CUTe: Corpus of Portuguese Undergraduates' Texts
Um recurso para a investigação em escrita académica em português

Adriana Cardoso[1, 2]
Catarina Magro[1, 2]
João Braz[2]
Teresa Nunes[2]
1. Centro de Linguística
2. Escola Superior de Educação de Lisboa
acardoso@clul.ul.pt; cmm@clul.ul.pt; 2011450@alunos.eselx.ipl.pt;
2011418@alunos.eselx.ipl.pt

Abstract:

The *Corpus of Portuguese Undergraduates' Texts* (CUTe) is an error-tagged learner corpus of Portuguese academic writing. This corpus gathers texts produced by undergraduate students and aims to contribute to the development of writing skills in mother tongue through a methodology traditionally used in studies of Second Language Acquisition and Foreign Language Teaching. This article intends to present the CUTe project, focusing on aspects related to corpus design, data collection methodology, error taxonomy and annotation, and corpus search tools.

Keywords: academic writing, learner corpus, error taxonomy, educational linguistics

Palavras-chave: escrita académica, corpus de aprendizagem, tipologia de erros, linguística educacional

1. Introdução

O ensino superior em Portugal sofreu mudanças significativas nas últimas décadas, que não se traduziram apenas no aumento das ofertas de ensino, mas também no alargamento das origens sociais, culturais e geográficas dos alunos (cf. Balsa *et al*., 2001). Esta mudança no perfil do aluno que acede ao ensino superior teve um grande impacto nas competências dos alunos, nomeadamente ao nível da escrita. No passado, a maioria dos docentes do ensino superior assumia que os seus alunos dispunham das habilidades necessárias para produzir os géneros textuais usados em contexto académico (e.g., artigo, dissertação, trabalho de investigação). Contudo, a realidade desmente continuamente este pressuposto. Como tem sido demonstrado por diversos autores, são muitas as dificuldades de escrita que os alunos enfrentam, mesmo à saída da licenciatura (cf. Carvalho, 1999; Cabral, 2003; Carvalho & Pimenta, 2005; Vasconcelos, Monteiro & Pinheiro, 2007; Rodrigues & Pereira, 2008; Rodrigues, 2010; Cardoso, Hortas *et al*., 2012; Cardoso, Magro *et al*., 2012; Cardoso & Magro, 2013).

***Textos Selecionados, XXIX Encontro Nacional da Associação Portuguesa de Linguística,* Porto, APL, 2014, pp. 169-184, ISBN 978-989-97440-3-5**

Ainda que estas dificuldades se manifestem nos diversos planos da construção de texto, investigação recente revela que é ao nível da microestrutura textual que os problemas são mais profundos, afetando a generalidade dos alunos e comprometendo a qualidade global das suas produções escritas (Cardoso, Hortas *et al*., 2012; Cardoso, Magro *et al*., 2012; Cardoso & Magro, 2013). Na verdade, embora falantes competentes, a maioria dos alunos do ensino superior em Portugal mostra não dominar o conhecimento gramatical e as convenções de representação gráfica da língua que garantem o cumprimento da norma – variedade valorizada no meio escolar (Costa, 2007).

Apesar de atualmente haver uma perceção generalizada deste problema, não há, até agora, investigação sistemática neste domínio: investigação que, com base em *corpora* de larga escala, proceda a uma análise qualitativa e quantitativa dos erros de microestrutura presentes nas produções escritas dos alunos do ensino superior.

É sabido que existe uma correlação entre desenvolvimento linguístico e desenvolvimento da escrita, como também se sabe que a aprendizagem formal de conteúdos gramaticais, sobretudo em idade adulta, contribui para o desenvolvimento de níveis de proficiência de escrita (Barbeiro, 1994, 1999, 2002; Rodrigues & Silvano, 2009; Costa, 2010). Assim, o sucesso das disciplinas de escrita académica que, desde há uns anos, integram os planos de estudo da grande maioria dos cursos de licenciatura em Portugal depende, em grande parte, da identificação das áreas críticas da expressão escrita dos alunos e de uma intervenção didática linguisticamente orientada.

A constituição de um *corpus* de escrita académica em português, à semelhança dos 'learner corpora' criados no âmbito de estudos em aquisição de L2 (ou em aquisição de língua estrangeira)[1], assume, neste contexto, especial relevância:

> By offering more accurate descriptions of learner language than have ever been available before, computer learner corpora will help researchers to get more of the facts right. (...) And in a more practical way, they will help to develop new pedagogical tools and classroom practices which target more accurately the needs of the learner (Granger, 1998: 17).

As vantagens desta metodologia, tradicionalmente utilizada na investigação sobre aquisição de L2 ou na didática de língua estrangeira, parecem ser extensíveis ao ensino da escrita académica em língua materna. Na verdade, quer a extensão e profundidade das dificuldades de escrita manifestadas pelos alunos portugueses, quer a necessidade de ensino formal das marcas linguísticas que caracterizam a escrita académica justificam uma estratégia em moldes semelhantes.

