criados e disponibilizados em acesso aberto *corpora* de produções de aprendentes. Note-se que o facto de o ensino de Português a estrangeiros na FLUC ter uma longuíssima tradição (Rio-Torto 2014) tem permitido aos investigadores o contacto com um grande número de potenciais informantes com perfis diversificados.

O projeto de criação de recursos desta natureza, com relevância para investigadores, mas também para os agentes de ensino, iniciou-se com o *Corpus de Produções Escritas de Aprendentes de Português L2* (PEAPL2), coordenado por Cristina Martins (Martins 2013). Esse acervo encontra-se, atualmente, organizado em 3 *subcorpora*:

i. *subcorpus* Português Língua Estrangeira (Martins *et al.*, 2019a);
ii. *subcorpus* Timor (Martins *et al.*, 2019b);
iii. *subcorpus* Guiné-Bissau (Martins *et al.*, 2019c).

No primeiro caso, disponibilizam-se produções textuais de aprendentes de português que frequentavam, à época da recolha, cursos de PL2 na FLUC. Nos outros dois *subcorpora*, reúnem-se produções escritas que, embora tenham sido motivadas de forma idêntica, foram obtidas num contexto completamente diferente (em Timor e na Guiné-Bissau), por falantes para os quais o português é, tipicamente, e em ambos os casos, língua segunda.

A escassez de dados orais, genericamente apontada pela literatura, como já referimos, verifica-se também no que diz respeito ao PL2[9].

Essa constatação motivou o desenvolvimento, no CELGA-ILTEC, do *Corpus Oral de Português L2-Coimbra* (COral-Co), coordenado por Isabel Santos (http://teitok2.iltec.pt/coralco/index.php?action=home), estando igualmente em curso a criação do *Corpus de Interações Orais* (COrIntO) de aprendentes de PL2 (coord. Conceição Carapinha) (Carapinha 2022).

O corpus CorIntO, ainda em fase-piloto, visa a obtenção de uma base de dados constituída por interações verbais orais, protagonizadas por dois ou três participantes – aprendentes adultos de PL2 – que discutem, num determinado intervalo temporal, pontos de vista acerca de um tema sensível. O objetivo é a obtenção de um discurso completo, quase espontâneo, ou seja, de uma atividade comunicativa global, poligerada, de natureza argumentativa, que permita análises de natureza pragmático-discursiva e a descrição das interlínguas dos participantes neste domínio.

9 Na verdade, e tanto quanto é do nosso conhecimento, dispúnhamos, até esse momento, de dois *corpora* com dados de produção oral de aprendentes de PLE: (i) o *Learner Corpus Português L2 – COPLE2* (Centro de Linguística da Universidade de Lisboa (CLUL)), disponível em http://teitok.clul.ul.pt/cople2/, que apresenta os textos orais transcritos e anotados, apesar de os ficheiros áudio se encontrarem inativos; e (ii) o Projeto CAL2 – *Corpus de Aquisição L2* – subcorpus *Produção Oral* (Centro de Linguística da Universidade Nova de Lisboa (CLUNL)), disponível mediante registo prévio em http://cal2.clunl.fcsh.unl.pt/imlogin.php, que apresenta apenas a transcrição das produções orais (constitui, portanto, um *mute spoken corpus* (Ballier & Martin 2015:110)).

No COral-Co, estão disponibilizados os dados relativos a 112 informantes, repartidos por 28 línguas maternas (LM) e 5 níveis de proficiência (de A1 a C1+), estipulados de acordo com o *Quadro Europeu Comum de Referência para as Línguas* (QECRL) (Conselho da Europa, 2001) (cf. Quadro 1). Tal como no *subcorpus* escrito PEAPL2-PLE, todos os informantes eram, à época da recolha, estudantes a frequentar o Curso Anual ou de Férias *de Língua e Cultura Portuguesas para Estrangeiros* ou as unidades curriculares de *Língua Portuguesa Erasmus* da FLUC (Santos *et al.* 2016: 748). A recolha e posterior disponibilização das produções foi autorizada por cada aprendente, que, à semelhança do procedimento seguido na construção dos *corpora* escritos, preencheu uma declaração de consentimento informado.

Nível de proficiência	**#**
A1	22
A2	20
B1	28
B2(+)	35
C1+	7
TOTAL	**112**

Quadro 1. Distribuição dos informantes que integram o COral-Co por nível de proficiência linguística

São igualmente disponibilizados os metadados relevantes, quer os relacionados com os informantes, quer os relacionados com as opções metodológicas seguidas na recolha e na transcrição, devidamente descritas e fundamentadas[10]. O pressuposto é o de que a partilha deste nível de informação permite ao utilizador avaliar a adequação qualitativa e quantitativa do recurso aos seus interesses de investigação (Mackey & Gass 2005: 98; Bell & Payant, 2021: 54).

