Automatização no diagnóstico de nível de língua: anotação e versatilidade dos recursos para PLE

Raquel Amaro[*], *Susana Correia*[*], *Carolina Gramacho*[♦] & *Amália Mendes*[♦]

[*]CLUNL – Centro de Linguística da Universidade NOVA de Lisboa, [♦]CLUL – Centro de Linguística da Universidade de Lisboa

Abstract:

The automatic diagnosis and analysis of the production of foreign language learners can contribute to overcome linguistic barriers that hinder the integration of migrant populations. The richness and complexity of the phenomena observed in this context and the multiplicity of objectives served by automatic analysis tools demonstrate the inevitability of manual annotation of data and the importance of producing versatile resources, in order to maximize their usability. The present study aims, therefore, to contrast the needs of automatic diagnosis systems and the analysis of the phenomena reflected in the annotations for European Portuguese, based on COPLE2 and the *corpora* analysis conducted within the scope of the POR Nível project, proposing a system of annotation that includes error annotation and annotation of structures associated with complexity. The results highlight the need to enhance the usability of resources, to acknowledge their value and to promote the necessary investment in their development.

***Keywords*:** Portuguese as a FL/L2 *corpora*, COPLE2, POR Nível, FL/L2 error annotation, automatization of proficiency diagnosis

***Palavras-chave*:** *corpora* de Português LE/L2, automatização para diagnóstico de nível de língua, COPLE2, POR Nível, anotação de erro de LE/L2, automatização no diagnóstico de nível de língua

1. Introdução

O diagnóstico e a análise automáticos da produção de aprendentes de língua estrangeira são atualmente um tópico de investigação muito relevante, na medida em que permitem responder diretamente e de forma mais imediata a necessidades de aprendentes de língua, nomeadamente as que decorrem de situações de migração e integração de populações. As técnicas de automatização deste diagnóstico e análise estão em pleno desenvolvimento desde há já alguns anos (Meurers, 2009) e inserem-se, *grosso modo*, em dois grandes grupos – análise de erro e análise de complexidade –, servindo-se de sistemas mais ou menos complexos do ponto de vista computacional e de processamento (Ripley, 2009; Amaral *et al*., 2006; Curto *et al*., 2014; Chen & Meurers, 2019; del Río, 2019).

As abordagens por análise de erro visam identificar o nível de proficiência linguística de um dado falante pela presença ou ausência de erros nas suas produções, com base em mapeamentos entre erros/acertos expectáveis e/ou típicos de um dado nível, tendo como referência, por exemplo, documentos orientadores, como o *Quadro Europeu Comum de Referência para as Línguas* ou o *Referencial Camões de Português Língua Estrangeira*, para o caso específico do português (Talhadas, 2016; Gramacho *et al.*, 2019). As abordagens por análise de complexidade procuram nos textos características de vários níveis (lexical, sintagmático, frásico, textual) e, de acordo com a presença destas, caracterizam-nos como mais ou menos complexos e, logo, mais ou menos adequados ao nível de proficiência do aprendente (DuBay, 2004; Vajjala & Meurers, 2012; Curto *et al*., 2014; Weiss & Meurers, 2018).

A automatização destas abordagens pode ser baseada em regras predefinidas, a partir das análises e mapeamentos acima referidos e com custos de programação mais elevados, ou baseada em métodos de

aprendizagem automática a partir de dados. No entanto, quer as técnicas baseadas em métodos de aprendizagem automática (supervisionada ou não) que extraem/identificam traços relevantes a partir de dados anotados, quer as técnicas de análise multidimensional de vetores de medida de complexidade linguística com base em regras, pressupõem a análise e a anotação manual de dados, seja para construir os *corpora* de treino e teste necessários aos sistemas de aprendizagem automática, seja para testar e indiretamente informar os sistemas de análise no que respeita aos vetores de complexidade.

Apesar disso, os fenómenos e as tipologias de anotação necessárias são muitas vezes distintos, como brevemente se exemplifica abaixo:

(i) fenómenos e tipologias consideradas para anotação de erro: p. ex., no domínio da ortografia: nasalidade, acentuação; da morfossintaxe: flexão verbal, concordância nominal (projeto POR Nível, CLUNL; COPLE2, Mendes *et al.*, 2016).
(ii) fenómenos e tipologias consideradas para anotação de complexidade: p. ex., dimensão média da oração; constituintes coordenados por oração; nomes complexos por oração; rácio de orações subordinadas (projeto SyB, EKUT).

Por sua vez, a análise de dados linguísticos de aprendizagem de língua estrangeira, e respetiva compilação e construção de *corpora* de aprendizagem, pela riqueza e complexidade dos fenómenos que abrangem e pela multiplicidade de objetivos que servem, são em si temas de estudo produtivos e, mais importante ainda, dependentes da(s) língua(s) em análise (por exemplo, Alexandre & Pinto, 2014; Alexandre & Gonçalves, 2015; Antunes & Mendes, 2015; Cabrera & Zubizarreta, 2005; Castelo *et al.*, 2015; Mendes *et al.*, 2016, Talhadas, 2016). Por isso, é essencialmente a partir desta investigação que os sistemas de anotação são desenhados (Tono, 2003; Nicholls, 2003; Dagneaux *et al*., 2005).

Esta conjugação de fatores demonstra-nos claramente dois aspetos interdependentes:

(i) a inevitabilidade da anotação humana dos dados, que por si é um processo moroso e dispendioso, mas pouco atrativo no que respeita à captação de financiamento para a sua realização e desenvolvimento;
(ii) a importância da versatilidade dos recursos criados, de modo a maximizar a sua usabilidade e o investimento realizado.

Neste contexto, a análise dos sistemas de anotação de *corpora* de aprendizagem e das necessidades dos sistemas automáticos é de grande relevância, quer para promover a melhor anotação possível, dado (i), quer para assegurar a viabilidade de todo o processo, pelo referido em (ii). Para tal, é necessário perceber que formato terá uma anotação que permita viabilizar as abordagens e as técnicas em prática, ou seja, como desenhar sistemas de anotação de modo a que estes permitam a anotação de erro (anotação *negativa*, i.e. identificação de estruturas não aceites) e de estruturas associadas à complexidade (anotação *positiva*, *i.e.* identificação de estruturas bem formadas) e que permitam também associar os dados de produção a níveis de proficiência, de modo a permitir o diagnóstico de nível e a adequação de materiais/leituras aos diferentes níveis de proficiência.

O presente artigo visa, assim, o contraste das necessidades dos sistemas de diagnóstico automático e a análise dos fenómenos refletidos nas atuais anotações para o português, tendo como base o COPLE2 (Mendes *et al*., 2016) e os resultados da análise conduzida no âmbito do projeto POR Nível (Gramacho *et al*., 2019), de modo a discutir um sistema de anotação que contemple a anotação de erro, mas também possibilite a anotação de estruturas associadas à complexidade. Para além de sistematizar os fenómenos em causa em ambas as estratégias de diagnóstico, o trabalho pretende também potenciar a usabilidade dos recursos, valorizando-os e fomentando o tão necessário investimento no seu desenvolvimento.

A secção seguinte foca os *corpora* e os dados necessários para permitir o desenvolvimento de sistemas de diagnóstico automático, descrevendo as estratégias e os requisitos dos sistemas de diagnóstico em termos de análise e anotação linguística, por um lado, e os sistemas usados no COPLE2 e no projeto POR Nível, considerando os seus objetivos, metodologias de anotação e *tagsets*, por outro. A secção 3 apresenta o contraste dos sistemas de anotação e a análise dos fenómenos apresentados, com vista à integração dos dados e à prossecução de um sistema de anotação compreensivo que inclua a anotação de erros e de estruturas-alvo. Finalmente, a secção 4 apresenta as nossas considerações finais.