É neste sentido que agora se apresenta o CUTe – *Corpus of Portuguese Undergraduates' Texts*: o primeiro *corpus* de aprendizagem de escrita académica em português. O presente artigo, que pretende dar a conhecer este projeto, tem a seguinte estrutura: a secção 2 é destinada à apresentação do CUTe, ocupando-se de aspetos relativos aos objetivos específicos do projeto, à composição do *corpus*, à metodologia adotada na compilação e tratamento dos dados, à tipologia definida para anotação de erros e ao modo de disponibilização e pesquisa do *corpus*. A secção 3 apresenta uma síntese dos primeiros resultados de análise do *corpus* e a secção 4 encerra o artigo com considerações finais.

[1] A constituição e utilização de 'learner corpora' no âmbito dos estudos em aquisição de L2 ou do ensino de língua estrangeira é uma prática que surgiu na europa no final da década de 80, tendo atualmente uma forte tradição. Para um estado da arte da investigação neste domínio, vejam-se, entre outros, os seguintes trabalhos: Granger (1998, 2004); Granger, Hung & Petch-Tyson (2002); Nesselhauf (2004); Gilquin, Granger & Paquot (2007); Granger, Gilquin & Meunier (2013).

2. Apresentação do CUTe (*Corpus of Portuguese Undergraduates' Texts*)

2.1. Objetivos do projeto

O CUTe – *Corpus of Portuguese Undergraduates' Texts* é um projeto sem financiamento em desenvolvimento na Escola Superior de Educação de Lisboa (ESELx) desde julho de 2012. Este projeto visa contribuir para o desenvolvimento de competências de escrita académica através da criação e disponibilização de um recurso que satisfaça os requisitos empíricos da investigação neste domínio.
Concretamente, o projeto tem como objetivo constituir e disponibilizar um *corpus* eletrónico de escrita académica em português com anotação de erros que envolvem os níveis micro e macroestruturais. A constituição de um *corpus* de aprendizagem com estas características permite avaliar qualitativa e quantitativamente os problemas presentes nas produções escritas dos alunos do ensino superior, fornecendo as bases para uma intervenção didática mais direcionada e linguisticamente orientada.
Através deste projeto, pretende-se ainda promover a investigação em linguística educacional centrada na escrita académica, uma área de estudos sem tradição em Portugal.

2.2. Composição do corpus

O CUTe é constituído por textos produzidos em contexto académico por alunos de licenciatura da ESELx, estando planeada, para uma segunda fase de desenvolvimento do projeto, a integração no corpus de textos de alunos de outras instituições de ensino superior.
Atualmente, os textos incluídos no corpus são representativos de dois géneros académicos – artigo de divulgação e artigo de opinião – e são produzidos em dois momentos distintos do percurso académico dos alunos da ESELx: (i) a unidade curricular de *Técnicas de Expressão Oral e Escrita* (UC do primeiro ano dos cursos de licenciatura) e (ii) a *Prova de Língua Portuguesa de Acesso aos Mestrados Profissionalizantes* (prova realizada no final da licenciatura para acesso aos mestrados que habilitam para a docência).[2]
Em ambos os contextos acima referidos, os textos são elaborados no âmbito de provas de avaliação, o que garante a uniformidade das condições de produção: a produção textual é vigiada e decorre num período temporal controlado; os textos são manuscritos, redigidos sem recurso a instrumentos de normalização linguística, tendo por base uma instrução de escrita que determina o seu género, tema e extensão.
O desenho do corpus respeita princípios de equilíbrio, assegurando-se a proporcionalidade entre número de textos por género, número de textos por estudante e número médio de palavras por texto. O Quadro 1, abaixo, apresenta a composição prevista para a primeira fase de desenvolvimento do projeto.

[2] A inclusão de textos no CUTe está dependente da autorização prévia dos seus autores. Os estudantes concedem esta autorização formalmente, assinando uma declaração e preenchendo um formulário em que fornecem informação relativa a idade, sexo e percurso académico. Estes dados são incluídos na base de dados que suporta o corpus e associados a um código de identificação do estudante. Os textos são anonimizados.

nº de estudantes	120
nº de textos por estudante	4 • 2 artigos de divulgação (1 produzido em TEOE, outro produzido na PLP) • 2 artigos de opinião (1 produzido em TEOE, outro produzido na PLP)
nº total de textos	480
nº médio de palavras por texto	350
nº total de palavras	168 000

Quadro 1: Composição do CUTe

Presentemente, o corpus conta com uma extensão de 40 000 palavras, das quais estão anotadas 17 500 (correspondentes a 50 textos).

2.3. Compilação e tratamento de dados

O CUTe é armazenado numa base de dados construída com o Programa FileMaker Pro Advanced. Este sistema de gestão de bases de dados é o *software* utilizado quer para a transcrição dos textos, quer para a anotação dos erros.

A transcrição dos manuscritos segue uma edição semiconservadora:

- respeita a marcação de parágrafo dos manuscritos originais e inclui indicação de quebra de linha e número de linha;
- respeita a ortografia e pontuação dos manuscritos originais; [3]
- não inclui indicações relativas a acidentes de escrita (anulações, entrelinhados, etc.).