Depois de validadas e editadas, as gravações foram disponibilizadas, procedendo-se imediatamente a seguir à sua transcrição e anotação através da plataforma TEITOK (*The Tokenized TEI Environment*) (Janssen 2016). Trata-se, portanto, de um verdadeiro *corpus* oral, já que se apresenta o sinal acústico e a respetiva transcrição, e não de um *spoken-based corpora* (tipo de acervo em que se fornecem as transcrições, mas não se disponibiliza o áudio correspondente – *mute spoken data*) (Ballier & Martin 2015). Pretendeu-se, desta forma, corresponder às necessidades e diferentes interesses dos investigadores, possibilitando assim, ao contrário dos *mute* spoken *corpora*, estudos no domínio das interfonologias.

Os dados do COral-Co foram elicitados mediante a apresentação aos informantes de um conjunto de 7 tarefas, de produção oral e de leitura oral, que criam contextos de produção progressivamente mais controlados (cf. Quadro 2). Como se depreende da leitura do Quadro 2, o número de ficheiros áudio disponibilizados por tarefa não é o mesmo, dado que, em alguns casos, não foi possível obter ou disponibilizar todas as produções orais de um dado informante.

10 Cf. http://teitok2.iltec.pt/coralco/index.php?action=descricao.

Natureza dos dados	Tarefa	N.º de ficheiros
Produção oral	1. Entrevista semiestruturada	106
	2. Elicitação de atos ilocutórios (a cada informante corresponde um número diverso de ficheiros)	104
	3. Construção de um texto narrativo a partir de uma sequência de imagens	109
	4. Nomeação de realidades apresentadas em suporte pictórico	109
Leitura oral	5. Leitura de texto	107
	6. Leitura de listas de palavras (morfofonologicamente aparentadas)	104
	7. Leitura de listas de palavras (pares mínimos)	105

Quadro 2. Distribuição do número de ficheiros por tarefa disponibilizados no COral-Co

Na apresentação dos materiais, o COral-Co tem alinhados áudio e excertos da fala (é, por isso, um *speaking corpus*), mas em duas das tarefas de leitura, nomeadamente as tarefas de leitura de palavras isoladas, a transcrição surge alinhada com unidades linguísticas mais pequenas, isto é, com as palavras, aproximando-se, nesse caso, de um "corpus fonético" (Ballier & Martin 2015: 110).

Replicando a forma de acesso aos dados já usada nos *corpora* escritos, o material compilado no COral-Co permite ao utilizador efetuar diferentes pesquisas: por Tarefa; por Nível de Proficiência e por LM. No que se refere às produções transcritas, é possível a pesquisa, por exemplo, por lema, por categoria gramatical, ou por terminação (cf. Abrantes 2019).

4. O *corpus* POPL2

4.1. Descrição

O *corpus* POPL2 (*Produção Oral em Provas de Português L2)* visa disponibilizar à comunidade científica, e em regime de acesso aberto, um acervo de dados orais produzidos por aprendentes tardios de PL2 em contexto de provas de avaliação final, previstas nos planos das unidades curriculares de *Comunicação Oral, Comunicação Oral e Escrita* e *Laboratório*, do *Curso Anual de Língua e Cultura Portuguesas para Estrangeiros* (CALCPE)[11] da FLUC. Neste projeto, coordenado

11 O *Curso Anual de Língua e Cultura Portuguesas para Estrangeiros* (CALCPE) está organizado em cinco níveis de proficiência linguística (A1, A2, B1, B2 e C1), em conformidade com os descritores propostos no QECRL (Conselho da Europa, 2001), sendo a formação semestral. A unidade curricular *Comunicação Oral* está disponível entre os níveis A1 e B1, com uma carga horária associada de 3h/semanais. A partir do nível B2 até ao nível C1+ os estudantes frequentam a unidade curricular *Comunicação Oral e Escrita,* com uma carga horária associada de 4h/semanais. Por sua vez, a unidade curricular de *Laboratório* integra os planos de estudos do nível A1 até B2+ (cf. Rio-Torto 2014: 35-36). No fim de cada semestre, os docentes destas unidades curriculares ministram, por nível de proficiência linguística, uma prova de avaliação oral.

por Tânia Ferreira, colaboram, como consultores científicos, investigadores do CELGA-ILTEC e, na qualidade de avaliadores, docentes do CALCPE, que ministram as provas orais.

Na linha do que se fez para outros *corpora*, pretende-se obter um acervo variado de dados no que respeita ao perfil dos informantes e à tipologia de tarefas de produção oral, devidamente articuladas com os programas e as práticas pedagógicas das unidades curriculares frequentadas pelos estudantes. Deste modo, e à semelhança do COral-Co, pretende-se disponibilizar os ficheiros áudio, os metadados, quer os relativos aos aprendentes quer os relativos ao processo de recolha, e a respetiva transcrição anotada, com recurso à plataforma TEITOK (Janssen 2016).

