DOI: https://doi.org//10.21747/16466195/ling17a1

Para a construção de um *corpus* de interações orais em Português Língua Não Materna (PLNM) – algumas reflexões

Conceição Carapinha
mccarapinha@fl.uc.pt
Centro de Estudos de Linguística Geral e Aplicada (CELGA-ILTEC)
Universidade de Coimbra

ABSTRACT.

As authentic data sets, different types of corpora increasingly support theoretical and applied linguistic studies. In the same vein, corpora may also constitute an appropriate working basis for second language acquisition and foreign language teaching, especially with regard to the uses of language. However, both Pragmatics, as a discipline focused on the analysis of meaning in context and the area of second language acquisition (SLA) have been quite reticent when it comes to using corpora, whether written or oral. In this text, we present a characterization of spoken learner corpora and advocate their relevance and usefulness to the teaching and learning of Portuguese as a Foreign Language, namely in what concerns Interlanguage Pragmatics. In the same line of thinking, the PL2 oral interaction corpus project presented in this text, with the challenges faced by its implementation, and the possibilities for future research that it will allow, aims at contributing to a better understanding of the scientific and pedagogical advantages of a database containing texts of a more interactive nature.

KEYWORDS.

Corpora; *spoken learner corpora*; acquisition; oral interaction; Portuguese as a Foreign Language.

RESUMO.

Enquanto acervos de dados autênticos, os diversos tipos de *corpora* sustentam, cada vez mais, estudos de natureza teórica e aplicada, podendo também constituir uma boa base de trabalho no âmbito do ensino/aprendizagem de uma L2, sobretudo no que tange ao domínio dos usos. No entanto, quer a Pragmática, enquanto disciplina focada na análise dos usos linguísticos contextualizados, quer a área de aquisição de línguas segundas (*SLA*) se têm mostrado bastante reticentes no que toca ao recurso a *corpora*, sejam eles escritos ou orais. Neste texto, apresenta-se uma caracterização dos *spoken learner corpora* e advoga-se a sua pertinência e a sua utilidade no processo de ensino/aprendizagem do PLNM, nomeadamente no que à área da Pragmática da Interlíngua diz respeito. Nessa linha, o projeto de constituição de um *corpus* de interações orais, em PLNM, que aqui se apresenta – descrevendo não apenas os desafios que se colocam à sua implementação, mas também as possibilidades de pesquisa futura que, uma vez constituído e disponibilizado, ele permitirá – visa contribuir com uma

reflexão sobre as valências científicas e pedagógicas de uma base de dados contendo textos de natureza mais interativa.

PALAVRAS-CHAVE.
Corpora; *spoken learner corpora*; aquisição; interação oral; Português Língua Não Materna (PLNM).

0. Introdução

Num momento em que se exige ao investigador a pesquisa a partir de dados autênticos, um *corpus*, enquanto "conjunto de textos escritos (ou excertos de textos) ou de transcrições de registos orais, tipicamente em formato electrónico" (Nascimento 2002), pode ser explorado em áreas tão diferentes como o desenho de materiais instrucionais, a construção de um dicionário ou a investigação fundamental, possibilitando, portanto, investigação quer de cunho mais empírico quer de caráter mais teórico.

É verdade que são muito diferentes os acervos de dados a que este rótulo é atribuído, quer relativamente ao número de *tokens,* quer ao tipo de dados, quer ainda à sua organização interna, por isso, e para que um conjunto de dados seja considerado um *corpus*, é necessário que esses dados sejam representativos do objeto que se pretende analisar – seja ele uma língua, uma variedade linguística, um género textual, um tipo de falante ou qualquer outro fenómeno –, que haja critérios consistentes no que toca à sua recolha e organização e ainda que esses dados respondam a um determinado objetivo de investigação.

Independentemente das questões relacionadas com os diferentes tipos de *corpora*, com a sua constituição, dimensão, objetivos e acessibilidade, são inúmeras as vantagens da sua utilização. Estes acervos traduzem usos autênticos da língua em apreço, quer na sua modalidade escrita quer na sua modalidade oral e o facto de serem informatizados aumenta exponencialmente as suas valências, permitindo pesquisas muito rápidas e confiáveis em largas quantidades de dados, facilitando o reconhecimento de frequências de uso, a identificação de padrões (que escapariam a uma análise mais casuística) e fenómenos de coocorrência em determinados contextos ou géneros.

Flowerdew (2009) chama até a atenção para uma outra mais-valia dos

corpora, quiçá mais invisível; com efeito, as pesquisas baseadas nestas bases de dados permitem evidenciar muitas das incongruências que persistem nas atuais descrições da língua, pois muitos dos fenómenos atestados em gramáticas e outro tipo de textos não encontram suporte nos *corpora*, ao mesmo tempo que estes revelam usos que não estão ainda descritos.

Não negligenciando a relevância teórica da discussão acerca do estatuto da Linguística de Corpus (LC), entendida ora como domínio de pesquisa, ora como metodologia de análise, podemos afirmar que esta área permite estudar muitos aspetos da pesquisa linguística a partir de uma abordagem empírica (McEnery & Wilson 2001). Neste sentido, os *corpora* assumem indubitável relevância, não só porque eles possibilitam uma descrição da língua centrada em usos atestados, como também porque permitem trabalhar variados fenómenos de diferentes planos de descrição linguística a partir de quadros teóricos diversos. Uma terceira e incontornável vantagem diz respeito à aplicabilidade dos *corpora* a áreas de investigação conexas, como a do ensino de línguas estrangeiras, domínio em que os *corpora* constituem também um recurso precioso.

Neste estudo, apresentar-se-á precisamente uma reflexão sobre a forma como a constituição de um *corpus* de interações orais de aprendentes de português língua não materna (PLNM), que se integra no âmbito, mais lato, dos chamados *corpora* orais de aprendentes (*spoken learner corpora*), pode vir a coadjuvar a investigação no âmbito da pragmática do PLNM, elencando, por um lado, os desafios que se colocam à sua constituição e, por outro, as possibilidades de análise pragmática abertas por um acervo deste tipo. Na secção 1., serão abordadas, ainda que de forma necessariamente breve, as relações entre a LC e a Pragmática, por um lado, e a área de investigação em Aquisição de Línguas Segundas (doravante *SLA*[1]), por outro. Na secção 2., caracterizar-se-ão os *corpora* orais de aprendentes (*spoken learner corpora*). Partindo da experiência obtida na constituição de um *corpus*-piloto, na secção 3., analisar-se-ão alguns problemas que podem dificultar a constituição de um *corpus* deste tipo, bem como algumas das vantagens deste inovador recurso metodológico para a investigação e para a pedagogia, sobretudo na área dos usos do PLNM.

1 Adota-se a sigla em inglês por ser de mais fácil identificação.

1. Linguística de Corpus e outras áreas de investigação

1.1. Linguística de Corpus e Pragmática

Centrada na análise de usos linguísticos contextualizados, a Pragmática parece ser a área de investigação preferencial para trabalhar com *corpora*. Na verdade, "[a] pragmatic perspective on language implies that the focus is on the use of lexical elements or grammatical structures in their linguistic, social and cultural context" (Aijmer 2020: 29). Ora, ao permitir identificar estruturas e usos no seu contexto de ocorrência, um *corpus* parece estar, de facto, ao serviço das pesquisas pragmáticas.