Para anotação de erros na base de dados, os textos são divididos nos períodos que os constituem. Cada período tem um número de entradas na base de dados correspondente ao número de erros que contém. O processo de anotação envolve: (i) a codificação manual do erro de acordo com a tipologia definida (ver secção 2.4.), (ii) o destaque gráfico da sequência em que o erro ocorre e (iii) a indicação do(s) número(s) de linha(s) respetivo(s).

2.4. Tipologia de erros

Na literatura têm sido propostas diversas tipologias de erros, que têm subjacentes diferentes pressupostos teóricos, objetivos e destinatários. Para o português europeu, encontram-se disponíveis tipologias que se centram em erros produzidos por alunos nos primeiros anos de escolaridade (cf., i.a., Cardoso, Costa & Pereira, 2002; Baptista, Viana & Barbeiro, 2011), não existindo porém uma uniformização nas classificações propostas. A nível internacional, também se verifica uma ausência de estandardização neste domínio. Mesmo considerando a anotação de erros desenvolvida no âmbito dos 'learner corpora', as propostas apresentadas pelos diferentes grupos de trabalho exploram diferentes modelos de anotação e diferentes tipologias (cf. Díaz-Negrillo & Fernández-Domínguez, 2006).

[3] Sobre a dificuldade da manutenção dos erros originais na transcrição dos textos de corpora de aprendizagem, veja-se Granger (1998).

No âmbito do presente projeto, não foi adotada *a priori* nenhuma das tipologias disponíveis na literatura, tendo sido construída uma tipologia para a análise dos erros especificamente encontrados no CUTe [4]. Trata-se, portanto, de uma tipologia em construção, que, até ao momento, dá conta apenas dos erros que ocorrem nos 50 textos anotados (com ca. de 17 500 palavras), mas que no futuro, à medida que novos textos forem anotados, poderá vir a ser alargada de forma a permitir a classificação de outros tipos de erro.

A tipologia proposta (cf. Quadro 2) tem por base a análise linguística dos textos, considerando apenas os erros que envolvem os níveis micro e macroestruturais. Encontra-se organizada em dois níveis diferentes de descrição/anotação: nível de análise e categoria.

Nível de análise	**Categoria**
Ortografia	Acentuação
	Maiúsculas/minúsculas
	Translineação
	Hifenização
	Processamento e representação gráfica de fonemas
	Acordo ortográfico
	Outros
Pontuação	Uso da vírgula
	Uso de outros sinais de pontuação
Morfologia	Flexão
	Derivação
Sintaxe	Ordem de palavras
	Topicalização
	Coordenação
Morfossintaxe	Concordância
	Segmentação de palavras
Sintaxe/semântica	Seleção categorial
	Seleção semântica
	Seleção de modo
	Outras incompatibilidades categoriais
	Outras incompatibilidades semânticas
Semântica	Referência nominal
	Referência temporal-aspetual
	Conetores
Gralhas	

Quadro 2: Tipologia construída para anotação do CUTe

[4] Esta opção prende-se com diferentes fatores: (i) as tipologias disponíveis para o português são orientadas para a categorização de erros que ocorrem tipicamente em textos produzidos por alunos a frequentar os primeiros anos de escolaridade (cf. Cardoso, Costa & Pereira, 2002; Baptista, Viana & Barbeiro, 2011); (ii) o grau de granularidade das tipologias construídas no âmbito de 'learner corpora' não permite anotar o corpus com o nível detalhe adequado aos objetivos do projeto.

Como se pode constatar pela análise do Quadro 2, a tipologia é constituída por sete grandes níveis de análise, que contemplam quer erros relativos às convenções de representação gráfica (*ortografia*, *pontuação*), quer erros que envolvem diferentes níveis de análise linguística (*morfologia*, *sintaxe*, *semântica*), organizados, por vezes, em categorias compósitas (*morfossintaxe*, *sintaxe/semântica*). Por fim, foi ainda considerado um nível suplementar para a anotação de *gralhas*.
Os erros codificados ao nível da *ortografia* são classificados de acordo com as seguintes categorias:

- *acentuação* (e.g. *obrigatóriamente*, CUTe F14 L14);
- *maiúsculas/minúsculas* (e.g., *diretor de Turma*, CUTe F10 L13);
- *translineação* (e.g., *a-/credito*, CUTe F09 L31-32);
- *hifenização* (e.g., *mais valia*, CUTe F46 L52);
- *processamento e representação gráfica de fonemas*[5] (e.g., *procorar*, CUTe F12 L33; *treceiro*, CUTe F46 L37);
- *acordo ortográfico*[6] (e.g., *lectivo*, CUTe F46 L03);
- *outros*[7] (e.g., *bulling*, CUTe F40 L33).