4.1.1. Recolha de dados

A compilação de dados de natureza oral produzidos por aprendentes apresenta, como vimos na secção 3., diversos constrangimentos que dizem respeito às características dos textos orais, às especificidades do próprio processo de recolha e à complexidade do processo de transcrição (Bell & Payant 2021: 57).

Como referido atrás, o contexto de gravação e de avaliação pode condicionar o grau de espontaneidade das produções orais dos aprendentes. Ou seja, as produções obtidas nestas situações apresentam marcas específicas do contexto comunicativo em que ocorrem, sendo previsivelmente diferentes de outras ocorridas em situações comunicativas mais informais. No *corpus* POPL2 em particular, o facto de os dados orais serem produzidos em momentos de avaliação, temporalmente limitados, poderá condicionar o desempenho linguístico dos informantes; em contrapartida, será estimulada a automonitorização da produção. Com efeito, é expectável que o aprendente, por saber que está a ser avaliado e por ter receio de falhar, evidencie um maior controlo da sua produção oral; esse controlo manifestar-se-á, por exemplo, em estratégias de evitação, isto é, na inibição do uso de estruturas da língua-alvo com as quais o aprendente não se sente seguro, limitando, dessa forma, a ocorrência de desvios. No entanto, dado que estes textos são recolhidos em situações semelhantes, isto é, de avaliação, e resultam, igualmente, de tipologias de atividades idênticas, obter-se-á, por esta via, um acervo de dados comparáveis e utilizáveis em diferentes trabalhos de investigação.

As provas orais são realizadas no final de cada semestre letivo pelos docentes das unidades curriculares de *Comunicação Oral*, *Comunicação Oral e Escrita* e *Laboratório*. Dos níveis A1 a B2+, a prova é ministrada conjuntamente pelos docentes de *Laboratório* e de *Comunicação Oral* ou de *Comunicação Oral e Escrita*, no caso das turmas de B2 e B2+. Nas turmas de C1 e C1+, uma vez que o plano curricular não contempla a disciplina de *Laboratório* (cf. Rio-Torto, 2014, p. 36), a prova é ministrada pelo docente de *Comunicação Oral e Escrita*, podendo ainda estar presente o docente responsável pela disciplina de *Estruturas da Língua Portuguesa* (C1 / C1+) que, assim, pode colaborar na condução da prova.

Tipicamente, os aprendentes realizam as provas em pares, mas, por vezes, em função do número de alunos por turma, há provas em que participa um único aluno ou em que estão presentes três ou mais alunos em simultâneo. Nestas situações, a presença de múltiplos falantes poderá condicionar o processo da transcrição; em função da qualidade final da gravação, assim se procederá ou não à respetiva transcrição. Com vista a assegurar a correta identifi-

cação dos intervenientes em prova, é solicitado aos aprendentes que, no início, façam uma breve apresentação, indicando explicitamente o número que lhes foi atribuído na ficha do informante, a(s) língua(s)-materna(s) e o nível de proficiência em português da turma em que se encontravam integrados. Estas informações servem apenas para auxiliar o transcritor no processo de identificação das vozes dos intervenientes e não farão, por isso, parte do ficheiro áudio disponibilizado.

No que concerne ao local de gravação, as provas decorrem em salas de aula comuns do edifício da FLUC que, tipicamente, não estão concebidas para esse tipo de utilização. No entanto, é possível recorrer a dispositivos de captação de som que garantem a qualidade adequada do ficheiro áudio para posterior transcrição (Adolphs & Carter 2013: 8; Bell & Payant 2021: 60; Santos 2020: 65). Além disso, de forma a minimizar o impacto da gravação para o normal funcionamento da prova oral, quer para os informantes quer para os docentes, é necessário que o dispositivo de gravação seja discreto e de fácil utilização. Assim, optou-se por gravar as provas com recurso a um dispositivo telefónico móvel, já que, dadas as suas características, este equipamento permite obter ficheiros áudio de qualidade, além de ser prático e de uso intuitivo. Já foram, aliás, realizadas gravações-piloto (cf. secção 4.2.) que confirmaram a adequação deste equipamento.

4.1.2. Tipologia de atividades

As atividades que motivam as produções orais em prova podem ser muito diversas e encontram-se devidamente articuladas com os programas das respetivas unidades curriculares (cf. secção 4.1.). Em linhas gerais, neste contexto pode verificar-se o recurso a dois grandes tipos de atividades: (i) de interação; e (ii) não-interativas (cf. Figura 1).

As atividades de *entrevista*, *discussão informal* e *role-play* integram o conjunto de atividades de interação por pressuporem a intervenção de, pelo menos, dois interlocutores.