Porém, esta óbvia convergência nem sempre foi explorada e, durante muito tempo, o trabalho da LC correu em paralelo às pesquisas em Pragmática, sobretudo devido às diferentes abordagens propostas pelas duas áreas. A natureza da investigação pragmática é de índole claramente qualitativa, individual e local, interessada que está na análise da função de uma determinada expressão ou estrutura inserida num cotexto específico (Aijmer & Rühlemann 2015; Jucker 2018; Aijmer 2020), ao passo que a LC, ao analisar grandes massas de dados, tem preferido uma metodologia mais quantitativa e vertical.

Estabelecer uma articulação entre os dois domínios tem sido, aliás, um desafio que enfrenta algumas resistências. Muitos investigadores ligados à Pragmática alertam para o facto de os *corpora* servirem 'apenas' como porta de entrada para os dados e necessariamente terem de ser sujeitos a um rigoroso escrutínio em função do quadro teórico (e das perguntas de investigação) adotado, sob pena de o estudo se tornar apenas estatístico e não incluir nenhum tipo de explicação causal sobre os factos, ou seja, não testar uma teoria. Outros assinalam que a escassez de dados relativos ao perfil sociológico dos informantes e às condições (contextuais) que envolveram a recolha de dados pode enviesar ou pelo menos não auxiliar as análises pragmáticas (McEnery & Wilson 2001). Por outro lado, sempre que houver uma discrepância entre o que se diz (e que consta do *corpus*, oral ou escrito) e o que se pretende dizer, ou seja, sempre que estiver em jogo um significado mais implícito, a análise pragmática não encontrará respostas no *corpus*. E esta é uma lacuna relevante, à qual vários investigadores são

sensíveis: "the relationship between pragmatics and corpus linguistics is not unproblematic. The reason is simple: corpora record text, not meaning, and they record context only crudely" (Rühlemann 2010: 289). Avulta ainda o problema das ferramentas de anotação pragmática. O desenvolvimento dessas metodologias ainda se debate com alguns desafios de difícil resolução, pois não havendo muitas vezes, no domínio dos usos, uma correlação fixa entre forma e função, nem sempre é fácil detetar determinados fenómenos pragmáticos.

Apesar destes obstáculos, o entrosamento das duas áreas tem vindo, segundo Romero-Trillo (2008), a dar-se paulatinamente e os contributos da LC trouxeram claros benefícios às análises qualitativas realizadas no domínio pragmático, uma vez que é possível combiná-las com metodologias mais quantitativas. Parece ser este, justamente, o grande contributo da LC para a Pragmática: oferecer-lhe uma metodologia de análise de base empírica (Rühlemann & Clancy 2018: 242).[2] Entretanto, neste recente trilho conjunto, as duas áreas têm redefinido os respetivos objetivos. Por um lado, a LC tem vindo a admitir a pertinência científica dos pequenos *corpora*; por seu turno, a metodologia investigativa da Pragmática acomoda-se bastante melhor com esses pequenos e *domain-specific corpora*, que lhe permitem manter a base empírica, enquanto busca padrões pragmáticos numa base de dados mais manuseável (Vaughan & Clancy 2013: 70).

Nesta sequência, a Pragmática de Corpus, jovem disciplina que interseta as valências das duas áreas, vem-se afirmando como a "science that describes language use in real contexts through corpora" (Romero-Trillo 2017: 1)

1.2. Linguística de Corpus e *Second Language Acquisition (SLA)*

A pesquisa realizada no âmbito da área de investigação em Aquisição de Línguas Segundas (*SLA*) visa descrever os conhecimentos linguísticos do aprendente, a progressão desses conhecimentos ao longo das diversas fases de construção da interlíngua, ou seja, a forma como esses conhecimentos

2 Note-se que os procedimentos controlados de elicitação de dados, habitualmente usados na área, têm sido criticados precisamente pela sua falta de autenticidade, o que tem levado os investigadores a preferir dados mais autênticos ou, pelo menos, a cruzar os resultados obtidos a partir dessas duas metodologias.

evoluem (ou estagnam, no processo conhecido por 'fossilização') e os fatores que afetam, limitando ou incentivando, o desenvolvimento dessa competência. Neste sentido, as dimensões cognitivas do processo de aquisição têm vindo a ser muito exploradas, sobretudo no que respeita ao processamento da informação linguística, aos recursos atencionais e ao papel da motivação na aprendizagem de línguas estrangeiras, análises muitas vezes domiciliadas num quadro teórico de base generativa.[3] Ora, o desenho de um quadro teórico deste tipo tende a afastar o interesse dos investigadores da análise de dados autênticos, e muitos destacam que a investigação realizada no âmbito da *SLA* não tem favorecido o recurso a *corpora* (Lozano & Mendikoetxea 2013: 65).

Entretanto, outros autores (Leech 1997; Mukherjee 2006; Johansson 2007) salientam o potencial contributo dos *corpora* no processo de ensino-aprendizagem de uma língua estrangeira, especificando que esse contributo pode concretizar-se sob três formatos distintos: os dados da língua alvo podem ser explorados na sala de aula de língua estrangeira e constituir a base de novas aprendizagens; podem também ser usados, de forma indireta, servindo como base para a construção de material instrucional; uma terceira hipótese prevê o desenvolvimento de *corpora* especificamente orientados para o ensino.

É precisamente nesta última área que avultam os *learner corpora*, acervos de dados produzidos por aprendentes de uma língua estrangeira. Inicialmente, os *learner corpora* englobavam pequenas bases de dados decorrentes da recolha de professores ou investigadores, em aula; no entanto, a investigação efetuada a partir desses dados não permitia a obtenção de resultados com elevado poder explanatório e foram razões desta índole que levaram à constituição de *learner corpora* de maiores dimensões, como o ICLE (International Corpus of English Learner)[4], o primeiro grande *learner corpus*, resultante da escrita de textos argumentativos produzidos por aprendentes de inglês como L2[5] e organizado em diferentes *subcorpora* em

3 Hondo (2014: 2) assinala, no âmbito desta linha cognitivista, a emergência de outras orientações. Outras abordagens, de cunho mais funcionalista, têm vindo, entretanto, a surgir neste domínio. Veja-se: Atkinson (2011) e VanPatten *et al.* (2020).

4 Sobre este corpus, ver Granger *et al.* 2002.

5 O termo L2 refere, aqui, qualquer língua adquirida após a nativa, evitando assim os problemas definitórios inerentes às expressões 'língua segunda' e 'língua estrangeira'.

função da língua materna dos falantes.[6]

Ainda assim, e nem mesmo com a crescente entrada das abordagens socioculturais na área da *SLA*, ou seja, com a tónica nos fatores sociais que influem na aquisição da L2, o interesse pelos *learner corpora* aumentou. Em sentido inverso ao que seria previsível, o impacto dos *learner corpora* na pesquisa em SLA tem sido bastante lento (McEnery *et al.* 2019).

A preferência pelos dados experimentais (elicitados) – em detrimento dos dados reais produzidos pelos aprendentes – revelada pela área da *SLA* (Granger 2002) deve-se a algumas limitações destes últimos, apontadas pelos investigadores. Em primeiro lugar, os *learner corpora* são constituídos por dados, orais ou escritos, que apenas refletem as competências produtivas dos aprendentes, não dando conta das suas competências recetivas, nem espelhando os processos cognitivos necessários à construção da interlíngua do aprendente, razão por que a sua utilidade é, sob este prisma, reduzida. Um segundo problema relaciona-se com a escassez dos dados; por muito vasto que seja um *learner corpus*, ele pode não representar, de forma adequada, as diversas fases de construção da interlíngua, pode não refletir os verdadeiros conhecimentos do aprendente (o *intake*) e pode não oferecer suficientes evidências de um determinado fenómeno que se pretende estudar. As ferramentas de anotação de *corpora* constituem um terceiro problema, uma vez que ainda não estão suficientemente desenvolvidas para identificar fenómenos numa base de dados que se afasta, em múltiplos aspetos, da variedade padrão, exigindo, portanto, uma metodologia de anotação manual para tratamento dos erros e desvios dos aprendentes.