O nível de análise *pontuação* contempla duas categorias: *uso da vírgula* e *uso de outros sinais de pontuação*. A categoria *uso da vírgula* inclui casos de omissão de vírgula, inserção de vírgula e uso de vírgula em vez de outros sinais de pontuação (cf., respetivamente, exemplos (1)-(3) abaixo). A categoria *uso de outros sinais de pontuação* contempla casos que envolvem: (i) falta de parêntesis; (ii) uso indevido de dois pontos; (iii) uso indevido de ponto e vírgula; (iv) uso indevido de ponto final (cf. exemplos (4) e (5) abaixo, que envolvem uso de ponto final em vez de vírgula).

(1) Tendo este aspeto em considera-/ção Ø sou a favor destas medidas (CUTe F22 L31)
(2) Esta nova lei, não reforça o papel dos professores (CUTe F04 L24)
(3) Teremos de ter em conta, no entanto, o que muitas vezes está por traz das diferentes situa-/ções, pais que querem e não conseguem controlar como deseja-/vam os seus filhos, filhos que não possuem quem os oriente, famílias destruturadas com graves problemas económicos (CUTe F36 L18)

(4) As actividades de recuperação são de total autonomia do professor titular e da escola. O que exije ainda mais trabalho dos professores. (CUTe F15 L44)
(5) É, neste sentido, que a escola tem o dever e a obrigação de fomentar e desenvolver a Ética Escolar. Devendo fomentar diversos valores como a solidariedade, a entreajuda, o respeito, a justiça e a igualdade em todos os seus alunos. (CUTe F28 L43)

O nível de análise *morfologia* contempla erros que resultam quer de problemas na representação gráfica de afixos, quer do uso de afixos na criação de novas palavras ou

[5] Na categoria *processamento e representação gráfica de fonemas*, são considerados casos que envolvem quer erros que resultam da inexistência de relação biunívoca entre som e grafia (e.g., *procorar*, CUTe F12 L33) (cf. Cardoso, Costa & Pereira, 2002) , quer erros que resultam de problemas de processamento dos sons (segmentação, identificação e ordenação) (e.g. *treceiro*, CUTe F46 L37) (cf. Baptista, Viana & Barbeiro, 2011).
[6] Nos textos anotados até ao momento, os alunos podiam optar pela grafia prevista no Acordo Ortográfico de 1945 ou no de 1990, devendo indicar explicitamente a sua opção. Assim, na categoria *acordo ortográfico* foram considerados apenas os casos de erros não consistentes com a opção escolhida.
[7] Neste momento, a categoria *outros* inclui apenas erros de representação gráfica de palavras estrangeiras.

em formas flexionadas de uma palavra. Neste nível, estão previstas duas categoriais principais: *flexão* (cf. exemplo (6) e *derivação* (cf. exemplo (7)).

(6) creio que só assim se obteram resultados. (CUTe F03 L11)
(7) uma agravação do seu desempenho escolar. (CUTe F11 L30)

O nível de análise *sintaxe* contempla as categorias: *ordem de palavras*, *topicalização* e *coordenação*. A categoria *ordem de palavras* é usada, entre outros, para classificar erros relativos à posição dos pronomes clíticos (cf. (8)), à extraposição ilegítima de orações (cf. (9)) e à posição/escopo de partículas focalizadoras (cf. (10)). A categoria *topicalização* é usada para codificar estruturas que, embora sejam produzidas no registo oral, não são aceitáveis na escrita académica. Tal é o caso das estruturas que envolvem a construção de tópico pendente (cf. (11)) e de deslocação à esquerda de tópico pendente (cf. (12)) (cf. Duarte, 2003). Por fim, a categoria *coordenação* contempla os casos de coordenação assimétrica, em que os termos coordenados não têm a mesma natureza categorial, como em (13)[8].

(8) Penso isso uma vez que tem-se assistido nos últimos anos a um aumento do incumprimento de regras (CUTe F25 L21)
(9) O objectivo principal do novo estatuto do aluno é combater a indisciplina, em que a assiduidade e a pontualidade devem ser respeitadas pelos alunos e respectivas famílias. (CUTe F15 L19)
(10) A edu-/cação não deveria ser apenas deixada para as escolas e os pais têm papel fundamental nesse aspeto, contudo infelizmente existem muitos pais que não procuram ter esse papel. (CUTe F22 L09)

(11) O aluno se não faz os TPC sistematicamente, é importante chamar a atenção do encarregado de educação, explicando também a importância de realização dos deveres em casa (CUTe F10 L19)
(12) Por vezes a violência que os alunos trazem para a escola por norma, eles observam e convivem com isso no seu dia-a-dia, (CUTe F10 L37)

(13) terão como intuito melhorar o rigor e a qualidade do ensino, através da responsabilidade atribuída aos alunos e seus encarregados, assim como, a valorização do papel dos profissionais da educação. (CUTe F04 L44-46)

No nível de análise *morfossintaxe*, são consideradas as categorias *concordância* e *segmentação de palavras*. São anotados como *concordância* os erros que envolvem falta de concordância: entre sujeito e verbo (cf. (14)); entre sujeito (ou objeto direto) e predicativo (cf. (15)); no interior do sintagma nominal (cf. (16)). Na categoria *segmentação de palavras* encontram-se anotados, até ao momento, apenas erros que envolvem hipossegmentação (i.e., união indevida de palavras), como em (17) e (18).