A *entrevista* é uma atividade de interação em que o(s) docente(s) coloca(m) questões ao(s) aprendente(s) sobre diferentes tópicos (por exemplo, apresentação pessoal, atividades de tempos livres, etc.). Os temas abordados na entrevista serão os adequados ao nível de proficiência dos informantes e, dada a amplitude de temas motivadores da produção, esta é uma atividade frequentemente utilizada em contextos de avaliação oral.

As atividades de *discussão informal* correspondem a trocas de pontos de vista entre os aprendentes a partir de um estímulo previamente apresentado pelo docente que pode ser visual (por exemplo, uma imagem) ou escrito[12] (por exemplo, um *slogan*). Espera-se que, neste contexto, o docente intervenha apenas na fase de apresentação do estímulo, de forma a não condicionar o diálogo que se estabelece entre os aprendentes.

O *role-play* corresponde à simulação de situações reais de comunicação (por exemplo, interação entre vendedor e comprador, entre médico e paciente, etc.). Na fase de preparação da atividade, o docente apresenta o contexto situacional que motiva o diálogo e indica o papel que cada aprendente vai assumir. As situações representadas abrangem diferentes graus de formalidade

12 O recurso a estímulos escritos deverá ser ponderado e atender ao nível de proficiência linguística dos aprendentes. Por exemplo, deve evitar-se, em níveis mais baixos, o recurso a textos longos, de forma a que a menor competência do domínio escrito não comprometa o desempenho oral.

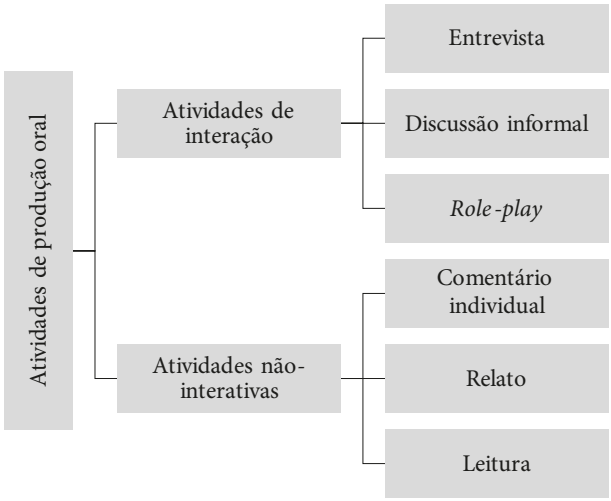

Figura 1. Tipologias de atividades de produção oral aplicáveis em contextos de provas orais

e de complexidade linguística e devem ser, por conseguinte, ajustadas aos níveis de proficiência em que se aplicam.

Em contrapartida, atividades de *comentário individual, relato* e *leitura* constituem exemplos de atividades não-interativas, visto que apenas intervém um único falante. O *comentário individual* corresponde à apresentação de um ponto de vista a propósito de um estímulo visual ou escrito previamente apresentado. O *relato* compreende a descrição, a partir de um estímulo pictórico, de objetos, de pessoas, de lugares e/ou a narração de sequências de imagens. Já a *leitura* pode, também, constituir um momento da prova oral, embora esta não seja uma atividade frequentemente utilizada em provas orais. Com efeito, apesar de permitirem "avaliar aspetos criteriosamente selecionados no domínio das competências fonológica e ortoépica, segmental ou prosódica" (Santos *et al.* 2016: 752), os exercícios de leitura oral não possibilitam aferir diretamente a competência comunicativa dos aprendentes.

Numa mesma prova, pode verificar-se a ocorrência de produções orais que resultam da aplicação de tarefas de natureza interativa e não-interativa. Assim, toda a informação relativa à prova é devidamente documentada e associada aos respetivos ficheiros na secção dos metadados (cf. 4.1.3.).

4.1.3. Metadados

As opções metodológicas adotadas neste projeto visam a articulação deste acervo com outros *corpora* de aprendentes já existentes no CELGA-ILTEC (cf. secção 3.), nomeadamente no que diz respeito ao processo de recolha e à disponibilização das informações dos aprendentes, isto é, dos metadados relativos aos sujeitos cujas produções se recolheram.

Para caracterizar o perfil do informante, utilizou-se o modelo do inquérito concebido no âmbito do *corpus* PEAPL2 e adaptado no *corpus* COral-Co (Santos *et al.* 2016). Esse inquérito surge organizado por secções que permitem recolher:

i. dados biográficos, sociolinguísticos e curriculares dos aprendentes;
ii. informação relativa à experiência de contacto com a língua portuguesa fora da sala de aula; e
iii. identificação da variedade (ou variedades) do português aprendida(s)[13].