2. *Corpora* e dados para sistemas de diagnóstico automático

Os *corpora* de língua estrangeira (LE) ou língua segunda (L2) são *corpora* especiais, no sentido em que são constituídos por dados de língua selecionados e organizados de acordo com critérios linguísticos explícitos, de modo a poderem ser usados como amostras representativas (Sinclair 1996:4), não de uma dada língua em geral (*corpora* de referência/*corpora* monitores), mas das produções de falantes não nativos, aprendentes da língua. Dados empíricos deste tipo servem diversos objetivos de investigação, tais como estudos sobre aquisição de LE/L2, multilinguismo, fenómenos de interferência linguística ou análise e diagnóstico de proficiência. A sua constituição, recolha e tratamento, no âmbito da disponibilização de recursos linguísticos, é em si mesma uma área de investigação, na medida em que os critérios de seleção e amostragem deste tipo de dados, bem como a sua compilação, codificação e anotação refletem necessariamente as características intrínsecas dos dados, por um lado, e os objetivos de investigação que os *corpora* visam servir, por outro (ver Díaz-Negrillo & Thompson, 2013; Granger *et al*., 2015, entre muitos outros).

Como tal, a constituição de *corpora* enquanto recursos linguísticos de referência e com potencial de representatividade não se limita à recolha e armazenamento de dados disponíveis (Biber, 1993; McEnery & Hardie, 2012). As várias etapas de trabalho implicam decisões e motivações de ordem teórica, metodológica e prática sobre aspetos como:

(i) a dimensão e organização dos dados: dimensão total do *corpus*, critérios de estratificação do universo disponível (ex.: níveis de proficiência/ensino e língua materna), distribuição dos dados/nível, etc.
(ii) a codificação[1] dos dados: informações a registar (ex.: idade do falante, língua nativa do falante, nível de escolaridade, tempo de permanência no território de acolhimento e produção livre/produção induzida), formatos de codificação (ex.: cabeçalhos dos ficheiros e ficheiros autónomos de metadados), etc.
(iii) a anotação[2] dos dados: tipos e níveis de anotação (ex.: morfossintática, sintática, semântica e erros), nível de especificidade das anotações (ex.: classe de palavras, flexão, grau e erro de concordância sujeito-verbo), formato e organização das anotações (ex.: simultâneas, numa mesma cadeia/etiqueta, *e*x.: *Nc#ms_SUJ*, e paralelas), etc.
(iv) os formatos de codificação e interfaces: TEI - *Text Encoding Initiative* e TEITOK - *Tokenized TEI Environment*, CPQ - *Corpus Query Processor*, CPQWeb, etc.

O potencial de utilização destes dados está diretamente relacionado com as suas particularidades e, naturalmente, com o seu tratamento e análise. Por exemplo, para o diagnóstico de proficiência, a anotação de

[1] A codificação diz respeito ao registo de propriedades intrínsecas aos dados. Podem ser relativos à situação de produção/comunicação (ex.: data de produção, autor/falante, idade do autor/falante, local de recolha, etc.), ou relativos a características do texto (ex.: mudança inesperada de linha, linha em branco, texto rasurado, texto em maiúsculas, etc.).
[2] A anotação diz respeito à caracterização de propriedades não intrínsecas aos textos, sejam elas mais ou menos implícitas, como a classe de palavras e informação morfossintática, sejam elas de ordem mais subjetiva, como marcação de ironia, metáforas, etc. No caso de *corpora* de LE/L2 são exemplos típicos de anotação a marcação do erro e a proposta de correção desse mesmo erro.

erros/desvios pode ser tão interessante como a anotação de acertos, tal como a análise quantitativa dos erros, associada a uma análise qualitativa de cada erro, pode ser um fator crucial de diagnóstico. Por exemplo, um texto sem erros de construção frásica, mas apenas com frases simples e curtas, pode não refletir um nível mais alto de proficiência do que um texto com erros de construção frásica, mas com frases complexas e constituintes deslocados. Neste contexto, é relevante perceber se os *corpora* de LE/L2 existentes para o português europeu poderão ser usados em sistemas de diagnóstico automático e que tipo de anotação parece ser necessária e/ou relevante para estes sistemas.

Nas secções abaixo descrevemos o que pode implicar um sistema de diagnóstico de proficiência automático e apresentamos 2 *corpora* de LE/L2 e respetivos sistemas de anotação para perceber a sua usabilidade nesses sistemas.

2.1. Sistemas de diagnóstico automático

Os sistemas de diagnóstico automático visam o diagnóstico do nível de proficiência linguística de forma totalmente automática e tendencialmente imediata de modo a servir objetivos diversos (Chapelle & Ghung, 2010), tais como:

(i) a análise e/ou validação *bottom-up* automática (isto é, a extração de características associadas a perfis de aprendizagem e de proficiência linguística a partir da análise automática de dados de LE/L2);
(ii) o auxílio a sistemas de ensino/aprendizagem de LE/L2 (CALL – *Computer-Aided Language Learning* e ICALL – *Intelligent Computer-Aided Language Learning*);
(iii) a deteção de produções de falantes não-nativos.

Tal como atualmente em várias outras áreas, os sistemas de diagnóstico automático usam, de modo disseminado, sistemas de aprendizagem automática. Em termos metodológicos, estes sistemas são concebidos e afinados considerando 3 aspetos essenciais: i) os dados; ii) o processo de aprendizagem automática; iii) a comparação/avaliação dos resultados.

A aprendizagem automática consiste numa estratégia de processamento para, a partir de dados, inferir padrões ou funções. Num sistema de aprendizagem automática supervisionado, é fornecida ao sistema uma amostra dos dados e dos resultados esperados. Por outras palavras, o sistema tem acesso a dados anotados/classificados e tem, então, de "aprender" qual a melhor função para aproximar os dados de *input* aos dados de *output*, com base nos dados de referência/resultados esperados. Num sistema de aprendizagem automática não supervisionada não há dados de referência ou resultados esperados; o sistema tem de inferir a estrutura natural presente em dados não anotados.

A estratégia mais usada nos sistemas de diagnóstico automático é a aprendizagem automática supervisionada que, com base em *corpora* anotados, divididos em *corpus* de treino e *corpus* de teste, aplica algoritmos e modelos estatísticos, sem instruções explícitas, mas com poder de inferência, para determinar qual a melhor função para chegar aos resultados esperados a partir de padrões/traços inferidos dos dados de treino. Após o treino, os resultados são então medidos relativamente aos dados de teste, para determinar quais os algoritmos de aprendizagem automática mais eficazes nos casos em estudo.

Dependendo dos métodos usados, os dados necessários para desenvolver os sistemas poderão ser diferentes. Por outro lado, dependendo das informações anotadas nos dados, o desempenho dos sistemas poderá também ser diferente. É neste contexto que a constituição e anotação de *corpora* de LE/L2 se torna tão relevante. É possível dizer-se que os sistemas de aprendizagem automática não têm regras de conhecimento linguístico explícitas, mas usam conhecimento linguístico implícito, refletidos nos padrões dos dados de língua, e explícitos, expressados nas anotações. Considerando o funcionamento dos sistemas e as particularidades dos

dados, foram sendo desenvolvidas várias estratégias de análise para chegar ao diagnóstico automático, brevemente descritas na secção abaixo.

2.1.1. Estratégias e traços

Podemos distinguir três grandes tipos de estratégias usadas no diagnóstico automático de proficiência, associadas a vários métodos de análise, por um lado, e a diferentes traços (*features*), por outro. Os traços dizem respeito a características ou funções/correlações que os sistemas de aprendizagem automática consideram relevantes para a determinação ou inferência de padrões e que, tipicamente, estão diretamente relacionados com as informações codificadas nos dados (ex.: número de verbos, concordância no sintagma nominal, etc.), mas não necessariamente (ex.: frequência de nomes, número de palavras por frase, etc.).

A tabela abaixo sistematiza as três estratégias: análise quantitativa, geral, também designada de linguística, análise de complexidade e coerência, e análise de erro.

Análise	***Métodos***	***Traços***
quantitativa/ geral/ linguística	*bag-of-words*/*n*-gramas	frequência, nº palavras/frase, diversidade lexical, ...
	n-gramas com POS	concordância no SN, sujeito-V contíguos, subcategorias, ...
	n-gramas c/ dependências	concordância no SN e sujeito-V, subcategorização, ...
complexidade (métricas)	descritiva	nº de sílabas, rácio sílabas/palavra, nº de palavras, nível de legibilidade
	morfológica	nº de verbos, nº de nomes, nº de formas, ...
	lexical	rácio lema/forma, nº de palavras lexicais
coerência	sintática	nº de SN; nº de modificadores por SN, ...
	Coh-Metrix[3]	
erro	linguístico	gramatical (verbo, tempo, morfema)
	modificação	omissão, adição, troca, ...