No global, o investigador que trabalha na área de investigação da *SLA* crê que é difícil, por um lado, controlar as numerosas variáveis que operam em contextos mais autênticos (Granger 2002; Mackey & Gass 2005) e, por outro, considera que os metadados que acompanham estes acervos não explanam exaustivamente todas as diferentes variáveis relevantes para a pesquisa.

Apesar destas fragilidades, a investigação realizada no domínio da *SLA* pode beneficiar do aporte trazido pela análise de um grande *learner corpus*.

---

6 O ICLE foi posteriormente valorizado com o advento de outros corpora, como o LOCNESS (Louvain Corpus of Native English Essays) que inclui um grande acervo de textos argumentativos produzidos por estudantes nativos ingleses e norte-americanos, possibilitando, assim, análises contrastivas (nativo vs. não nativo).

A vastidão dos dados disponíveis pode permitir ao investigador encontrar tendências, estruturas, expressões, usos e desvios que, de outra forma, lhe escapariam. Com as atuais valências tecnológicas, será certamente possível criar um instrumental de anotação adequado, que permita uma fácil identificação de padrões de coocorrência e de listas de frequência. Estes resultados poderiam alimentar muitos estudos nesta área, nomeadamente no que toca aos aspetos gramaticais mais prontamente adquiridos, mais vulneráveis ou mais sujeitos a fossilização e, por outro lado, ao papel das interfaces na aquisição de uma língua segunda (Lozano & Mendikoetxea 2013).

2. Os *corpora* orais de aprendentes de L2 (s*poken learner corpora*)

Embora o campo de investigação conhecido como *Learner Corpus Research* (LCR)[7], tenha direcionado a sua atenção sobretudo para os *learner corpora* contendo dados escritos, os *spoken learner corpora* têm vindo a ganhar cada vez mais protagonismo. Ainda raros e principalmente pouco explorados, estes *corpora* resultam da gravação e posterior compilação de discursos orais, produzidos na língua alvo por falantes não nativos.[8] Estes acervos podem apresentar-se sob três formatos distintos: apenas na versão transcrita, habitualmente apelidados de *mute spoken corpus* (Ballier & Martin 2015); na versão transcrita acompanhada dos respetivos áudios; e, na versão mais moderna, sob formato multimodal, combinando dados provenientes de áudio e de vídeo, na tentativa de apreender gestos, expressões faciais, olhares, comportamento proxémico e háptico e todo o tipo de sinais comportamentais que enriquecem e dão sentido à interação verbal. Muito recentes, estes *corpora* multimodais são, todavia, ainda mais raros, tendo em conta as dificuldades de natureza ética e legal que envolvem a recolha e a divulgação pública da imagem e dos dados pessoais.

A carência de *spoken learner corpora* deve-se a razões várias, relacionadas com a difícil tarefa de coletar os dados, uma vez que cada informante ou

7 Nesta área, pontuam os trabalhos de Sylviane Granger.

8 Assinale-se que, ao mencionarmos os *spoken learner corpora*, tal não significa, como muito bem assinalam Lozano *et al*. (2021), que estamos a reportar-nos a dados necessariamente autênticos, uma vez que, não raro, estes dados são produzidos em contexto de aula e em condições controladas.

grupo de informantes tem de ser gravado individualmente, com o tempo consumido na respetiva transcrição e com a relativa ausência de ferramentas que deem conta das especificidades do oral na escrita (Luzón *et al*. 2007: 3). A este respeito, é indubitável que a análise de *spoken learner corpora* se torna ainda mais difícil por requerer uma atenção particular aos "elements beyond the text, such as intonation, gesture and discourse structure, which cannot easily be explored with the use of the kinds of frequency-based techniques used in the analysis of written corpora" (Adolphs & Knight 2010: 38). Com efeito, e tal como foi assinalado por diversos investigadores, um *corpus* oral é sempre multimodal e, portanto, engloba também elementos paraverbais e não verbais que complexificam a tarefa da transcrição. Por este motivo, é sempre necessário tomar decisões quanto ao grau de granularidade da transcrição de um *spoken learner corpus*, a qual pode envolver diferentes níveis de complexidade.

Não por acaso, muitos *spoken learner corpora* foram criados para possibilitar, sobretudo, pesquisas de natureza fonética e fonológica, com exercícios orientados de leitura e com um elevado controlo do *output*, de modo a permitir a comparabilidade dos dados e análises quantitativas rigorosas (Díaz-Negrillo & Thompson 2013), e não tanto para permitir, por exemplo, a descrição de uma interação verbal em língua segunda, mais imprevisível, mais complexa e em que o contexto (conceito difícil de definir, nas suas múltiplas vertentes – epistémica, linguística, interacional e social) é determinante.

De qualquer modo, os *spoken learner corpora* podem fornecer uma valiosa base empírica para analisar o desenvolvimento das competências orais do aprendente, nos diversos planos de descrição linguística. Se forem concebidos com objetivos mais ambiciosos, nomeadamente o de incluírem diferentes amostras de discurso espontâneo e finalístico, monogerado e poligerado, então as suas valências aumentarão consideravelmente, pois permitirão obter materiais adequados à descrição das interlínguas dos aprendentes no plano pragmático-discursivo, possibilitando o acesso a informações diversas que um *corpus* escrito tipicamente não traduz: silêncios, hesitações, interrupções, sobreposições e todos os traços vocais e prosódicos que refletem as dimensões emocionais e atitudinais que habitualmente acompanham a fala (MacWhinney 2021).

3. A pertinência de um *learner corpus* de interações orais em PLNM

No que tange à língua portuguesa na sua variante europeia, estão disponíveis três *corpora orais* de aprendentes de Português como L2, que envolvem, aliás, dados de natureza bastante díspar. Os dados constantes do COPLE2[9] resultam da gravação de conversas entre dois candidatos, ocorridas durante um exame de natureza oral, do CAPLE, e mediadas por um avaliador (Mendes *et al*. 2016). No caso do projeto CAL2[10], os dados orais correspondem a entrevistas realizadas entre um aprendente e um falante nativo do português sobre temas diversos (Ferreira *et alii*, no prelo). O *corpus* COral-Co[11] apresenta um conjunto bastante variado de dados de produção oral, que decorrem da aplicação de um inquérito, envolvendo distintas tarefas, orientadas para permitir análises em diferentes domínios de funcionamento da língua portuguesa (fonético-fonológico; morfológico; sintático; semântico; lexical; pragmático; textual).[12]

Estes *corpora* resultam, assim, de atividades de avaliação, da interação entre nativo e não nativo, e de tarefas de elicitação de dados (respetivamente). Não escamoteando a sua inequívoca relevância, é notória a ausência de um *corpus* que dê conta da interação verbal entre aprendentes de PLNM.

Apesar de muitas abordagens ao ensino/aprendizagem de L2 estarem, em si mesmas, enformadas pela centralidade atribuída ao conceito de interação verbal, como é o caso das teorias sociointeracionistas (Mondada & Doehler 2004) e, particularmente, da *Communicative Language Teaching* (Richards 2006) e apesar de um bom domínio da competência interacional ser o objetivo último de qualquer processo de ensino/aprendizagem de uma L2, esta é uma área em que os aprendentes exibem, normalmente, grandes dificuldades comunicativas, quer a interação verbal em que participam seja mais espontânea ou mais finalística.