Antes do momento da gravação da prova, o aprendente preenche, por escrito, a ficha do informante e também assina, nessa altura, uma declaração de consentimento informado em que autoriza, no âmbito do projeto, a gravação e a disponibilização dos dados. A cada ficha está associado um número que servirá, posteriormente, para aceder aos dados do informante. Também se disponibiliza, entre os docentes / avaliadores, uma ficha com instruções gerais relativas ao processo de gravação da prova, com a declaração de consentimento informado que também deverá ser assinada.

Tratando-se de um *corpus* de produções orais recolhidas em contextos de provas de avaliação final realizadas na FLUC, todos os informantes são estudantes estrangeiros, adultos, em contexto de imersão linguística, tendo já frequentado, pelo menos, um semestre de aulas de língua portuguesa. Não obstante a sua diversidade há, assim, alguma homogeneidade no perfil dos informantes.

Além dos dados relativos aos aprendentes, também é feita a documentação e codificação de tudo o que diz respeito aos ficheiros áudio[14]. Assim, os metadados das gravações integram as seguintes informações: (i) data da recolha (dia.mês.ano); (ii) tempo total da gravação transcrita (hh:mm:ss); (iii) tipo(s) de atividade(s) realizada(s) (atividade(s) de interação e/ou atividade(s) não-interativa(s)) (cf. secção 4.1.2.); e (iv) o(s) códigos(s) numérico(s) do(s) informante(s) que participa(m) na prova.

4.1.4. Transcrição

Como já referido, no POPL2 prevê-se disponibilizar o ficheiro áudio e a respetiva transcrição. Atualmente, existem diferentes programas informáticos especificamente vocacionados para a transcrição de segmentos orais, bem como listas de convenções de transcrição que permitem a uniformização dos dados transcritos (Adolphs & Knight 2010; Ballier & Martin 2015; Bell & Payant 2021). Apesar das várias opções disponíveis, transcrever segmentos orais constitui, indubitavelmente, "[o]ne of the biggest challenges in corpus linguistic research" (Adolphs & Knight 2010: 44). A transcrição de segmentos orais produzidos por aprendentes, falantes não nativos, é

13 Este dado figura-se como especialmente relevante para os trabalhos de investigação com dados orais "dada a existência de características fónicas diferenciadoras no universo de língua portuguesa" (Santos *et al.* 2016: 748).

14 "Metadata (background information) refers to all data aside from language samples (the corpus) collected from participants (e.g., learner's age, gender, proficiency, level, language background). To increase the rigor, transparency, and usability of learner corpora, collecting and publishing substantial metadata (including how the metadata were collected) is essential […]." (Bell & Payant 2021: 54).

particularmente complexa, sobretudo no que concerne às opções a tomar para a representação, na escrita, de segmentos que resultam da pronúncia não convergente com a língua-alvo (Guilquin 2015: 20; Bell & Payant 2021: 61)[15].

Existem diferentes convenções de transcrição que se distinguem quanto ao grau de detalhe que permitem no registo de traços característicos do discurso oral (sobreposições, hesitações) e de aspetos fonológicos, tanto segmentais como prosódicos (Adolphs & Knight 2010). Além da representação textual da fala, com recurso aos símbolos gráficos e convenções oficiais (uma das modalidades mais comuns de transcrição), traços da pronúncia podem ser transcritos com recurso ao Alfabeto Fonético Internacional e com o apoio de programas informáticos (Adolphs & Knight 2010: 44; Ballier & Martin 2015: 118) como, por exemplo, o PRAAT (Boersma & Van Heuven 2001; Boersma 2014; Brinckmann 2014; Boersma & Weenink 2022), o ELAN (https://archive.mpi.nl/tla/elan) ou o EXMARaLDA (Schmidt & Wörner 2009).

Com vista à compatibilização dos dados do POPL2 com os disponibilizados no COral-Co (http://teitok2.iltec.pt/coralco/index.php?action=home), pretende adotar-se uma parte substancial das convenções de transcrição utilizadas nesse acervo (cf. Santos *et al.* 2016)[16]. Portanto, pondera-se transcrever, com recurso aos símbolos e normas da ortografia oficial, as alterações fonológicas evidentes que podem resultar do desconhecimento, por parte do informante, (i) da estrutura morfológica das palavras-alvo e (ii) da estrutura fonológica de constituintes morfológicos, lexicais, flexionais ou derivacionais, que conduzem a formas não convergentes com a língua-alvo, mas, ainda assim, identificáveis. Reconhecemos, porém, que o recurso aos símbolos ortográficos convencionais para representação de traços claramente desviantes que parecem relevantes para a descrição da interfonologia complexifica o processo de transcrição, correndo-se o risco da proliferação de situações que resultam da interpretação do transcritor[17] (Mendes *et al.* 2016: 3209). Ainda assim, e à semelhança do COral-Co, os dados transcritos serão sujeitos, posteriormente, a um processo de lematização e a pesquisa por lema permitirá ao investigador ter acesso à forma-alvo (cf. Abrantes 2019). Por outro lado, a disponibilização do áudio permitirá ao investigador aceder às formas efetivamente produzidas pelos aprendentes.