Tabela 1. Estratégias de análise para diagnóstico automático de proficiência

A estratégia de análise quantitativa/geral/linguística pode ser mais ou menos sofisticada usando métodos de representação e análise de texto, como conjuntos múltiplos de palavras, sem informação acerca de posição relativa, ordem ou gramatical (*bag of words*); conjuntos múltiplos de palavras em que duas ou mais palavras são tomadas como elementos do conjunto, permitindo informação acerca de ordem relativa/posição (*n*-gramas), *n*-gramas com informação morfossintática (*Parts Of Speech* - POS), mais ou menos fina que, que para além de permitirem analisar a posição relativa ou a ordem, permitem analisar fenómenos como concordância determinante-Nome ou Nome-Adjetivo, concordância Pronome-Verbo ou Nome-Verbo (*n*-gramas com POS); *n*-gramas com mais de um elemento em que há relações de dependência sintática tal como núcleo-modificador; núcleo-sujeito; núcleo-objeto (*n*-gramas com dependências). Nesta estratégia, os traços podem ser de natureza mais ou menos linguística (por exemplo, concordância sujeito-verbo ou número de palavras/frase, respetivamente).

[3] http://www.cohmetrix.com/

A estratégia de análise de complexidade ou coerência usa típica e simultaneamente vários métodos, focando aspetos lexicais, sintáticos e não linguísticos, como número de palavras por frase ou número de sílabas por palavra. Estas características são emanadas de análises linguísticas, orientações de escrita clara e guias de simplificação de textos escritos para maior legibilidade. Um dos métodos mais usados é o *Coh-Metrix* (Graesser *et al.,* 2004), que lista 108 traços, agrupados em:

1) traços descritivos (ex.: número de parágrafos e número de letras/palavra-desvio médio)
2) cálculos e percentis de facilidade/simplicidade textual (ex.: narratividade, simplicidade sintática, concretude das palavras, coesão referencial e temporalidade)
3) coesão referencial (ex.: sobreposição de nomes em frases adjacentes e sobreposição de anáforas em frases adjacentes)
4) medidas de LSA (*Latent Semantic Analysis*, Landauer *et al.,* 2007) (ex.: sobreposição semântica em frases adjacentes e sobreposição semântica em parágrafos adjacentes)
5) diversidade lexical (ex.: rácio *de types*/*tokens*[4] e palavras lexicais)
6) conectores (ex.: incidência de conectores causais/temporais/lógicos, coesão temporal e repetição de tempo e aspeto)
7) complexidade sintática (ex.: número de frases encaixadas antes de verbo principal, similaridade sintática entre frases adjacentes e número de modificadores em SN)
8) densidade de padrões sintáticos (ex.: densidade de SV, densidade de SAdv e incidência de negação)
9) informação lexical (ex.: incidência de nomes, verbos, adjetivos, polissemia de verbos e hiperónimos verbais)
10) legibilidade (ex.: *Flesch Reading Ease* - Flesch, 1948, *Flesch-Kincaid Grade Level*, Kincaid *et al.*, 1975).

O método *Coh-Metrix* utiliza todos os traços disponíveis, propondo várias medidas, análises e combinações de traços simples, tais como percentis, médias e desvios padrões de traços. É um sistema feito para o inglês e já adaptado a várias línguas, incluindo o português do Brasil (Scarton & Aluísio, 2010).

Finalmente, a terceira estratégia foca a análise e a anotação dos erros, considerando dois métodos essenciais: a identificação e anotação de erro linguístico, por um lado, e a modificação inesperada do texto, considerando omissões, adições e trocas, por outro. Este método será descrito em maior pormenor nas secções seguintes.

2.1.2. Requisitos típicos

Esta breve descrição das estratégias e propriedades visadas pelos sistemas de diagnóstico automático permite-nos perceber que estes implicam necessariamente:

(i) a disponibilidade de grandes quantidades de dados linguísticos, de modo a abranger um leque suficiente de fenómenos e em número representativo;
(ii) a análise de dados linguísticos que cubra os traços considerados relevantes para o diagnóstico;
(iii) a definição e operacionalização de metodologias e sistemas de anotação que permitam dar conta de traços de vários níveis, cobrindo unidades de dimensão variável (do morfema ao parágrafo);
(iv) trabalho de anotação/revisão manual dos dados.

[4] *Types* correspondem *grosso modo* a lemas, sem diferenciação semântica (ex.: as formas *rato*, *ratos* e *ratinho* correspondem ao *type* 'rato'); *tokens* correspondem às formas que objetivamente ocorrem no texto (ex.: a frase "O gato comeu os ratos e os ratinhos" contém oito *tokens* e cinco *types* (o, gato, comer, e, rato).

Como descrito acima, a constituição de *corpora* contempla, desde logo, tarefas de normalização, codificação e anotação morfossintática, sintática e semântica, independentemente dos dados linguísticos em causa. No entanto, a constituição e tratamento de *corpora* de LE/L2, pelas suas propriedades intrínsecas, em particular a presença de erros, não permite a aplicação direta de anotadores concebidos e treinados para outros segmentos da língua, para além de implicar a anotação de erro e a anotação da respetiva correção, se possível.

Nas secções seguintes iremos descrever dois exemplos de análise de dados de aprendentes do português europeu LE/L2, tratados em *corpora* de LE/L2 formais, no caso do COPLE2, e focando a análise e anotação de erro para diagnóstico de proficiência, no caso do projeto POR Nível, com vista ao seu contraste e integração e de modo a perceber se respondem aos requisitos dos sistemas de diagnóstico automático.

2.2. COPLE2: um *corpus* de LE/L2

2.2.1. Objetivos

O COPLE2 – *Corpus de Português Língua Estrangeira / Segunda* é um conjunto de materiais escritos e orais produzidos por alunos estrangeiros que estão a aprender português como LE ou L2 (PLE/L2), bem como por candidatos a exames de certificação de proficiência de nível de língua. A recolha dos materiais é feita na Faculdade de Letras da Universidade de Lisboa (FLUL), no âmbito dos cursos de Português Língua Estrangeira, ministrados pelo Instituto de Cultura e Língua Portuguesa, e dos exames de acreditação realizados pelo Centro de Avaliação de Português Língua Estrangeira (CAPLE) (Mendes *et al*., 2016; del Río & Mendes, 2018a).

O COPLE2 constitui um recurso importante para o estudo do ensino e da aprendizagem do PLE/L2. Está disponível gratuitamente *online* e apresenta informação detalhada ao nível dos metadados (do informante e do texto produzido), bem como diversos tipos de anotação linguística (del Río & Mendes, 2018b). O sistema de opções de pesquisa possibilita combinar os diferentes tipos de variáveis, facilitando a busca de fenómenos específicos.

A inclusão de uma grande variedade de línguas maternas (L1) permite a realização de estudos com base na Análise Contrastiva Interlíngua (Granger, 1996): comparando dados produzidos por falantes com diversas línguas maternas, torna-se mais fácil avaliar se alguns dados são influenciados pela L1 dos informantes ou se são produzidos por outros aprendentes, em geral, independentemente da L1, despistando, assim, falsos fenómenos de transferência (Jarvis, 2000; Paquot, 2013). Deste modo, este *corpus* visa fornecer dados acessíveis a professores e/ou investigadores que permitam realizar trabalhos de natureza linguística variada, tais como a identificação de erros gerais na aprendizagem de PLE/L2, ou a identificação de erros específicos que possam resultar de transferências da língua materna ou de outras línguas estrangeiras previamente adquiridas. Estudos desta natureza possibilitarão o desenvolvimento de aplicações e materiais didáticos na área do ensino de PLE/L2, adequando estratégias de ensino a um público-alvo específico.

2.2.2. Metodologia na anotação do erro

O conjunto de etiquetas definido (*tagset*) de anotação do erro encontra-se hierarquizado em dois níveis de informação:

(i) a área linguística em que ocorre o erro: ortografia, gramática e léxico;
(ii) o tipo de erro específico para cada área: a ortografia contém 11 etiquetas, e inclui erros relacionados com a própria ortografia da palavra, a pontuação, a acentuação, a fronteira de palavra e o uso de maiúsculas e minúsculas; a gramática contém 24 etiquetas, e inclui erros relacionados com a concordância em género ou número, o modo e tempo verbais, a omissão de palavras, a categoria gramatical, a ordem das palavras, a sufixação, o processo de cliticização, entre outros; o léxico contém 2 etiquetas que dizem respeito ao uso de uma palavra num contexto errado e ao uso de uma palavra inexistente em português (37 etiquetas no total).