A competência interacional (Celce-Murcia 2007) integra a capacidade

---

9 *Learner Corpus Português L2 – COPLE2*, do Centro de Linguística da Universidade de Lisboa (CLUL), com coordenação de Amália Mendes.

10 *Projeto CAL2 – Corpus de Aquisição L2 - subcorpus Produção Oral*, do Centro de Linguística da Universidade Nova de Lisboa (CLUNL), com coordenação de Ana Madeira. Para mais informações, consultar: http://cal2.clunl.fcsh.unl.pt/index.html.

11 *Corpus Oral de Português L2 – Coimbra (COral-Co)* – do CELGA-ILTEC, da Universidade de Coimbra, com coordenação de Isabel Almeida Santos.

12 Sobre o *corpus COral-Co*, ver Santos *et al.* (2016).

de construir uma conversa, desde a secção de abertura até à fase de encerramento, gerindo todas as vertentes próprias de uma interação verbal que é, também, sempre social. Neste âmbito, o aprendente terá de fazer uso da competência acional, que prevê a construção de atos ilocutórios adequados aos vários momentos da interação, mas também terá de dominar a competência conversacional, gerindo, de forma apropriada, as regularidades inerentes à organização sequencial do diálogo, nas fases mais relacionais e mais transacionais da interação, e ver-se-á ainda instado a organizar adequadamente todo o comportamento paraverbal, cinésico e háptico (gestos, olhares, distância do interlocutor e sinais de retroação, por exemplo). Todas estas imposições representam, assim, um ónus considerável para um aprendente de uma L2, na medida em que lhe exigem a gestão adequada e simultânea dos planos transacional e relacional, ainda por cima enquadrados por normas reguladoras que são, não raro, culturalmente dependentes.

Razões desta natureza exigem, por um lado, um conhecimento profundo e relativamente sistematizado das características e do funcionamento das interações orais, em PLNM, produzidas por aprendentes de português europeu e impõem, por outro, um trabalho intensivo em torno do desenvolvimento da competência interacional dos aprendentes. Com efeito, é necessário ter uma descrição fundamentada dessas produções orais, com o objetivo de identificar as áreas de maior dificuldade e fragilidade, as subcompetências mais carenciadas de intervenção, em suma, para realizar um levantamento exaustivo dos aspetos que terão de ser alvo de instrução explícita e de ser contemplados em contexto instrucional. [13]

A constituição de um *corpus* de interações orais produzidas por aprendentes de PLNM constitui um primeiro passo para dar consecução a este desiderato.

13 Reconhece-se também, naturalmente, a necessidade de obter dados provenientes de falantes nativos. Numa fase posterior deste projeto, poderá equacionar-se a hipótese de vir ainda a incluir subcorpora de controlo de falantes nativos das LM dos informantes.

3.1. Desafios na compilação de um *corpus* de interações orais em PLNM – resultados preliminares

O objetivo primordial da constituição deste *corpus* oral é a obtenção de uma base de dados constituída por interações verbais orais, protagonizadas por uma díade ou tríade de participantes – aprendentes adultos de PLNM – que discutem, num determinado intervalo temporal, pontos de vista acerca de um tópico sensível, e foi com este propósito que se iniciou, em maio de 2022, a fase-piloto da recolha de dados. As opções metodológicas adotadas nesta fase fizeram emergir algumas questões que serão, agora, objeto de análise.

Ao ambicionar aceder a discurso oral espontâneo, foi necessário refletir sobre a metodologia de recolha de dados. A exigência, epistemológica, de trabalhar com dados autênticos deveria obrigar o investigador a coletar discurso espontâneo, sem o conhecimento dos envolvidos na recolha e, obviamente, sem qualquer intervenção da sua parte, evitando, assim, o recurso a procedimentos de recolha de dados que permitissem ao informante preparar (ou refletir sobre) o seu próprio desempenho. Todavia, de um ponto de vista legal, tal metodologia é impraticável, razão por que a elicitação de dados, enquanto metodologia orientada, tem sido a mais utilizada no domínio da *SLA*.

Nesta linha de raciocínio, e considerados todos os potenciais problemas, a metodologia utilizada na fase-piloto foi a da *elicited conversation*, método que, segundo Taguchi & Roever (2017), constitui um procedimento pouco utilizado, mas promissor. Neste caso, e uma vez lançado, pelo investigador, o tópico de discussão, toda a interação fica a cargo dos intervenientes; a ausência do investigador do contexto conversacional e a não existência de um *script* rígido deixam aos participantes o ónus de coconstruir a sua interação. Esta opção possibilita uma produção oral (quase-) espontânea, a decorrer em situação (quase-) informal e com a participação mínima do moderador (o investigador), razão por que pode aproximar-se bastante do perfil de um verdadeiro debate na língua alvo. Não sendo dados autênticos, estes poderão, seguramente, ser caracterizados como *near-natural*.

A metodologia a utilizar envolveu também a decisão relativa ao assunto que constituiria o móbil do debate. A escolha recaiu sobre um tópico

acessível, mas suficientemente instigante para motivar a discussão, suscitado por um título do jornal Observador: "O Parlamento criou um grupo de trabalho para discutir o fim dos animais nos circos e nos jardins zoológicos. Manter os animais nos jardins zoológicos e nos circos deve ser proibido?".

Aos intervenientes foram dadas indicações genéricas sobre a atividade, de modo a permitir que esta decorresse da forma mais natural possível e sem a subsequente intervenção do investigador. Para enquadrar a interação, foi criado um cenário verosímil – apresentado aos participantes, pelo investigador, no início da atividade – que contemplava um contexto específico: um encontro entre três colegas, no bar da faculdade, durante um intervalo de 10 minutos entre duas aulas, no âmbito do qual, e a propósito do título do Observador que um deles deveria ler em voz alta, decorreria o debate.

Não tendo previsto fornecer aos aprendentes, previamente, quaisquer informações acerca da atividade em que iam participar, para promover a autenticidade da interação oral, o primeiro grupo de três informantes cuja produção oral foi recolhida na fase-piloto pediu, no entanto, algum tempo de preparação antes do início da tarefa. A imprevisibilidade da situação causou manifesta ansiedade em todos os membros do grupo, evidenciando a complexidade da tarefa e o risco de o grupo poder vir a ter sérias dificuldades em produzir dados orais, e consequentemente, em cumprir os objetivos pretendidos.[14]

Estas dificuldades obrigaram a repensar toda a planificação inicial e conduziram à sua reformulação. Num segundo momento da fase-piloto, já com outros informantes, optou-se pela diferenciação de contextos de recolha, exigindo a cada um dos grupos participantes um grau de preparação distinto. Na segunda fase, participaram nove aprendentes, oito de nível C1+ e um de nível B2, constituindo três grupos de debate, e a cada um desses grupos foi apresentado um cenário diferente. Ao primeiro grupo foram dadas instruções prévias (com uma semana de antecedência) sobre os objetivos da atividade, o tema e a forma como iria processar-se o debate, instruções que incluíam, também, o papel interacional a desempenhar por cada um (um dos participantes teria de defender um determinado ponto

14 Esta dificuldade foi também reiterada na resposta à pergunta – sobre propostas de melhoria da atividade – que constava da ficha de dados individuais solicitada a cada um dos participantes no fim da atividade.

de vista; outro deveria contra-argumentar; o terceiro elemento funcionaria como voz do consenso, tentando equilibrar e harmonizar a discussão). O segundo grupo diferiu do anterior pelo facto de ter conhecimento do tema e das instruções apenas no momento do debate. Também ao terceiro grupo foi apresentado o tema-estímulo pelo investigador, imediatamente antes da interação, não tendo sido fornecida qualquer outra instrução, podendo cada um dos participantes defender o ponto de vista que entendesse.[15]

Sendo certo que, segundo Biber (1993), o processo de compilação de um *corpus* deve ser recetivo aos problemas surgidos aquando da coleta de dados, com esta diversidade de opções pretendeu-se calcular em que medida a pré-preparação e a pré-determinação de papéis interacionais do debate 'contaminariam' a naturalidade da interação, por um lado, e avaliar qual o risco de o debate fracassar face à inexistência de instruções rígidas, por outro, ou seja, avaliar qual a melhor metodologia, tendo em conta os objetivos do *corpus*.