4.2. Gravações-piloto

Entre os dias 18 e 19 de janeiro de 2022, realizaram-se gravações das provas orais a duas turmas dos níveis B2 e B1, respetivamente. Além das gravações, foram recolhidas as informações relativas aos perfis sociolinguísticos dos informantes e as respetivas autorizações para a disponibilização da gravação no âmbito do projeto POPL2.

No total, obtiveram-se 21 ficheiros áudio que resultam da gravação de provas aplicadas a 34 aprendentes. Os informantes são de ambos os sexos (53% do sexo feminino e 47% do sexo mas-

15 Estas formas não convergentes podem decorrer do desconhecimento, quer da estrutura fonológica, quer da estrutura morfológica das palavras da língua-alvo usadas pelo aprendente.

16 O documento com as convenções de transcrição seguidas no COral-Co está disponível para consulta na página *web* do projeto, na secção "Transcrição".

17 Como afirmam Mendes *et al.* (2016, p. 3209): "faithful transcription would lead to a multiplication of different writing options and to many inconsistencies depending on the transcriber".

LM	Nível QECRL	
	B1	B2
Árabe	1	1
Chinês	8	5
Coreano	2	-
Espanhol	1	-
Francês	-	1
Inglês	2	2
Japonês	3	3
Mandarim	1	-
Tailandês	1	-
Tétum	2	-
Turco	-	1
TOTAL	**21**	**13**

Quadro 3. Distribuição do número de participantes nas gravações-piloto do POPL2 por LM e nível de proficiência da turma frequentada

culino) e têm uma média de idades de 24 anos (distribuídos entre os 20 e os 44 anos). O tempo médio da duração das provas foi de 12 minutos. Apresenta-se, no Quadro 3, a caracterização da amostra, com a distribuição dos aprendentes por níveis de proficiência e respetiva LM.

As informações relativas às LM foram dadas pelos informantes antes da prova, mediante o preenchimento do inquérito disponibilizado. Nas provas ministradas à turma do nível B1, foram, em geral, aplicadas atividades de interação (*entrevista* e *discussão informal)* e atividades não-interativas (*comentário individual* e *relato*). Na turma de B2, as produções orais resultaram, sobretudo, do recurso à *entrevista*.

Em média, participaram em cada prova dois informantes em simultâneo. Na turma de B2, houve uma prova em que participaram três aprendentes e na turma de B1 houve uma prova em que participou um único aprendente.

Estas gravações permitiram validar as opções metodológicas estipuladas pela equipa na fase de preparação do *corpus.* Assim, com vista à obtenção de um acervo robusto de produções orais resultantes de tarefas aplicadas em provas de avaliação, prosseguir-se-á o processo de recolha e posterior disponibilização do POPL2.

A realização das gravações-piloto permitiu também confirmar a qualidade da gravação feita com recurso a um dispositivo telefónico móvel e a adequação das condições em que decorre a recolha dos dados.

5. Considerações finais

A elaboração de *corpora* de produções de aprendentes de L2 decorre da convicção de que o conhecimento do processo de aquisição / aprendizagem e das interlínguas que os aprendentes vão construindo ao longo desse processo não se fará adequadamente sem a observação do seu comportamento linguístico. Neste âmbito, considera-se que, não sendo exequível a obtenção de dados longitudinais, a construção de amostras *cross-sectional* (constituídas por produções de aprendentes em diferentes níveis de proficiência) permitirá observar em tempo aparente (Labov 1972) os processos de reestruturação. Por outro lado, a aquisição/aprendizagem de uma L2 desenvolve-se, habitualmente, no domínio da escrita e no domínio da oralidade; assim, tanto os dados de produção escrita como os dados de produção oral se afiguram essenciais para o trabalho de descrição das interlínguas. Além disso, por via do conhecimento do desempenho (convergente ou não convergente) do aprendente, poder-se-ão definir estratégias pedagógicas adequadas e poder-se-ão conceber os materiais instrucionais mais eficazes. É, no entanto, mais complexa a organização de *corpora* de produções orais e tal facto justifica a menor abundância e a menor dimensão deste tipo de recurso.