(14) É importante que seja aplicada sansões/coimas às famílias (CUTe F06 L04)
(15) As famílias que negligenciam a educação dos seus filhos, privando-os de frequência diária escolar devem ser responsabilizados. (CUTe F32 L17)
(16) A meu ver acho que é uma lei com muito exageros (CUTe F41 L07)

[8] Nos casos em que os membros coordenados são selecionados por predicadores, as agramaticalidades que envolvem coordenação assimétrica são codificadas no nível *sintaxe/semântica*, em *seleção categorial*.

(17) Esta é uma das razões porque penso ser importante a comunicação (CUTe F22 L22)
(18) Em segundo lugar, considero bastante interessante o facto da responsabilidade não apenas dos alunos (CUTe F31 L19)

No nível *sintaxe/semântica*, são consideradas as categorias *seleção categorial*, *seleção semântica, seleção de modo, outras incompatibilidades categoriais* e *outras incompatibilidades semânticas/lexicais*. A categoria *seleção categorial* é usada para codificar os erros que decorrem da violação das restrições impostas por um predicador à natureza categorial dos seus argumentos (cf. exemplos (19) a (21), que envolvem, respetivamente, um predicador nominal, verbal e adjetival). A categoria *seleção semântica* contempla erros que envolvem alteração do número de argumentos selecionados pelo predicador (cf. (22)), bem como casos de violação das restrições de seleção semântica, em que não são respeitados os traços semânticos que os predicadores impõem aos seus argumentos (cf. (23)). A categoria *seleção de modo* é usada para codificar os casos em que o modo selecionado não é o que ocorre, como é o caso do uso do indicativo pelo conjuntivo ou vice-versa (cf. (24)-(25))[9]. Por fim, as categorias *outras incompatibilidades categoriais* (cf. (26)) e *outras incompatibilidades semânticas/lexicais* (cf. (27)) são utilizadas na codificação de erros que envolvem incompatibilidades entre constituintes que não estabelecem entre si uma relação de predicador-argumento.

(19) Em suma, sou da **opinião** Ø que, este novo estatuto se focaliza em algumas das maiores fragilidades do ensino, em Portugal. (CUTe F04 L39)
(20) Em suma, a escola não deve **consti-/tuir** num lugar de autoridade e disciplina (CUTe F12 L36-37)
(21) sendo, contudo, necessário que as famílias estejam **dispostas** para tal. (CUTe F35 L38)

(22) e as "más companhias" anteriormente referidas chegam de ambientes familiares pouco sau-/dáveis e **débeis** em educação. (CUTe F11 L40)
(23) Se as famílias não estão preparadas para educar as suas crianças, vamos ajudá-las a **atingir** as ferramentas que necessitam para o fazerem, ao invés de as castigarmos e de as submeter à humilhação social. (CUTe F08 L39)

(24) Em conclusão, penso que o novo "Estatuto do Aluno e Ética Escolar" deveria ser reavaliado e modificado para que existem alterações nos problemas de assiduidade e disciplina dos alunos. (CUTe F13 L33)
(25) Em primeiro lugar, julgo que seja extremamente essencial o reforço dos docentes e de todos os agentes educativos (CUTe F31 L10)

(26) para que haja, um maior sucesso e empenho escolar e um melhor **bem-/-estar** nas escolas por parte dos alunos e professores (CUTe F37 L04-05)
(27) Não posso concordar com **este mecanismo**, onde só são dados apoios de educação parental aos pais (CUTe F08 L22)

O nível *semântica* centra-se em problemas que afetam a coesão e a coerência textuais, incluindo as categorias *referência nominal, referência temporal-aspetual* e *conetores*. A categoria *referência nominal* contempla problemas de dependência referencial – que

[9] Nesta categoria são também codificados erros que envolvem o uso de formas finitas por formas não finitas e vice-versa.

envolvem a construção de referência anafórica (cf. (28)), referência dêitica (cf. (29)) e referência anafórica/dêitica (cf. (30)) –, bem como casos de determinação nominal (cf. (31)). A subcategoria *conetores* é usada para codificar erros que envolvem o uso inadequado/ausência de conetores, como ilustrado em (32).

(28) Por tudo isto, na minha opinião, as novas medidas não solucionam os problemas em questão. Os deveres não são todos iguais, as suas necessidades não são iguais, as soluções não podem ser as mesmas para todos. (CUTe F35 L49-52)
(29) Para além do já referido, a situação agrava-se quando o aluno é maior, porque incorre na retenção do presente ano lectivo e ainda se vê impedido de voltar à escola nos dois anos seguintes. (CUTe F34 L41-42)
(30) Há casos de pais que se preocupam e acompanham o percurso escolar dos seus filhos, contudo, esses continuam a cometer atos de indisciplina e de mau comportamento. (CUTe F07 L18-19)
(31) Nesse sentido, acredito que uma diminuição do horário laboral de um dos pais nos primei-/ros anos da infância será hipótese (CUTe F50 L25)
(32) Com o passar do tempo, os alunos (alguns) tornaram-se indisciplinados, não respeitando as regras que lhes são impostas. Para tal e na minha opinião é importante criar medidas para travar a indisciplina. (CUTe F46 L24)

2.5. Disponibilização e pesquisa do corpus

O CUTe está disponível através da página *web* do projeto: http://www.cute.org.pt. A partir de um *interface* de pesquisa *online*, construído com o programa FileMaker Pro Advanced, é possível descarregar e pesquisar o *corpus*.