Para além desta avaliação, há ainda algumas questões mencionadas pelos participantes e sobre as quais urge também refletir, antes de tomar uma decisão definitiva, nomeadamente as que respeitam: à necessidade de introduzir um momento inicial, que permita contextualizar o assunto e familiarizar os aprendentes com a temática, contrariando a impositividade da tarefa; à possibilidade de formar grupos mais alargados, para que cada participante se sinta mais apoiado na defesa da sua opinião; à eventual rotatividade dos aprendentes nos vários grupos, permitindo a cada um participar no debate mais do que uma vez, mas com diferentes interactantes.[16]

A seleção dos informantes constituiu um outro aspeto importante na fase-piloto da recolha de dados.

Contrariamente ao desenho inicial do projeto, que previa a recolha de dados envolvendo informantes com diferentes línguas maternas e de distintos níveis de proficiência – a iniciar no nível B1, dada a complexidade da tarefa –, foi apenas possível obter, na fase-piloto, os contributos de

15 Este terceiro cenário correspondia, como é óbvio, ao inicialmente previsto no protocolo.

16 Se a primeira sugestão parece ser de fácil execução, as duas seguintes podem ser problemáticas. Um grupo mais alargado implicará mais dificuldades na identificação das vozes e só resultará se os dados forem recolhidos em vídeo, o que nos reconduz às questões legais atrás identificadas. Por outro lado, e embora a dinâmica das interações orais, sempre diferente, obrigue os aprendentes a reagir verbalmente de forma diferenciada, a possibilidade de um mesmo aprendente poder participar no debate, em grupos distintos, pode vir a diminuir o grau de naturalidade da sua participação.

quatro grupos, ou seja, de doze aprendentes de PLNM, onze de nível C1+ e um de nível B2+,[17] que se encontravam em contexto instrucional em Portugal e, portanto, em regime de imersão, a frequentar cursos ou unidades curriculares de língua, literatura ou cultura portuguesas.[18] De igual modo, as línguas maternas destes doze informantes são também pouco variadas (nove falantes de chinês mandarim; um falante de árabe; um falante de italiano e um falante de japonês). Tal assimetria pode, por um lado, vir a revelar alguns dados interessantes, no caso dos grupos constituídos por falantes que partilham a mesma língua materna,[19] mas, por outro, pode vir a obstaculizar o equilíbrio do *corpus* relativamente às LM nele representadas, pode vir a inviabilizar a constituição de tríades que partilhem outras LM e pode até dificultar a constituição de grupos com grande diversidade de LM. Esta é uma clara desvantagem que resulta do facto de a base de recrutamento de informantes ser escassa e pouco diversificada.

Um outro potencial problema relacionado ainda com os informantes é o que decorre do número, desigual, de alunos inscritos nos diferentes níveis de aprendizagem, havendo bastante menos alunos a frequentar os níveis mais altos, o que pode dificultar, de novo, a constituição de um *corpus* equilibrado, desta feita no que concerne aos níveis de proficiência.

Finalmente, ao assumir que os contributos dos intervenientes são voluntários, e foi desse pressuposto que partiu o convite que lhes foi endereçado para, de forma espontânea, participarem numa atividade extracurricular, é ainda de equacionar a possibilidade de só os alunos mais proficientes se terem apresentado, o que pode enviesar os dados obtidos e retirar representatividade ao *corpus* (Gilquin 2015).

Um outro aspeto importante no que concerne à recolha de dados orais diz respeito ao próprio processo de gravação, aqui incluindo o espaço e o equipamento. Tentou evitar-se a captação de ruídos de fundo, cuidado fundamental para preservar a qualidade da gravação e, mais ainda, para permitir uma transcrição do oral que pode vir a revelar-se dilatada no tempo

17 Idealmente, os grupos deveriam ser constituídos por informantes com o mesmo nível de proficiência.

18 As produções orais dos quatro grupos da fase-piloto perfizeram um total de cinquenta minutos de gravação, com uma média de doze minutos e meio por grupo.

19 Num dos grupos cuja produção oral foi recolhida na fase-piloto do projeto, constituído por estudantes que partilhavam precisamente o chinês mandarim, como LM, é audível um fenómeno de *code-switching*, em que um dos participantes sussurra algo em mandarim, à espera de obter, dos colegas de grupo, a tradução portuguesa da expressão chinesa.

e exigente quanto à discriminação vocal dos intervenientes, sobretudo se estes forem mais do que dois. Com efeito, e considerada a participação de três intervenientes do mesmo sexo, como aconteceu em dois dos grupos da fase-piloto, ficou patente que o equipamento de gravação a utilizar deveria permitir identificar as diferentes vozes. Para coadjuvar o processo de reconhecimento vocal, foi solicitado às participantes nestas interações que se identificassem no início da gravação (embora essa informação não venha a ser divulgada aquando da futura disponibilização dos ficheiros áudio).[20] Por outro lado, e embora tal não seja o objetivo primeiro da constituição deste *corpus* oral, uma boa qualidade sonora permitirá também outras possibilidades de pesquisa. No que concerne ainda ao equipamento, é importante procurar um adequado equilíbrio entre a qualidade e a discrição. Segundo Golato (2017: 23), "the very presence of recording equipment may alter subjects' speech production", por isso foi necessário ponderar qual dos mecanismos de captação e de gravação de voz – gravador, computador ou telemóvel – seria o mais fidedigno (em termos de registo sonoro) e o menos invasivo. Durante a fase-piloto, a recolha dos dados foi feita por uma dupla via (telemóvel e *tablet*), precisamente para permitir analisar a qualidade sonora dos dois dispositivos e fundamentar uma decisão posterior.

Apesar de todos estes procedimentos cautelares, no sentido de melhorar o desenho do protocolo e de otimizar o processo de recolha, não descurando os objetivos iniciais do projeto, será ainda necessário avaliar o peso da própria situação de gravação, uma vez que também esta pode influenciar a interação oral que vai ser gravada. A marcação de um dia e de uma hora para reunir os participantes e dar consecução à tarefa retira-lhe também alguma genuinidade. A imposição de um tópico de debate, pelo investigador, e a obrigatoriedade de participação de cada um dos intervenientes, associadas ao facto de os participantes estarem cientes de que o seu desempenho será avaliado por um investigador, ainda que essa entidade se ausente do contexto enquanto a interação decorre, constituem outras constrições que, previsivelmente, terão consequências na produção oral dos falantes.

Tentar minimizar o impacto de todos estes fatores é um objetivo difícil

20 Prevê-se a disponibilização do material áudio acompanhado da respetiva transcrição.

de atingir, mas tal desiderato deve permanecer sempre no horizonte de preocupações do investigador e, mais do que isso, levá-lo a documentar bem todo o contexto em que a gravação tem lugar, com informação acerca dos participantes, do espaço, do tempo, do equipamento e de todos os dados que possam vir a ser relevantes numa audição e análise posteriores (Adolphs & Knight 2010).