Não obstante as dificuldades e as limitações que tipicamente se verificam no processo de recolha de produções orais, é indubitável a sua relevância para a investigação no âmbito do ensino e aprendizagem de uma L2. Neste contexto, o *corpus* POPL2 visa, assim, reforçar os acervos de produções de aprendentes de PL2, pelo que contribuirá para diminuir a assimetria no que diz respeito ao volume de dados de produções escritas e de produções orais. Nele integram-se dados recolhidos em provas de avaliação, aplicadas a aprendentes tardios que frequentam, na FLUC, cursos de português em diferentes níveis de proficiência linguística. As produções orais são motivadas por diferentes tipologias de atividades, quer de natureza interativa quer de natureza não interativa, permitindo, assim, a constituição um acervo de produções orais diversificadas, que pode ser continuamente ampliado. Além disso, o facto de as produções orais resultarem de tarefas ministradas em situações idênticas, isto é, de avaliação, permitirá a constituição de uma base de dados de natureza similar, que, por sua vez, complementará as que resultam de produções escritas.

As opções metodológicas seguidas neste projeto refletem as diferentes especificidades relativas à elaboração de bases de dados de produções de aprendentes, nomeadamente no que respeita aos procedimentos de recolha, à compilação dos metadados, e ao processo de transcrição e disponibilização dos dados transcritos. Espera-se, por fim, que este seja um recurso fundamental para a investigação no âmbito do PL2, estando igualmente disponível para a fundamentação de materiais e práticas em contexto instrucional.

Referências bibliográficas

Abrantes, C. (2019). *Investigação em corpora informatizados de produções orais e escritas de aprendentes de PLNM: FAQ e orientações para a exploração de valências*. Universidade de Coimbra, Projeto de Mestrado.

Adolphs, S.; & Carter, R. (2013). *Spoken Corpus Linguistics. From Monomodal to Multimodal.* New York / London: Routledge.

Adolphs, S.; & Knight, D. (2010). Building a spoken corpus. What are the basics? In A. O'Keeffe, & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 38–52). New York / London: Routledge.

Ballier, N.; & Martin, P. (2015). Speech annotation of learner corpora. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 107–134). Cambridge: Cambridge University Press.

Bell, P.; & Payant, C. (2021). Designing Learner Corpora: Collection, Transcription, and Annotation. In N. Tracy-Ventura, & M. Paquot (Eds.), *The Routledge Handbook of Second Language Acquisition and Corpora* (pp. 53–67). New York / London: Routledge.

Boersma, P. (2014). The use of PRAAT in corpus research. In J. Durand, U. Gut, & G. Kristoffersen (Eds.), *The Oxford Handbook of Corpus Phonology* (pp. 342–360). Oxford: Oxford Academic. <https://doi.org/10.1093/oxfordhb/9780199571932.001.0001>

Boersma, P.; & Van Heuven, V. (2001). Speak and unSpeak with PRAAT. *Glot International*, *5*, 9/10, 341–347.

Boersma, P.; & Weenink, D. (2022). PRAAT: doing phonetics by computer [Computer program]. Version 6.2.14. <http://www.praat.org/>

Brinckmann, C. (2014). PRAAT scripting. In J. Durand, U. Gut, & G. Kristoffersen (Eds), *The Oxford Handbook of Corpus Phonology* (pp. 361-379). Oxford: Oxford Academic. <https://doi.org/10.1093/oxfordhb/9780199571932.001.0001>

Carapinha, C. (2022*).* Para a construção de um *corpus* de interações orais em Português Língua Não Materna (PLNM) – algumas reflexões. *Linguística, Revista de Estudos Linguísticos da Universidade do Porto* (vol. 17).

Conselho da Europa (2001). *Quadro Europeu Comum de Referência para as Línguas*. Edições Asa.

ELAN (Version 6.4) [Computer software]. (2022). Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. <https://archive.mpi.nl/tla/elan>

Félix-Brasdefer, J. C. (2007). Natural speech vs. elicited data: A comparison of natural and role play requests in Mexican Spanish. *Spanish in Context,* 4, 2, 159–185.

Flores, C. M. M. (2013). Português Língua Não Materna. Discutindo conceitos de uma perspetiva linguística. In R. Bizarro, M. Moreira, & C. Flores (Orgs.), *Português língua não materna: investigação e ensino* (pp. 35–46). Lisboa: Lidel. <https://repositorium.sdum.uminho.pt/bitstream/1822/23009/1/C.Flores_PLNM%20Discutindo%20conceitos%20de%20uma%20perseptiva%20lingu%C3%ADstica.pdf>

Gilquin, G. (2015). From design to collection of learner corpora. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 9–34). Cambridge: Cambridge University Press.

Granger, S. (2002). A bird's eye view of learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* (pp. 3–33). Amsterdam & Philadelphia: Benjamins.

————. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching: a critical evaluation. In K. Aijmer (Ed.), *Corpora and Language Teaching* (pp. 13–32). Amsterdam & Philadelphia: Benjamins. <https://doi.org/10.1075/scl.33.04gra>

Granger, S.; Giquin, G.; & Meunier, F. (2015). Introduction: learner corpus research – past, present and future. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 1–5). Cambridge: Cambridge University Press.