Os dois níveis de informação são expressos em etiquetas de tipo *position-based* (i.e., cada categoria e subcategoria ocupa sempre a mesma posição na etiqueta de anotação), em que a primeira letra corresponde à área geral apresentada em (i), e as restantes letras aos tipos de erro referidos em (ii). O processo de anotação funciona em duas etapas. Na primeira etapa, procede-se à normalização do erro nos três níveis linguísticos e à atribuição do lema e etiqueta morfossintática corretos. Em (1)-(3) exemplifica-se o processo de normalização para os três níveis.

(1) Normalização no nível ortográfico: o aluno produz a forma 'vigilância', que é normalizada para 'vigilância', com a atribuição do lema 'vigilância' e da etiqueta morfossintática NFS (nome, feminino, singular).
(2) Normalização no nível gramatical: o aluno produz 'o televisão', em que a forma 'o' é normalizada para 'a', com a atribuição do lema 'a' e a etiqueta morfossintática DAFS (determinante, definido, feminino, singular).
(3) Normalização no nível lexical: o aluno produz 'quando nós visitaramos nestes dois passeios', em que a forma 'visitaramos' é normalizada para 'fizemos', com a atribuição do lema 'fazer' e a etiqueta morfossintática VMIS1P (verbo, principal, indicativo, passado, 1ªp, plural), sendo a forma 'nestes' normalizada para 'estes'.

Na segunda etapa de anotação, identifica-se o fenómeno específico de cada erro. Partindo dos exemplos apresentados em (1)-(3), a forma 'vigilância receberá a etiqueta SGS (*Spelling Grapheme Substitution* – erro ortográfico em que um grafema é usado em vez de outro); a forma 'o' receberá a etiqueta GAG (*Grammar Agreement Gender* – erro de concordância que afeta o género); a forma 'visitaramos' receberá a etiqueta LC (*Lexical Choice* – a palavra existe em português, mas está a ser usada num contexto errado).

Os erros que abranjam mais do que um *token* ('Cada vez que falo sobre esta história, não posso controlar o meu próprio que ser bravo') são anotados em *stand-off*, isto é, em vez de adicionar uma etiqueta de erro no texto, a anotação é guardada num ficheiro independente e é indexada com o ficheiro de texto.

Para a obtenção desta versão do conjunto de etiquetas, realizou-se uma primeira experiência de anotação de 36 textos (7.073 *tokens*), que permitiu determinar as áreas linguísticas que deviam ser tratadas, bem como a otimização das etiquetas de anotação. Foi, também, realizada uma experiência de anotação para avaliar a concordância/divergência entre anotadores (*inter-annotator agreement*). Para o efeito, duas anotadoras anotaram 10 textos do *corpus*, tendo os resultados sido avaliados com a medida *Cohen's kappa*, obtendo-se o resultado $\kappa = 0.85$, considerado muito positivo. A análise dos aspetos divergentes na anotação permitiu um aperfeiçoamento das normas de anotação.

2.2.3. *Tagset* de anotação de erro

Categorias linguísticas		***Exemplos de erros e normalização***[5]
Ortografia	Spelling_StressMark	*diferentes paises e povos -> países*
	Spelling_Grapheme_Addition	*practicamente -> praticamente*
	Spelling_Grapheme_Deletion	*qerem -> querem*
	Spelling_Grapheme__Substitution	*spportei -> suportei*
	Spelling_Grapheme_Transposition	*apíses -> países*

[5] Os exemplos são direta e integralmente retirados dos *corpora* em análise. No caso do COPLE2, as propostas de normalização apresentadas são baseadas na proposta do professor que primeiro avaliou a produção do aprendente (informação alvo de codificação) podendo ser completadas e/ou alteradas pelo anotador, com base numa análise global do texto nem sempre completamente visível no excerto apresentado.

	Spelling_Capitalization	*a liberdade de que fala pessoa -> Pessoa*
	Spelling_WordBoundarySplit	*última mente não falamos -> ultimamente*
	Spelling_WordBoundaryMerged	*fimdesemana -> fim de semana*
	Spelling_PunctConfused	*a quem lhe ouvir. por exemplo -> ouvir, por*
	Spelling_PunctRedundant	*Assim, foi a minha modesta leitura -> Assim foi*
	Spelling_PunctMissed	*é tal grande que às vezes > que, às vezes*
Gramática	Grammar_UnnecessaryWord	*eu vou a organizar uma festa -> vou organizar*
	Grammar_OmittedWord	*fala do dia a dia do cidadão -> a fala*
	Grammar_WrongWord	*numa altura em que ninguém sem tempo por nada -> tem*
	Grammar_WrongCategory	*não vive nas selvagens com tantos riscos- > nas selvas*
	Grammar_WrongStructure	*A industria de seviços (hotel, restaurante, transporte, etc)*
	Grammar_Agreement_Gender	*os ideáis humanitarias -> humanitários*
	Grammar_Agreement_Number	*tem paisagens lindíssima -> lindíssimas*
	Grammar_Agreement_Gender&Number	*pode ser palavras bo -> boas*
	Grammar_Agreement_Person	*os portugueses não podia -> podiam*
	Grammar_WordOrder	*não lêem livros muitos -> muitos livros*
	Grammar_Verb_Tense	*Sempre havia e sempre havrá -> houve*
	Grammar_Verb_Mode	*uma battaria nova deva durar -> deve/deverá*
	Grammar_Verb_Tense&Mode	*senhor prometeu-me que irão funcionar -> iriam*
	Grammar_Verb_FiniteNoFinite	*cada vez mais melhorar*
	Grammar_Verb_Aspect	*o quarto estivo muito frio*
	Grammar_VerbalConstruction_Periphrasis	*E espero que não va acontecer*
	Grammar_VerbalConstruction_Clitization	*descobrimos-as -> -las*
	Grammar_PronounClitic_Case	*visitou-lhe ontem -> -a/o*
	Grammar_PronounClitic_Person	*Liguei ás meninas e disse-lhe que vieram*
	Grammar_PronounClitic_Position	*Ele não disse-me nada. -> não me disse nada*
	Grammar_Noun_Number	*Minha última féria esteve -> férias*
	Grammar_SuffixDerivation	*a legalizacion desse assunto -> legalização*
	Grammar_SuffixInflection	*e uma familia aberte às outras pessoas -> aberta*
Léxico	Lexical_LexicalChoice	*Se não tiver medidas de proteção > meios*
	Lexical_UnexistentWord	*e estabilitamos a melhor relação -> estabelecemos*

Tabela 2. *Tagset* para anotação de erro do *corpus* COPLE2

2.3. Análise de produções escritas de PLE/L2 para o projeto POR Nível

2.3.1. Objetivos

O POR Nível é um projeto que visa a construção e validação de um teste de colocação em nível de língua para aprendentes de PLE, do nível A1 ao nível C1 do *Quadro Europeu Comum de Referência para as Línguas*. Integrando as dimensões de gramática, vocabulário, compreensão oral e compreensão escrita, o seu desenvolvimento e aplicação têm como propósito, além da colocação do aprendente num nível de língua adequado, comparar as competências dos alunos de diferentes instituições de ensino e contribuir para o desenvolvimento de estratégias e métodos de ensino, servindo as necessidades curriculares dos alunos e as necessidades metodológicas e científicas daqueles que se dedicam ao ensino e à investigação na área do PLE/L2.

O teste de colocação em nível de língua que está a ser desenvolvido no âmbito do projeto POR Nível baseia-se em três tipos de dados:

(i) documentos orientadores: *Principles of Good Practices for ALTE Examinations (Out. 2001), ALTE's Manual for Language Test Development and Examination (Abr. 2011), Educational Testing System (ETS) Guidelines (2013), Cambridge English Principles of Good Practice (Maio 2011)*), perfis de língua (*Plan Curricular del Instituto Cervantes, English Profile, Référentiel de l'Alliance Française*) e referenciais nacionais (*Referencial Camões – PLE)* e europeus (*QECR)*;
(ii) testes de colocação validados desenvolvidos para outras línguas: o teste do *Goethe Institut* para o alemão, o teste do *Instituto Cervantes* para o espanhol, o teste da *Alliance Française* para o francês e o teste do *Cambridge e do British Council* para o inglês;
(iii) análise de *corpora* da produção escrita de aprendentes de PLE, com o objetivo específico de construir uma tipologia de erros e competências na produção escrita de aprendentes de PLE que informasse a construção dos itens do teste POR Nível.