A opção "pesquisa por erro" permite efetuar a pesquisa de erros de qualquer nível de análise ou categoria previstos na tipologia (cf. Figura 1). Nesta modalidade de pesquisa, a busca pode incidir sobre todo o corpus ou apenas sobre um subcorpus definido.

Para além da listagem das ocorrências de erro em contexto, o *output* da pesquisa inclui informação sobre o número total de ocorrências e o número total de textos em que o erro ocorre (cf. Figura 2).

Em alternativa, a opção "pesquisa por texto" opera sobre a versão não anotada do corpus. Nesta modalidade, o âmbito de pesquisa pode ser restringido em função das variáveis *texto*, *aluno*, *género textual*, *contexto de produção* e *data de produção* (cf. Figura 3). O *output* da "pesquisa por texto" corresponde à lista de textos, completos e não anotados, que satisfazem os requisitos da pesquisa realizada (cf. Figura 4).

Os dois modos de busca, que podem aliás funcionar em combinação, estão pensados para responder a diferentes objetivos de utilização do *corpus*, sendo a "pesquisa por erro" útil para os utilizadores que pretendam analisar transversalmente uma determinada categoria de erro e a "pesquisa por texto" útil para quem se interesse, por exemplo, pelo estudo do conjunto de textos produzidos por um mesmo aluno.

Os resultados de pesquisa do CUTe podem ser consultados *online* ou descarregados em formato txt, estando preparados para serem submetidos a outras ferramentas de tratamento de texto, como programas de concordância, etiquetadores ou anotadores sintáticos.