Esta atenção aos metadados orientou já todo o processo de recolha de dados na fase preliminar e a elaboração do documento intitulado *Perfil do Informante* foi pensada nesse sentido. Para além de um bem detalhado consentimento informado – na medida em que os dados orais vão tornar-se públicos – que todos os envolvidos tiveram de ler e assinar, e que já visava obter informação pormenorizada sobre aspetos biológicos, geográficos, sociológicos e linguísticos dos envolvidos, foi explicitado o quadro, tão aprofundado quanto possível, das diversas vertentes envolvidas no processo de gravação dos dados orais. De facto, "[t]o 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).[21]

Importa, por fim, considerar ainda um aspeto fundamental: a representatividade de um *corpus* deste tipo. Embora esta questão possa levantar dúvidas, nomeadamente quanto à quantidade de textos/discursos que representam, de forma fidedigna e completa – se é que é possível falar de completude, neste caso –, um determinado género textual, considerando os onerosos procedimentos a que obriga (em termos de tempo, de espaço e de equipamento), um *corpus* oral deste tipo tende necessariamente a ser mais pequeno (Adolphs & Knight 2010), sendo que esta menor dimensão não invalida, necessariamente, a sua validade. De facto, e dependendo dos objetivos visados, um *corpus* mais pequeno pode ser suficientemente representativo de uma determinada população; por outro lado, justifica-se

21 Estas autoras sugerem que é necessário disponibilizar informação sobre: a instituição em que decorre a gravação; o tipo de conhecimento prévio dos informantes acerca do evento em que vão participar; o seu grau de conhecimento do género textual em causa; eventuais pesquisas previamente realizadas acerca do tópico a discutir; o contexto sociointeracional desenhado pelo investigador no início de cada uma das sessões de gravação; o hipotético recurso dos participantes a auxiliares escritos, durante a interação; o tempo que demora a contextualização inicial e a atividade interacional em si mesma, entre outras informações possíveis (Bell & Payant 2021). Somente desta forma se avaliará o grau de autenticidade dos dados recolhidos e o peso detido pelo enquadramento, simulado, que os enquadra.

também o seu menor tamanho se considerarmos que se trata de um *corpus* mais especializado; por fim, havendo um rico conjunto de metadados, como acima se referiu, os investigadores terão à sua disposição todas as informações de que necessitam para avaliar as instâncias discursivas presentes.[22]

3.2. Possibilidades de investigação

Conquanto alguns investigadores considerem que um *corpus* deste tipo é pouco interessante para a investigação no domínio da *SLA*, por conter mais informações sobre os usos efetivos que os aprendentes fazem da língua alvo do que sobre o processo de aprendizagem dessa língua, as valências de um *corpus* de interações orais em PLNM são inúmeras, devem ser ressaltadas e podem mesmo revelar-se aliciantes para o tratamento das questões que, tipicamente, têm ocupado a agenda investigativa na área da *SLA*.

Em primeiro lugar, e tratando-se de um tipo de interação oral relativamente mais espontânea, exterior ao contexto letivo, os dados obtidos serão mais naturais do que os disponibilizados pelos diferentes *corpora* orais existentes, uma vez que grande parte destes últimos foi recolhida em contexto de avaliação e em contexto de sala de aula. É também crível que, neste tipo de trocas verbais, de natureza dialógica e em contexto *near-natural*, seja visível o manifesto e genuíno interesse dos falantes em defender uma causa, o que concorrerá para a maior autenticidade dos dados obtidos. No mesmo sentido, o conforto psicológico advindo de uma interação entre pares, sem a presença e a intervenção do professor/investigador, será, previsivelmente, um adjuvante não apenas de maior produção linguística, como de menor inibição em participar (Philp *et al*. 2014).

O facto de se tratar de uma *elicited conversation* concorre, de igual forma, para a riqueza e para a originalidade dos dados obtidos, pois, neste caso, a produção oral dos participantes não se reduz a palavras ou sequer a frases, mas traduz-se em textos/discursos, ou seja, aquilo a que acedemos é ao produto de uma atividade comunicativa completa, a um conjunto

22 Por limitações de espaço, não serão aqui considerados os constrangimentos impostos pela transcrição destes dados.

de intervenções poligeradas, devidamente articuladas entre si e inseridas num contexto que, não sendo efetivamente real, constitui, nas palavras de Kramsch (1986: 367), "a shared internal context or 'sphere of inter-subjectivity' that is built through the collaborative efforts of the interactional partners".

Por outro lado, não estando dirigido para a elicitação de determinadas estruturas ou para a verbalização de funções linguísticas específicas, este *corpus* oferece uma base de dados mais ampla, relativa ao uso oral do português como língua não materna. Neste mesmo sentido, e dado estar centrado num género específico, que envolve sequências textuais de natureza argumentativa, será um *corpus* inovador, dotado de amplas potencialidades no que diz respeito à investigação em PLNM.

Como facilmente se depreende, o protocolo que orientará a recolha de dados foi pensado, em primeiro lugar, para permitir análises de natureza pragmático-discursiva que possibilitem a descrição das interlínguas dos participantes neste domínio. De facto, o objetivo primeiro deste *corpus* é o de permitir a análise da competência interacional dos aprendentes de PLNM, área ampla que contempla: a capacidade de realizar atos de fala e sequências de atos de fala em interações na língua alvo, que visem a troca de informações, a expressão de opiniões, a explicitação de problemas, a criação de cenários hipotéticos, entre outras possibilidades (âmbito da competência acional); a capacidade de gerir a interação verbal, no plano da organização dos turnos de fala, da abertura e do fecho das trocas verbais, da introdução, fecho, alteração ou retoma de tópicos (âmbito da competência conversacional); o domínio dos mecanismos de coesão lexical e gramatical, sobretudo no que concerne à adequada utilização de cadeias anafóricas e de marcadores discursivos; o domínio dos deíticos enquanto marcadores de ancoragem situacional; o domínio dos esquemas formais e das estratégias retóricas que permitem identificar uma determinada estrutura genológica (âmbito da competência discursiva); e o domínio da competência paraverbal (silêncios, pausas e sinais de retroação).

Neste sentido, são evidentes as potencialidades de um *corpus* oral deste tipo. As interações verbais a gravar, mais ou menos extensas, conterão um manancial de dados que refletirá a forma como estes falantes lidam com as exigências de uma interação oral em curso, em que têm de dar consecução

aos seus objetivos comunicativos, integrá-los na sequência discursiva a decorrer, negociar e gerir tópicos, construir argumentos e contra-argumentos, manter a vertente relacional, adotar ora o papel de locutores, ora o papel de ouvintes. No fundo, o acervo exibirá a forma como os participantes interagem, defendem pontos de vista e contra-argumentam numa L2, bem como a forma como coconstroem, momento a momento, uma interação verbal. Neste âmbito particular, há, aliás, uma outra dimensão que deve ser salientada. Na verdade, este *corpus* oral possibilitará a compreensão da forma como as competências de cada um se conformam e se ajustam reciprocamente à medida que a própria interação vai decorrendo. Subscrevemos, assim, as palavras de Young (2011: 428), quando afirma que a competência interacional (IC) "is not the knowledge or the possession of an individual person, but is co-constructed by all participants in a discursive practice, and IC varies with the practice and with the participants."