Janssen, M. (2016). TEITOK: Text-Faithful Annotated Corpora. In N. Calzolari *et al.* (Eds.), *LREC 2016. Tenth International Conference on Language Resources and Evaluation. May 23-28, 2016, Portorož, Slovenia*. <http://www.lrec-conf.org/proceedings/lrec2016/pdf/651_Paper.pdf>

Labov, W. (1972). *Sociolinguistic Patterns.* University of Pennsylvania Press.

Mackey, A.; & Gass, S. M. (2005). *Second Language Research. Methodology and Design*. London: Lawrence Erlbaum Associates, Publishers.

Madeira, A. (2017). Aquisição de língua não materna. In M. J. Freitas, & A. L. Santos (Eds.), *Aquisição de língua materna e não materna: questões gerais e dados do português* (pp. 306–330). Berlin: Language Science Press.

Martins, C. (2013). O Corpus de Produções Escritas de Aprendentes de PL2 (PEAPL2/CELGA). Caracterização e desenvolvimento de uma infra-estrutura de investigação. In R. Bizarro, M. A. Moreira, & C. Flores (Eds.), *Português Língua Não Materna: Investigação e Ensino* (pp. 69–80). Lisboa: Lidel.

Martins, C.; Ferreira, T.; Sitoe, M.; Abrantes, C.; Janssen, M.; Fernandes, A.; Silva, A.; Lopes, I.; Pereira, I.; & Santos, J. (2019a). *Corpus de Produções Escritas de Aprendentes de PL2 (PEAPL2): Subcorpus Português Língua Estrangeira*. Coimbra: CELGA-ILTEC.

Martins, C.; Pereira, I.; Melo, D.; Shanna, X.; Ximenes, M.; & Janssen, M. (2019b). *Corpus de Produções Escritas de Aprendentes de PL2 (PEAPL2): Subcorpus Timor.* Coimbra: CELGA-ILTEC.

Martins, C.; Santos, I.; Marques, M.; Abrantes, C.; Neves, A.; & Janssen, M. (2019c). *Corpus de Produções Escritas de Aprendentes de PL2 (PEAPL2): Subcorpus Guiné-Bissau.* Coimbra: CELGA-ILTEC.

Mauranen, A. (2004). Spoken corpus for an ordinary learner. In J. McH. Sinclair (Ed.), *How to Use Corpora in Language Teaching* (pp. 89–105). Amsterdam & Philadelphia: John Benjamins Publishing Company.

Mendes, A.; Antunes, S.; Janssen, M.; & Gonçalves, A. (2016). The COPLE2 Corpus: A Learner Corpus for Portuguese. In N. Calzolari *et al.* (Eds.), *LREC 2016. Tenth International Conference on Language Resources and Evaluation. May 23-28, 2016, Portorož, Slovenia.* <http://www.lrec-conf.org/proceedings/lrec2016/pdf/439_Paper.pdf>

Meunier, F. (2021). Introduction to Learner Corpus Research. In N. Tracy-Ventura, & M. Paquot (Eds.), *The Routledge Handbook of Second Language Acquisition and Corpora* (pp. 23–36). New York: Routledge.

Myles, F. (2015). Second language acquisition theory and learner corpus research. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of Learner Corpus Research* (pp. 309–331). Cambridge: Cambridge University Press.

Rio-Torto, G. (2014). Passado e presente dos Cursos de Férias. Da edição de 1924-1925 à de 2014. In G. Rio-Torto (Coord.), *90 anos de ensino de língua e cultura portuguesas para estrangeiros na Faculdade de Letras da Universidade de Coimbra* (pp. 13–38). Coimbra: Imprensa da Universidade de Coimbra.

Santos, G. (2020). Designing and building SCoPE: A spoken corpus of Brazilian Portuguese and L2-English. *Research in Corpus Linguistics*, 8, 49–64.

Santos, I. A.; Pereira, I.; Martins C.; Lopes, A.C.M.; Carapinha, C.; & Silva, A. (2016). Corp-Oral: PL2 – Um novo recurso para o estudo do português língua não materna. In A. Moreno, F. Silva, & J. Veloso (Eds), *Textos Selecionados do XXX Encontro Nacional da Associação Portuguesa de Linguística* (pp. 103–112). Lisboa: Associação Portuguesa de Linguística.

Schmidt, T.; & Wörner, K. (2009). EXMARaLDA–Creating, analysing and sharing spoken language corpora for pragmatic research. *Pragmatics*, 19, 4, 565–582.

Tracy-Ventura, N.; & Myles, F. (2015). The importance of task variability in the design of learner corpora for SLA research. *International Journal of Learner Corpus Research*, 1, 1, 58–95.