2.3.2. Metodologia na anotação de *corpora*

Os materiais que deram origem ao *corpus* do projeto POR Nível consistiram em setenta e cinco textos dos géneros "carta" e "texto argumentativo", em registo formal e informal, retirados da parte de produção escrita de exames certificados do CAPLE, de candidatos nativos de inglês, mandarim e espanhol, três das línguas maternas mais representadas nos cursos de PLE em duas instituições de ensino superior nacionais (FLUL e Faculdade de Ciências Sociais e Humanas da Universidade Nova de Lisboa (NOVA FCSH)). Foram analisados vinte e cinco textos por língua (cinco textos por nível em cada língua), correspondendo a um total de 20.178 *tokens*.

As competências e dificuldades dos aprendentes foram codificadas manualmente numa análise em dois passos que consistiu:

(i) na identificação de erros e de usos convergentes com as estruturas-alvo, nos domínios de ortografia, morfossintaxe e vocabulário, tendo sido criado um *subcorpus* para cada uma das componentes referidas.
(ii) na categorização, quer dos erros, quer dos usos convergentes com as estruturas-alvo.

Os critérios que presidiram à definição das estruturas-alvo e dos bons usos têm como base o conteúdo curricular dos documentos orientadores mencionados em 2.3.1. e os programas dos cursos do Instituto de Cultura e Língua Portuguesa (FLUL) e do Curso de Língua e Cultura Portuguesa (NOVA FCSH).

Nas Tabelas 3 e 4, apresentam-se as categorias linguísticas que orientaram a nossa análise, bem como os erros e respetiva normalização encontrados nos *corpora* analisados para a construção dos itens que integram o

POR Nível. Nos *corpora* analisados, foi frequente a ocorrência de palavras ou estruturas nas quais existe mais do que um erro. Assim, estas tabelas incluem estruturas ou palavras em que pode haver mais erros do que o erro referido na tabela. Para além deste aspeto, nem sempre é possível restringir um erro a um determinado domínio da língua, pelo que a mesma palavra ou estrutura pode ocorrer em diferentes domínios, ou em diferentes categorias linguísticas. Por limitação de espaço, nas secções seguintes são apresentados dados que correspondem apenas a produções do nível A1[6].

2.3.2.1. Erros e normalização: domínios fonológico e ortográfico, morfossintático e lexical

No domínio fonológico-ortográfico, os passos (i) e (ii) descritos em 2.3.2. foram efetuados em paralelo, através das ferramentas *comparar* e *pesquisar* do editor Microsoft Word. Estas ferramentas permitiram-nos contabilizar, mediante cruzamento dos dados dos *corpora* normalizados e dos dados do documento original, o número exato de usos-alvo e de erros neste domínio. A tipologia utilizada para a categorização (feita de base a partir dos dados) e contabilização, quer dos erros, quer dos usos convergentes, apresenta-se na coluna *Categorias linguísticas*. Na coluna *Exemplos de erros e normalização* apresentam-se exemplos de erros para cada categoria e a respetiva a correção por nós efetuada com a ferramenta *rever* do editor Microsoft Word à qual adicionámos, sempre que julgámos pertinente, um comentário à natureza do erro.

Para o domínio morfossintático, o *subcorpus* foi construído através do uso paralelo das ferramentas *rever*, *registar alterações* e *novo comentário* do editor Microsoft Word. Procurámos apresentar para cada erro uma proposta de normalização e um pequeno comentário acerca da sua natureza.

No que respeita ao domínio lexical, embora o objetivo ideal deste *corpus* fosse analisar o domínio descrito para cada nível no *Referencial Camões*, a natureza dos textos em análise (limitados às temáticas em enunciado no exame do CAPLE) fez com que esta análise consistisse em identificar e isolar erros que tivessem como origem questões de transferência, decalque e problemas de aquisição de expressões fixas e idiomáticas.

A tabela abaixo apresenta, de forma sistematizada as categorias linguísticas utilizadas para a categorização dos erros, ilustradas com exemplos de erros encontrados e respetiva normalização.

	Categorias linguísticas		***Exemplos de erros e normalização***[7]
Fonologia/ Ortografia	correspondência grafema-fonema: consoantes		*duaz* (EN) -> *duas* *ingrês* (CH) -> *inglês*
	correspondência grafema-fonema: vogais e ditongos		*féiras* (ES) -> *férias* *tudos* (CH) -> *todos*
	marcação gráfica do acento		*más* (ES) -> *mas* *tambem* (CH) -> *também*
	marcação gráfica da nasalidade		*bõa* (ES) ->*boa* *televisao* (EN) -> *televisão*
Morfossintaxe	Tempo/Modo/Forma verbal	Tempo/Modo em contexto de uso de outro Tempo/Modo	*tenho a graduação da minha irmá e* ***tenho dito*** *que iría* (ES) -> *tinha dito* *para* ***venho*** *à Grécia* (ES) -> *ir*
		forma	*mas* ***e***[8] *a minha cadeira favorita* (ES) -> *é* *a minha mãe* ***dize****-me* (ES) -> *diz-me*
	seleção verbal		*não posso* ***vir*** *ao Instituto* (EN) -> *ir* *E tu? Onde* ***é****?* (CH) -> *estás*

[6] Os dados relativos aos restantes níveis podem ser fornecidos mediante pedido dirigido à equipa do projeto POR Nível.

[7] Os exemplos são direta e integralmente retirados dos *corpora* em análise. No caso do POR Nível, as propostas de normalização apresentadas são as registadas pelo anotador que avaliou a produção do aprendente, com base na análise global da tarefa, nem sempre completamente percetível pelo excerto apresentado.

[8] Importa referir que se considerou o facto de este ser um erro também relacionado com a marcação gráfica do acento.

	artigo	omissão	*Eu estou em Brasil* (ES) -> *no* *Há muitas praias* ***e temperatura*** *é* (EN) -> *e a temperatura* *também* ***seus*** *exames* (EN) -> *os seus*
		adição	*para a prova* ***d****o inglés* (ES) -> *de* *aulas* ***nas*** *todas 4ªs feiras* (EN) -> *aulas todas* *vou escrever-te* ***as*** *outras cartas* (EN) -> *escrever-te outras*
	concordância interna ao NP		*com* ***o meus novos oculos*** *de sol* (ES) -> *os* ***dois vezes*** *por semana* (EN) -> *duas* ***comeres tradicionais inglês*** (EN) -> *ingleses* ***meu amigos*** (EN) -> *meus* *um* ***dia outra*** (CH) -> *outro* *as pessoas são* ***bonitas e agradável*** (CH) -> *agradáveis*
	preposição	omissão	*porque tenho de assistir* ***o*** *casamento* (EN) -> *ao* *podemos mudar* ***um*** *dia outra* (CH) -> *para um*
		adição	***Em*** *na praia* (ES) -> *Na* *vou com a minha irmã* ***para*** *jogar futebol* (EN) -> *irmã jogar*
		preposição em contexto de uso de outra preposição	*podemos ir no 5ª feira* ***n****o salão* (ES) -> *ao* *tenho que ir* ***n****o medico* (ES) -> *ao* *vou ver-te* ***até*** *breve* (EN) -> *em* *posso procurar* ***à*** *internet* (CH) -> *na*
	preposição e outras categorias	contração	*a comida* ***de aqui*** (ES) -> *daqui* *gostas* ***de eles*** (ES) -> *deles* *estou* ***em o*** *Caribe* (ES) -> *no*
	pronome clítico (acusativo, dativo, reflexo, recíproco, intrínseco, *se* apassivante, *se* impessoal)	colocação	*Na praia* ***se está*** *muito bem* (ES) -> *está-se* *Se faz favor de* ***dizer-me*** (EN) -> *me dizer*
		omissão	*Peço favor* (EN) -> *Peço-lhe*
		adição	*cada um traze-****se*** *o seu dinheiro* (ES) -> *traz*
	pronome pessoal tónico		***Iu*** *estou* (EN) -> *Eu* *eu vejo* ***ele*** (EN) -> *vejo-o*
	nome	género	***na*** *bar* (CH) -> *no* ***próximo*** *semana* (EN) -> *próxima*
		número	***a féria*** (CH) -> *as férias*
	sistema dos demonstrativos	variáveis *vs*. invariáveis	***Este*** *é muito divertido* (ES) -> *Isto*
	sistema dos advérbios locativos	dêixis *vs*. anáfora	*Podes estar* ***ali*** *na 6ª feira* (ES) -> *lá estar/estar lá*
	ordem		***Também nós*** *comemos* (ES) -> *Nós também* *um* ***dia outra*** (CH) -> *outro dia* ***não também*** *podemos* (CH) -> *também não*
Léxico	decalque		*alumna* (ES) -> *aluna* *pelicula* (ES -> *filme* *Julio* (EN) -> *Julho*
	expressão fixas/fórmula de tratamento		***Imenso*** *Professor Sr. Santos* (CH) -> *Caro/Prezado*