Outra das valências deste *corpus* oral – e seguramente uma das mais importantes – diz respeito aos processos de negociação de sentido, isto é, aos momentos de incompreensão, de ocorrência de erros pragmáticos e, até, de quebras comunicativas, que seguramente sucederão, e aos efeitos dessas dificuldades na própria interação (Walsh 2010), bem como à forma como os participantes ultrapassarão (ou não) esses incidentes. As estratégias de evitação e de paráfrase, os mecanismos de reformulação, as autocorreções, os pedidos de clarificação, os comentários de natureza metadiscursiva, o recurso ao *code-switching* ou os fenómenos de transferência a que recorrerão para remediar ou resolver esses momentos de impasse poderão constituir, para o investigador, informação relevante para compreender qual a relação entre as estratégias escolhidas e o nível de proficiência, por exemplo, e até para avaliar quais as características desta competência estratégica em cada uma das fases de desenvolvimento da interlíngua destes aprendentes.

Concomitantemente, um *corpus* deste tipo possibilitará também a análise das competências de natureza sociolinguística (exigidas pela contextualização inerente ao uso da língua) exibidas pelos aprendentes. Considerado o contexto de debate, enquanto confronto de pontos de vista (pelo menos parcialmente) divergentes sobre um determinado tópico, é previsível que surja a necessidade de discordar, de contra-argumentar, de contradizer, e, nessa tentativa de deslegitimar o discurso do(s) outro(s),

emergem as questões ligadas à cortesia, à potencial agressão à face do(s) outro(s), à necessidade de mitigar algum desse conflito ou, pelo contrário, de o extremar para vincar posições. Num contexto como este que aqui se desenha, será pertinente analisar a ocorrência de *face-threatening acts,* entre outro tipo de possíveis expressões de descortesia, mas também o recurso a estratégias atenuadoras e intensificadoras. A definição da relação interpessoal e, no fundo, o sistema de cortesia que vigora no início da interação, enquanto conjunto de direitos e deveres tacitamente aceite por todos os participantes, constituirão, de igual forma, um apelativo objeto de análise, sobretudo se os termos em que o debate decorre se forem alterando e negociando ao longo da interação.

Em rigor, o interesse pelas características e pelo funcionamento da interação, assim como pelos fatores que nela repercutem, já tinha atraído a atenção dos investigadores no âmbito das orientações mais interacionistas da área da SLA – veja-se o trabalho de Long (1981), por exemplo – mas o foco recaía, sobretudo, na interação verbal entre nativo e não nativo e a interação verbal entre dois não nativos, enquanto contexto de aquisição, ainda não tinha sido plenamente explorada. Por esta razão, um *corpus* de interações orais em PLNM pode revelar-se fundamental na investigação em L2, na medida em que possibilita aos investigadores a oportunidade de estudar o papel da própria interação na aprendizagem da língua alvo.

Quanto aos benefícios dos *corpora* orais para a área de investigação na área da *SLA*, é inegável que a análise destes textos/discursos autênticos permite dar mais consistência à investigação sobre o desenvolvimento das competências linguísticas do aprendente nas diferentes fases de construção da sua interlíngua. Com efeito, algumas das questões que têm estado sob o foco da *SLA*, pelo menos na sua linha mais cognitivista, podem ser reavaliadas se observadas à luz dos dados empíricos produzidos pelos aprendentes. O contraste entre textos/discursos de aprendentes de diferentes línguas maternas pode auxiliar a perceber a existência de tendências universais no processo de aprendizagem. A comparação das produções orais de aprendentes com a mesma língua materna, de diferentes níveis, pode permitir captar a influência dessa mesma língua no processo de aprendizagem, bem como avaliar o maior grau de complexidade e de fluência que essas produções exibem, à medida que o grau de proficiência

aumenta, o que pode ajudar a definir parâmetros avaliativos e a perceber quais as etapas e qual a natureza do processamento da L2. De igual forma, a identificação do erro constitui um precioso auxiliar no acesso às áreas mais críticas da aprendizagem da língua alvo (Myles 2015), ao mesmo tempo que permitirá, se analisado em diferentes níveis de proficiência, retirar ilações acerca dos processos de fossilização. Um bom conjunto de metadados pode também constituir um manancial de informações para estudar variáveis que são relevantes nesta área de investigação, quer se trate de traços de natureza sociolinguística, como a idade, o percurso académico, o número e a ordem de aprendizagem de outras línguas para além da materna, ou contextual, como o modo de comunicação (McEnery *et al*. 2019).

Mas outras questões relacionadas com as teorias de aprendizagem podem também beneficiar do aporte trazido pelos *corpora* dos aprendentes e, sobretudo, dos *corpora* orais. A forma como estes falantes acedem à L2 e a usam em tempo real (Myles 2015) pode ser clarificada com a análise de dados orais não planeados, na medida em que o investigador tem, neste caso, acesso privilegiado ao conhecimento linguístico implícito do aprendente. De facto, a dinâmica e a pressão de uma interação oral não permitem ao aprendente nem a reflexão prévia nem a correção subsequente, razão pela qual MacWhinney (2021: 159), citando Myles (2015: 314) considera até que "[f]rom the viewpoint of SLA theory, oral production is 'a better window into implicit knowledge'".

3. Palavras finais

A constituição de um *corpus* de interações orais em PLNM assume grande relevância em vários planos. Em primeiro lugar, na área do ensino/aprendizagem do português como língua não materna. Observar e analisar o comportamento linguístico dos não nativos neste tipo de contexto interacional permitirá obter uma compreensão mais apurada e rigorosa de uma área de investigação em que há evidentes lacunas descritivas e que não é trabalhada, de forma sistemática, em contexto de instrução formal. Analisar a interação oral (informal) em L2 revela-se, assim, uma preciosa pista de trabalho tendente a evidenciar em que medida o processo de

ensino/aprendizagem de uma gramática da L2 serve (ou não) os propósitos e as exigências de uma verdadeira interação oral. Os dados resultantes dessa análise podem auxiliar na melhoria das práticas pedagógicas, bem como ajudar a repensar muitos materiais instrucionais.

Naturalmente, e para que os objetivos delineados neste estudo sejam plenamente atingidos é necessário haver um conhecimento sistematizado das características e do funcionamento do mesmo tipo de interações orais entre nativos. Embora alguns autores questionem a existência de uma 'norma nativa', a possibilidade de detetar desvios e problemas comunicativos no *corpus* de interações orais em PLNM só é possível através do contraste com um *corpus* proveniente de falantes nativos.

Em segundo lugar, é também possível abordar algumas das questões de investigação da área da *SLA* através da análise de um *corpus* de interações orais em PLNM, uma vez que os problemas e os erros nele detetados permitem aceder às estratégias usadas pelos aprendentes no processamento da L2.

Os benefícios que um *corpus* oral deste tipo pode trazer à investigação em PLNM são evidentes, razão por que é necessário continuar a recolher *corpora* orais de diferentes géneros e a enriquecer as bases de dados existentes. Ao avançar neste sentido, um *corpus* de natureza multimodal seria de toda a utilidade, para dar mais consistência às análises do oral, mas também um *corpus* longitudinal, na medida em que seria possível averiguar, a longo prazo, a evolução da competência interacional dos aprendentes. Uma outra via que se afigura essencial diz respeito à necessidade e à urgência de desenvolver novas ferramentas digitais que permitam a análise destes *corpora* orais de aprendentes.

REFERÊNCIAS

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). Routledge.

Aijmer, K. (2020). Contrastive pragmatics and corpora. *Contrastive Pragmatics*, *1*(1), 28-57.

Aijmer, K., & Rühlemann, C. (Eds.). (2015). *Corpus Pragmatics: A Handbook*. Cambridge University Press.

Atkinson, D. (Ed.). (2011). *Alternative Approaches to Second Language Acquisition*. 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 University Press.

Bardovi-Harlig, K. (2013). Developing L2 pragmatics. *Language Learning*, *63*(s1), 68-86.