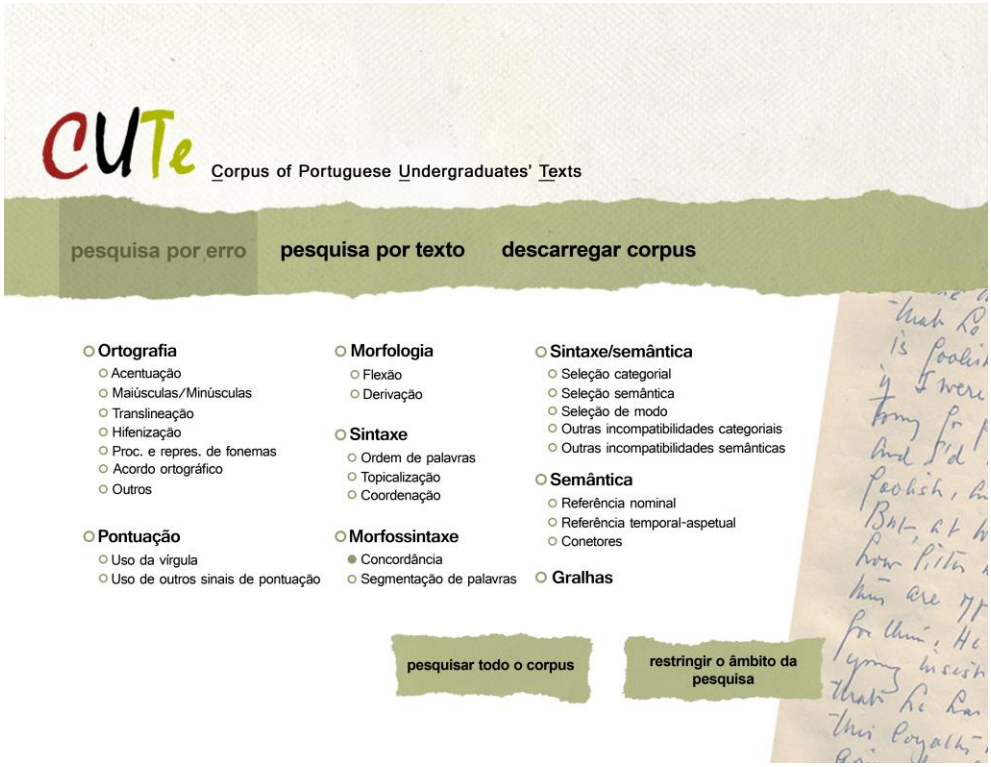

Figura 1: Interface de pesquisa por categoria de erro

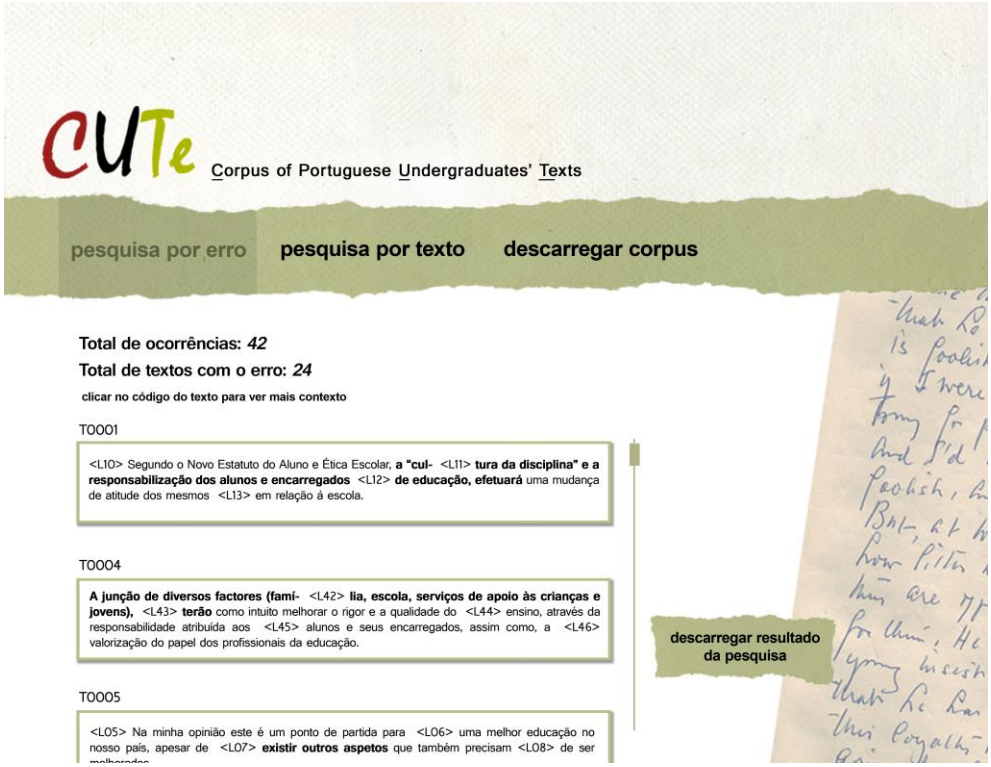

Figura 2: Visualização dos resultados da pesquisa por categoria de erro

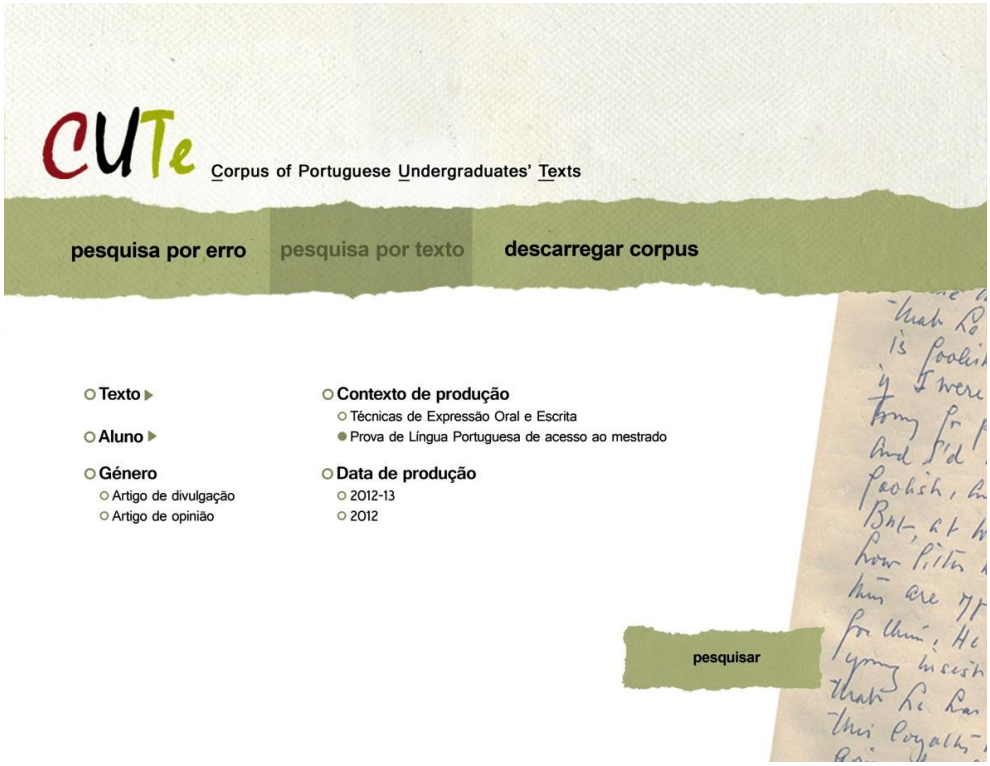

Figura 3: Interface de pesquisa por texto

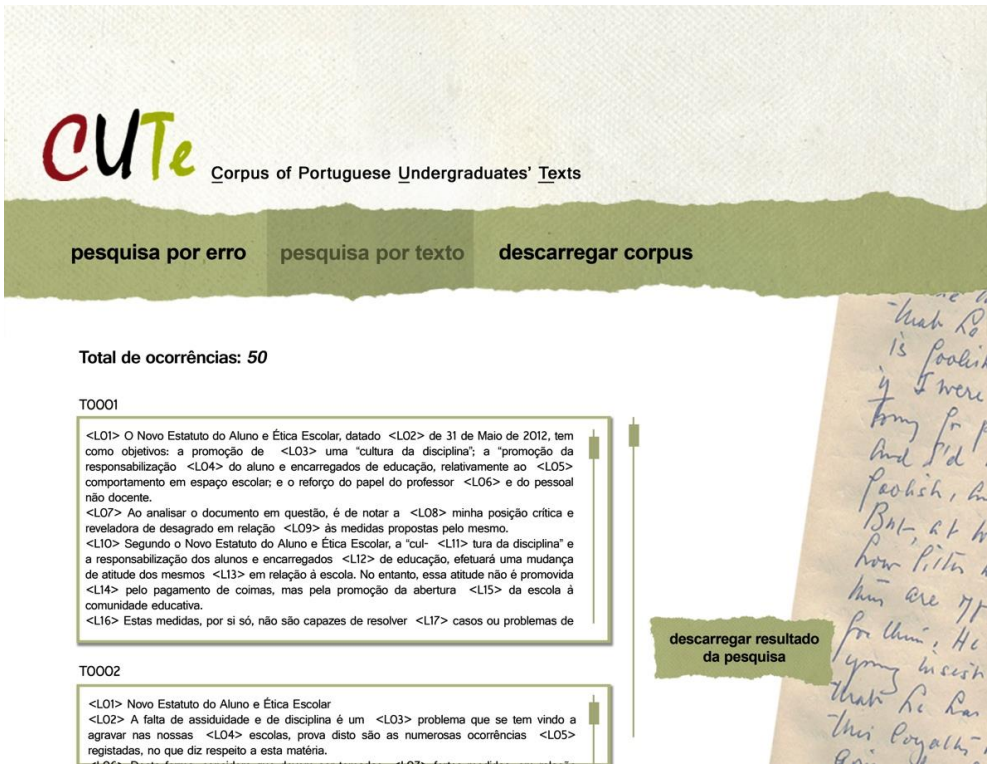

Figura 4: Visualização dos resultados da pesquisa por texto

3. Primeiros resultados

Nesta secção, apresentamos os dados quantitativos que resultam da análise do conjunto de cinquenta textos do CUTe atualmente anotados.
Como se pode observar no Quadro 3, no subcorpus considerado, registam-se 1069 ocorrências de erro, o que corresponde a uma média de 21,4 erros por texto. O número de erros por texto oscila entre 7 e 49.

Número total de erros	**Média de erros por texto**	**Texto com menor número de erros**	**Texto com maior número de erros**
1069	21,4	7	49