Tabela 3. Produções com identificação de erros e normalizações nos *corpora* analisados para o POR Nível (ES – L1 espanhol; EN – L1 inglês; CH – L1 mandarim)

2.3.2.2. Identificação e marcação de bons usos

A identificação e categorização de usos convergentes com as estruturas-alvo identificadas contemplam os 3 domínios também usados na identificação de erros. A Tabela 4, abaixo, apresenta as categorias linguísticas identificadas, bem como exemplos de bons usos para cada uma delas.

Categorias linguísticas		***Exemplos de bons usos***
Fonologia/ ortografia	correspondência grafema-fonema: consoantes – os sons de <x>; o grafema <ç>	*e**x**ijo* (EN) *pe**ç**o* (CH)
	correspondência grafema-fonema: vogais e ditongos	*ár**ea*** (EN) *organiz**ei*** (CH)
	marcação gráfica do acento	*vigil**â**ncia* (ES) *respons**á**vel* (CH)
	marcação gráfica da nasalidade – a nasalidade antes de consoante	***com**plicado* (EN) *c**ã**es* (EN)
Morfossintaxe	preposição: regência nominal	*introdução de um imposto* (CH) *instalação de câmaras* (CH)
	Tempo Modo e Aspeto: uso do Condicional Simples uso do Pretérito Perfeito Composto do Conjuntivo	*gostaria de exprimir* (CH) *Ainda que o mundo tenha feito muitos progressos* (EN)
	concordância: concordância Sujeito-Verbo	*quando alguém conhece um novo destino* (ES) *que vocês me tivessem enganado* (EN)
	conjunções/conetores: *por isso; mas*	*por isso tenho de ir* (CH) *Desculpe, mas não posso* (CH)
Léxico	"falsos amigos"	*O autor deste **romance*** (ES)
	expressões fixas	***curto prazo*** (EN) ***deixar de lado** todos os outros filmes* (EN)
	vocabulário específico: vida académica; espaço urbano	***cadeira** favorita* (ES) *nos **arredores** das cidades* (EN)

Tabela 4. Produções com identificação de bons-usos nos *corpora* analisados para o POR Nível (ES – L1 espanhol; EN – L1 inglês; CH – L1 mandarim)

3. Para um sistema de anotação compreensivo: anotação de fenómenos e sobreposições

À primeira vista, os sistemas e métodos de anotação para o PLE/L2 descritos compreendem a anotação de fenómenos muito semelhantes, havendo grande sobreposição. Esta coincidência é natural e expectável tendo em conta os objetivos de investigação propostos e as opções metodológicas seguidas, em particular a análise de produções para identificação dos casos relevantes para a determinação dos sistemas e etiquetas de anotação. No que respeita à tipologia de fenómenos a considerar na anotação, a grande diferença entre as propostas parece estar na anotação de estruturas-alvo, ou bons usos, nos *corpora* analisados para o POR Nível.

Nesta secção iremos analisar a anotação dos fenómenos e a sobreposição dos sistemas propostos, com vista a um sistema de anotação compreensivo, que permita integrar os dados sem perda, ou com perda mínima, de informação.

3.1. Anotação de erros

A Tabela 5 abaixo sistematiza a informação coincidente e não coincidente entre ambos os sistemas, no que respeita à anotação de erros. As diferentes etiquetas e fenómenos foram agrupados por nível/domínio – i) fonológico e ortográfico, ii) gramatical (ou morfossintático) e iii) lexical, tal como indicado na primeira coluna. As divergências entre os sistemas são assinaladas pela cor de preenchimento das células, assinalando maior ou menor custo de integração dos sistemas: i) cinzento claro indica divergência na etiqueta, sendo a anotação do fenómeno coberta ou parcialmente coberta por anotações a outro nível, com poucos ou nenhuns custos de transposição; ii) cinzento escuro indica divergência no fenómeno a anotar, o que requer definição de novas opções de anotação, com custos de implementação).

	COPLE2	**POR Nível**
Fonologia e Ortografia	*StressMark*	Nasalidade
	Grapheme_Addition	Vogal Ditongo
	Grapheme_Deletion	
	Grapheme__Substitution	
	Grapheme_Transposition	
	Capitalization	
	WordBoundarySplit	
	WordBoundaryMerged	
	PunctConfused	
	PunctRedundant	
	PunctMissed	
Gramática	*UnnecessaryWord*	Artigo Preposição Clítico
	OmittedWord	
	WrongWord	Artigo Preposição Pronome Verbo Demonstrativo Advérbio
	WrongCategory	Nome-Adjetivo
	WrongStructure	Comparativa Relativa Interrogativa *wh*- Elipse
	Agreement_Gender	SN Sujeito-Verbo Sujeito-Predicativo do Sujeito Sujeito
	Agreement_Number	
	Agreement_Gender&Number	
	Agreement_Person	
		Concordância negativa
	WordOrder	
	Verb_Tense	Forma
	Verb_Mode	
	Verbe_Tense&Mode	
	Verb_FiniteNoFinite	
	Verb_Aspect	

	Verbal_Construction	*Voice*	
		Periphrasis	
		ComplexForm	
		Clitização	
			Argumento duplo
	PronounClitic	*Case*	
		Person	
		Position	
	NounNumber		
	SuffixInflection		Nome plural
	SuffixDerivation		
			Contração
Léxico	*LexicalChoice*		Formas de tratamento
	UnexistentWord		Expressão fixa Decalque

Tabela 5. Comparação dos sistemas de anotação COPLE2 e POR Nível

A análise da Tabela 5 permite verificar que, ao nível fonológico e ortográfico, a anotação do COPLE2 é mais exaustiva, cobrindo todos os casos analisados no POR Nível. Há alguma disparidade na marcação de erros no que respeita à nasalidade, explícita na anotação do POR Nível, mas esta é parcialmente coberta pela presença/ausência do diacrítico nos dados do COPLE2. No entanto, a transposição da anotação de erros respeitantes a vogal e ditongo, proposta no POR Nível, requer a análise dos casos em concreto, não sendo direta.

Ao nível da gramática/morfossintaxe, a identificação da categoria proposta no POR Nível para os casos de omissão de palavra, palavra desnecessária, palavra errada, bem como categoria errada é diretamente transponível para o sistema de anotação COPLE2, uma vez que a informação de categoria é coberta pela anotação de POS. Por outro lado, a categorização fina de estruturas erradas proposta pelo sistema POR Nível (comparativa, relativa, interrogativa *wh*-, elipse), bem como a anotação de argumento duplo e concordância negativa requerem a definição de novas opções de anotação, que poderão passar pela adição de novas subetiquetas ao *tagset* do COPLE2, pela anotação sintática dos dados, com risco de falha de desempenho do *parser* provocado pelos desvios das estruturas, ou pela análise sistemática de estruturas de superfície com base em pesquisas inteligentes, permitidas pelo tratamento e codificação do *corpus* e pelas ferramentas de pesquisa associadas ao COPLE2 (ver secção 3.2).