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). Routledge.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, *8*(4), 243-257.

Celce-Murcia, M. (2007). Rethinking the role of communicative competence in language teaching. In E. Alcón Soler, & M. P. Safont-Jordà (Eds.), *Intercultural Language Use and Language Learning* (pp. 41-58). Springer.

Díaz-Negrillo, A., & Thompson, P. (2013). Learner corpora: Looking towards the future. In A. Díaz-Negrillo, N. Ballier, & P. Thompson (Eds.), *Automatic Treatment and Analysis of Learner Corpus Data* (pp. 9-30). John Benjamins.

Ellis, R., & Barkhuizen, G. (2005). *Analysing Learner Language*. Oxford University Press.

Ferreira, T.; Santos, I.; Carapinha, C.; Martins, C.; Pereira, I.; Rio-Torto, G.; Pereira, R.; Inverno, L.; Ferreira, C.; Sousa, S. & Chapouto, S. (no prelo). *Building a spoken learner corpus of Portuguese L2: objectives and difficulties*.

Flowerdew, J. (2009). Corpora in Language Teaching. In M. H. Long, & C. J. Doughty (Eds.), *The Handbook of Language Teaching* (pp. 327-350). Blackwell.

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 University Press.

Golato, A. (2017). Naturally Occurring Data. In A. Barron, Y. Gu, & G. Steen (Eds.), *The Routledge Handbook of Pragmatics* (pp. 21-26). Routledge.

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). John Benjamins.

Granger, S. (2008). Learner corpora in foreign language education. In N. Van Deusen-Scholl, & N. H. Hornberger (Eds.), *Encyclopedia of Language and Education* (Vol. 4, pp. 337-351). Springer.

Granger, S., Dagneux, E., & Meunier, F. (2002). *The International Corpus of Learner English. Version 1.1*. Handbook and CD-ROM. Presses Universitaires de Louvain.

Hondo J. (2014). Synthesizing Social and Cognitive Approaches in SLA. *Working Papers in Educational Linguistics*, *29*(2), 1-23.

Johansson, S. (2007). Using Corpora: from learning to research. In E. Hidalgo, L. Quereda, & J. Santana (Eds.), *Corpora in the Foreign Language Classroom: Selected Papers from the Sixth International Conference on Teaching and Language Corpora (TaLC 6)* (pp. 17-30). Rodopi.

Jucker, A. H. (2018). Data in pragmatic research. In A. H. Jucker, K. P. Schneider, & W. Bublitz (Eds.), *Methods in Pragmatics* (pp. 3-36). De Gruyter Mouton.

Kramsch, C. (1986). From language proficiency to interactional competence. *The Modern Language Journal, 70*(4), 366-372.

Leech, G. (1997). Teaching and language corpora: A convergence. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (Eds.), *Teaching and Language Corpora* (pp. 1-23). Longman.

Long, M. H. (1981). Input, interaction, and second language acquisition. *Annals of the New York Academy of Sciences*, *379*(1), 259-278.

Lozano, C., & Mendikoetxea, A. (2013). Learner corpora and Second Language Acquisition: The design and collection of CEDEL2. In A. Díaz-Negrillo, N. Ballier, & P. Thompson (Eds.), *Automatic Treatment and Analysis of Learner Corpus Data* (pp. 65-100). John Benjamins.

Lozano, C., Teixeira, J., & Madeira, A. (2021). Corpora and L2 acquisition: the L1 Portuguese – L2 Spanish subcorpus of CEDEL2. *Revista da Associação Portuguesa de Linguística*, (8-10), 137-154.

Luzón, M. J., Campoy, M. C., Sánchez, M. M., & Salazar, P. (2007). Spoken Corpora: New Perspectives in Oral Language Use and Teaching. In M. C. Campoy, & M. J. Luzón (Eds.), *Spoken Corpora in Applied Linguistics* (pp. 3-26). Peter Lang.

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

MacWhinney, B. (2021). TalkBank for SLA. In N. Tracy-Ventura, & M. Paquot, (Eds.), *The Routledge Handbook of SLA and Corpora* (pp. 158-172). Routledge.

McEnery, T., & Wilson, A. (2001). *Corpus Linguistics. An Introduction*. Edinburgh University Press.

McEnery, T., Brezina, V., Gablasova, D., & Banerjee, J. (2019). Corpus Linguistics, Learner Corpora, and SLA: Employing Technology to Analyze Language Use. *Annual Review of Applied Linguistics*, *39*, 74-92.

Mendes, A., Antunes S., Janssen, M., & Gonçalves, A. (2016). The COPLE2 Corpus: A Learner Corpus for Portuguese. In *Proceedings of the Tenth Language Resources and Evaluation Conference – LREC'16* (pp. 3207-3214). European Language Resources Association.

Mondada, L., & Doehler, S. P. (2004). Second Language Acquisition as Situated Practice: Task Accomplishment in the French Second Language Classroom. *Modern Language Journal, 88*(4), 501-518.

Mukherjee, J. (2006). Corpus linguistics and language pedagogy: the state of the art – and beyond. In S. Braun, K. Kohn, & J. Mukherjee (Eds.), *Corpus Technology and Language Pedagogy* (pp. 5-24). Peter Lang.

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

Nascimento, M. F. B. (2002). O lugar do corpus na investigação linguística. In A. Mendes, & T. Freitas (Orgs.), *Actas do XVIII Encontro da Associação Portuguesa de Linguística* (pp. 601-605). APL.

Orletti, F. (1984). Some methodological problems in data gathering for discourse analysis. *Journal of Pragmatics*, *8*(4), 559-567.

Philp, J., Adams, R., & Iwashita, N. (2014). *Peer interaction and second language learning*. Routledge.

Richards, J. C. (2006). *Communicative Language Teaching Today*. Cambridge University Press.

Romero-Trillo, J. (2017). Editorial. *Corpus Pragmatics*, *1*(1), 1-2.

Romero-Trillo, J. (Ed.). (2008). *Pragmatics and corpus linguistics. A mutualistic entente*. Mouton de Gruyter.

Rühlemann, C. (2010). What can a corpus tell us about pragmatics? In A. O'Keeffe, &

M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 288-301). Routledge.

Rühlemann, C., & Clancy, B. (2018). Corpus linguistics and pragmatics. In C. Ilie, & N. Norrick (Eds.), *Pragmatics and its Interfaces* (pp. 241-266). John Benjamins.

Santos, I. A., Pereira, I., Martins, C., Lopes, A. C. M., Carapinha, C., & Silva, A. (2016). *Corpus* oral de PL2: um novo recurso para a investigação e ensino. *Revista da Associação Portuguesa de Linguística*, (1), 745-760.

Taguchi, N., & Roever, C. (2017). *Second language Pragmatics*. Oxford University Press.

VanPatten, B., Keating, G. D., & Wulff, S. (2020). *Theories in Second Language Acquisition: An Introduction* (3.ª ed.). Routledge.

Vaughan, E., & Clancy, B. (2013). Small Corpora and Pragmatics. In J. Romero-Trillo (Eds.), *The Yearbook of Corpus Linguistics and Pragmatics* (Vol. 1, pp. 53-73). Springer.

Walsh, S. (2010). What features of spoken and written corpora can be exploited in creating language teaching materials and syllabuses? In A. O'Keeffe, & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 333-344). Routledge.

Young, R. F. (2011). Interactional Competence in Language Learning, Teaching and Testing. In E. Hinkel (Ed.), *Handbook of Research in Second Language Teaching and Learning* (Vol. 2, pp. 426-443). Routledge.