A transposição da anotação das contrações no sistema POR Nível indicada na tabela acima requer algum custo e diz respeito à anotação de falta de contração, como nos exemplos dados na Tabela 3 "*a comida* ***de aqui*** (ES) -> *daqui; gostas* ***de eles*** (ES) -> *deles; estou* ***em o*** *Caribe* (ES) -> *no*". No entanto, os casos de não contração de preposição com pronome/demonstrativo/artigo/... são facilmente recuperáveis nos dados do COPLE2, considerando que o COPLE2 não desfaz as contrações e anota a contração com PREP+DEM/PRS/DA/... (preposição + demonstrativo/pronome pessoal/artigo definido/...).

Ainda no que diz respeito ao nível da gramática, importa referir que a anotação de erros nas formas nominais (nome plural) e formas verbais propostas no POR Nível são diretamente transponíveis para o sistema de anotação do COPLE2 pela anotação de POS fina.

Finalmente, no nível lexical, o nível de especificação da análise proposta pelo POR Nível, nomeadamente no que respeita à identificação de decalques ou erros em expressões fixas, requer uma análise caso a caso com custos elevados e com riscos de baixo nível de concordância entre anotadores, pelo grau de interpretação que pode implicar, pelo que não é aconselhada. No entanto, no que respeita à anotação de erros em fórmulas de tratamento, esta poderá implicar menores custos de transposição na medida em que, por um lado, os erros podem estar parcialmente cobertos pela anotação de erros relativos aos pronomes pessoais do COPLE2 (*WrongWord*

+ *anotação de POS*) e, por outro, pelo facto de as fórmulas de tratamento poderem ser sistematicamente pesquisáveis (ex.: *Senhor/Doutor/Professor*).

3.2. Anotação de estruturas-alvo

A anotação das estruturas-alvo, ou bons usos, nos dados do POR Nível parece ser o fator mais divergente na comparação entre os sistemas. A anotação dos bons usos reflete a boa utilização de unidades/construções de diferentes categorias linguísticas e está associada a níveis de língua diferentes. No entanto, e apesar de não ser explicitamente marcada com etiquetas de anotação, a identificação de bons usos nos dados do COPLE2 é possível através do mapeamento das estruturas-alvo/nível consideradas no POR Nível para extração no COPLE2 e pela codificação explícita do nível de língua do falante nos dados do *corpus*.

As figuras abaixo exemplificam este processo de mapeamento para extração de estruturas-alvo no COPLE2, para o caso de uso de *ir* seguido de infinitivo, nível A2 (Figura 1) e para o caso de utilização de *ser* e *estar*, nível A1 (Figura 2).

Figura 1. Exemplo de pesquisa de utilização de "ir" seguido de infinitivo no COPLE2 (http://teitok.clul.ul.pt/learnercorpus/i)

Os diferentes níveis de tratamento e anotação do COPLE2 cobrem a informação e os requisitos de pesquisa necessários para verificar a presença das estruturas-alvo em causa. A cada linha de contexto apresentada está também associada informação acerca da língua materna do falante, outras línguas estrangeiras, nível de proficiência, género textual, tópico da produção e número de *tokens*.

Com algum investimento na interface, o mapeamento sistemático das estruturas-alvo poderia resultar numa nova opção de pesquisa que permitisse extrair de forma sistemática as estruturas de cada nível, e sem necessidade de construir manualmente a fórmula de pesquisa.

Por sua vez, a Figura 2 ilustra a possibilidade que o sistema oferece de extrair contextos de erros na utilização de *ser* e *estar*, por exemplo, pela conjugação dos campos anotados, nomeadamente a forma (*form = est.**), que nos permite listar os caracteres do radical de *estar* e o lema normalizado (*lemma = "ser"*).

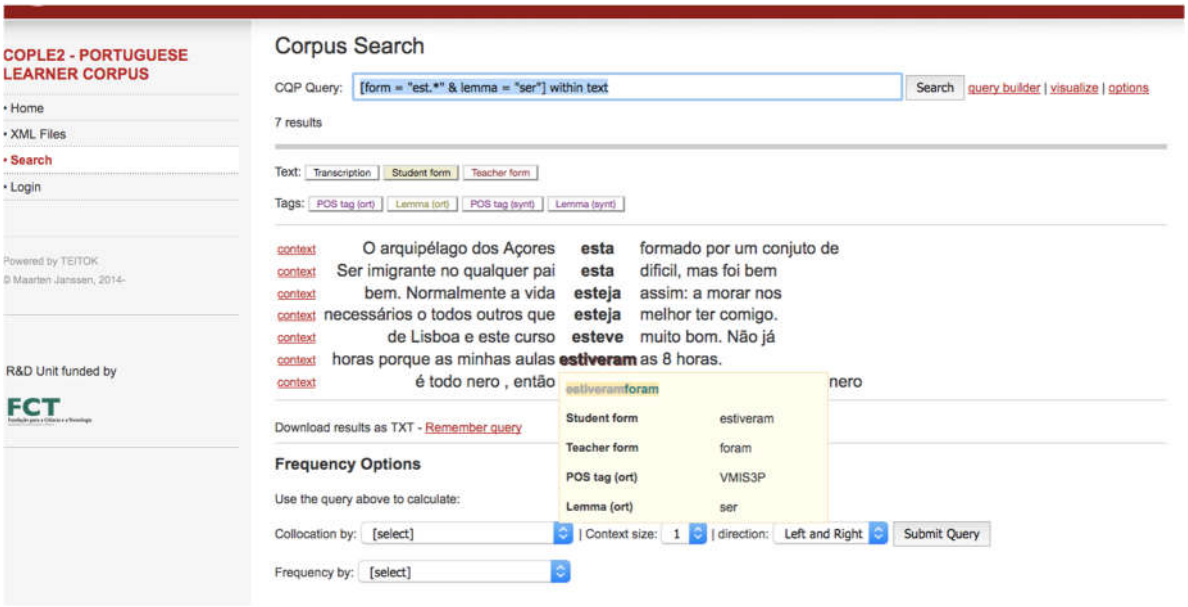

Figura 2. Exemplo de pesquisa de utilização de "estar" em vez de "ser" no COPLE2 (http://teitok.clul.ul.pt/learnercorpus/i)

4. Considerações finais

O trabalho aqui apresentado permite propor de modo sustentado: 1) a integração dos recursos analisados através da adoção do sistema de anotação definido para o COPLE2 aos dados do POR Nível, com perdas mínimas de especificação da anotação de erros ao nível fonológico/ortográfico e com recuo na especificação (problemática) de erros lexicais de alta subjetividade; 2) o enriquecimento do sistema de anotação do COPLE2 através da definição de novas subetiquetas para os casos de erro em comparativas, relativas, interrogativas *wh-*, elipse, argumento duplo e concordância negativa – propostos pelo sistema POR Nível – fazendo uso da codificação e sistema de pesquisa sofisticados do COPLE2; 3) o enriquecimento do COPLE2 pela identificação de estruturas-alvo nos dados, através do mapeamento das estruturas propostas no POR Nível em estruturas de superfície codificáveis em CQL e organizadas em *scripts* de pesquisa.

Deste modo, podemos concluir que a integração dos dados do COPLE2 e do POR Nível não só é desejável pelos benefícios que traz a cada um dos recursos, como é possível sem grandes custos associados. Em termos gerais, esta integração implica a inclusão dos *corpora* do POR Nível no COPLE2, respeitando embora os critérios de inclusão do COPLE2, e a operacionalização das novas funcionalidades de extração e marcação dos dados e das novas opções de anotação, considerando sempre a viabilidade da sua extensão à totalidade do *corpus*.

Retomando o objetivo inicial da análise dos recursos para o diagnóstico automático de proficiência em PLE/L2, e considerando que a automatização no diagnóstico de nível de língua depende fundamentalmente dos recursos disponíveis, verificamos que o PLE/L2 se encontra já dotado de recursos que sustentam a

automatização do seu diagnóstico e motivam o trabalho sobre a adaptação das métricas de análise (ex.: *Coh-Metrix*) ao português europeu. Os sistemas de anotação analisados viabilizam as abordagens e as técnicas em prática, seja através da anotação de erro seja pela identificação e mapeamento de estruturas-alvo associadas à complexidade e a níveis de proficiência. Nos recursos analisados, há algumas informações e análises em falta, nomeadamente informação para a LSA e informação de ordem lexical no que respeita a cadeias de hiperonímia ou marcação de polissemias. No entanto, essas análises recorrem tipicamente à utilização integrada de recursos externos ao *corpus*, como *wordnets* ou a *wikipedia*, sendo, por isso, um tópico de trabalho futuro em si mesmas.

Como nota final, importa salientar novamente o potencial de usabilidade destes recursos e chamar a atenção para o seu real valor, no que respeita aos vários níveis de tratamento e análise que refletem e que condicionam crucialmente o desempenho dos sistemas de tratamento automático da língua.

Referências:

Alexandre, N. & Gonçalves, A. (2015) Copular constructions in Portuguese as a second language (PL2) by Chinese learners: Do typological differences matter? In: *Workshop on Copulas across Languages*. June 18-19, University of Greenwich, London, England.

Alexandre, N. & Pinto, J. (2014) Aspects of relative clauses in Portuguese as a foreign language by Chinese learners. In: *20th Conference of the European Association for Chinese Studies*. July 22-26, Braga, Coimbra.

Amaral, L., Meurers, D. & Silva, G. (2006) Using Intelligent Computer-Assisted Language Learning (ICALL) Systems to Support Portuguese Instruction. In: *The 5th International Conference of the American Portuguese Studies Association (APSA)*. University of Minnesota. Minneapolis, Minnesota, October 5 - 7, 2006.

Antunes, S., & Mendes, A. (2015) Portuguese Multiword Expressions: data from a learner corpus. In: *LCR2015: 3rd Learner Corpus Research Conference*. September 11-13, Radboud University, Nijmegen, The Netherlands.

Ballier, N., Diaz-Negrillo, A. & P. Thompson (eds.) (2013) *Automatic treatment and analysis of learner corpus data*. Amsterdam & Philadelphia: John Benjamins, pp. 249–264.

Biber, D. (1993) Representativeness in Corpus Design. In: *Literary and Linguistic Computing*, vol.8 (4). Oxford University Press, pp. 243-257.

Cabrera, M. & Zubizarreta, M. L. (2005) Overgeneralization of Causatives and Transfer in L2 Spanish and L2 English. In: D. Eddington (ed.). *Selected Proceedings of the 6th Conference on the Acquisition of Spanish and Portuguese as First and Second Languages*. Somerville, MA: Cascadilla Proceedings Project, pp. 15-30.

Castelo, A., Santos, R. & Freitas, M. J. (2015) O uso de vogais ortográficas por aprendentes de Português como língua estrangeira: unidade na diversidade. In: *Língua Portuguesa: Unidade na diversidade – Cultura, Literatura, História, Linguística, Tradução e Ensino*. November 5-6, Lublin, Poland.

Chapelle, C. A., & Chung, Y.-R. (2010) The promise of NLP and speech processing technologies in language assessment. In: *Language Testing*, 27(3), pp. 301–15.

Curto, P., Mamede, N., Baptista, J. (2014) Automatic readability classifier for European Portuguese. In: *INFORUM 2014 – Simpósio de Inf*ormática, pp. 309–324.

Dagneaux, E., Denness, S., Granger, S., Meunier, F., Neff, J. & Thewissen, J. (Eds.) (2005) *Error Tagging Manual*. Version 1.2. Centre for English Corpus Linguistics, Université Catholique de Louvain.

del Río, I. (2019) Automatic proficiency classification in L2 Portuguese. In: *Procesamiento Del Lenguaje Natural*, 63.

del Río, I., & Mendes, A. (2018a) Error annotation in a Learner Corpus of Portuguese. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA).

del Río, I., & Mendes, A. (2018b) Error annotation in the COPLE2 corpus. In: *Revista da Associação Portuguesa de Linguística*, (4), pp. 225-239. https://doi.org/10.26334//2183-9077/rapln4ano2018a42.

DuBay, W. H. (2004) *The Principles of Readability*. Costa Mesa, California: Impact Information.

Flesch, R. (1948) A new readability yardstick. In: *Journal of Applied Psychology*, 32, pp. 221-233.

Graesser, A.C., McNamara, D.S., Louwerse, M.M. & Cai, Z. (2004) Coh-Metrix: Analysis of text on cohesion and language. In: *Behavior Research Methods, Instruments, & Computers* 36(2), pp. 193–202.

Gramacho, C., Madeira, A., Martins, C., Alexandre, N., Pinto, J . & Correia, S. (2019) Por Nível: Construção e validação de um teste de colocação para o Português Língua Estrangeira–resultados de um estudo-piloto. In: *Revista da Associação Portuguesa de Linguística*, (5), pp. 172-189. https://doi.org/10.26334/2183-9077/rapln5ano2019a13

Granger, S. (1996) From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In: K. Aijmer, B. Altenberg & M. Johansson (eds.) *Languages in Contrast. Text-based cross-linguistic studies*. Lund Studies in English 88. Lund: Lund University Press, pp. 37-51.

Granger, S., Gilquin, G. & F. Meunier (eds.) (2015) *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press.

Jarvis, S. (2000) Methodological rigor in the study of transfer: Identifying L1 influence in the interlanguage lexicon. In: *Language Learning* 50(2), pp. 245-309.

Kincaid, J., Fishburne, R., Rogers, R., & Chissom, B. (1975) *Derivation of new readability formulas for navy enlisted personnel*. Branch Report 8-75. Millington, TN: Chief of Naval Training.

Landauer, T., McNamara, D. S., Dennis, S., & Kintsch, W. (2007) (eds.). *Handbook of Latent Semantic Analysis*. Mahwah, NJ: Erlbaum.

McEnery, T. & A. Hardie (2012) *Corpus Linguistics: Method, theory and practice*. Cambridge University Press.

Mendes, A., Antunes, S. Janssen, M. & Gonçalves, A. (2016) The COPLE2 Corpus: A Learner Corpus for Portuguese. In: *Proceedings of the 10th Language Resources and Evaluation Conference – LREC'16*, 23-28 May 2016, Portoroz, Eslovénia, pp. 3207-3214.

Meurers, D. (ed.) (2009) Automatic Analysis of Learner Language. In: *CALICO Journal* 26(3). Equinox Publishing Ltd.

Nicholls, D. (2003) The Cambridge Learner Corpus – error coding and analysis for lexicography and ELT. In: D. Archer, P. Rayson, A. Wilson & T. McEnery (eds.). *Proceedings of the Corpus Linguistics 2003 Conference*. Lancaster University, pp. 572-581.

Paquot, M. (2013) Lexical bundles and L1 transfer effects. In: *Language Learning and technologt* 14(2), pp. 30-49.

POR Nível - Construção e validação de um teste de colocação em nível para o PLE, projeto de investigação do Centro de Linguística da Universidade Nova de Lisboa (CLUNL), em parceria com o Centro de Linguística da Universidade de Lisboa. http://fabricadesites.fcsh.unl.pt/por_nivel/

Ripley, M. (2009) *JISC case study: Automatic scoring of foreign language textual and spoken*. http://community.dur.ac.uk/smart.centre1/jiscdirectory/media/JISC%20Case%20Study%20-%20Languages%20-%20v2.0.pdf

Scarton, C.E., Aluísio, S.M. (2010) Análise da inteligibilidade de textos via ferramentas de Processamento de Língua Natural: adaptando as métricas do Coh-Metrix para o Português. In: *Linguamática* 2(1), pp. 45–61.

SyB – Complexity, projeto de investigação da Universidade de Tübingen (EKUT), http://sifnos.sfs.uni-tuebingen.de/SyB-0.1/.

Talhadas, R. (2016) Mapping Grammatical Structures onto Proficiency Levels. In: *Proceedings of 12th International Conference on Computational Processing of the Portuguese Language*, http://propor2016.di.fc.ul.pt/wp-content/uploads/2016/07/RuiTalhadasPROPORSRW2016.pdf.

Tono, Y. (2003) Learner corpora: Design, development and applications. In: D. Archer, P. Rayson, A. Wilson e T. McEnery (eds.), *Proceedings of the Corpus Linguistics 2003 Conference*. Lancaster University, pp. 800-809.

Vajjala, S., Meurers, D. (2012) On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition. In: *Proceedings of BEA*. Montréal, Canada: ACL, pp. 163–173.

Weiss, Z., Meurers, D. (2018) Modeling the readability of German targeting adults and children: An empirically broad analysis and its cross-corpus validation. In: *Proceedings of COLING'2018*, pp. 303-317.