Quadro 3: Ocorrência de erros no CUTe – dados globais

A Figura 5 apresenta a distribuição dos textos em função do número de erros que atestam. Como se pode verificar, a grande maioria dos textos apresenta entre 11 a 30 erros, o que revela que o corpus em análise constitui uma amostra com uma distribuição normal.

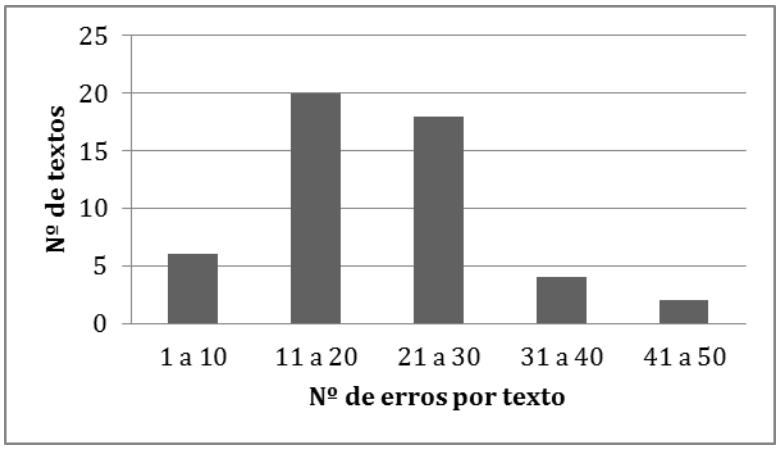

Figura 5: Ocorrência de erros no CUTe – distribuição dos texto por intervalo de erro

Em Portugal, não existe nenhum estudo quantitativo que incida sobre problemas de escrita no ensino superior, pelo que não é possível comparar os resultados do CUTe com outros dados nacionais. A título indicativo, apresentamos no Quadro 4 uma comparação entre a taxa de erro do CUTe e a obtida num estudo levado a cabo por Lunsford & Lunsford (2008) sobre a escrita de estudantes universitários americanos. Tomando como referência a média de erro por cem palavras, observa-se que existe um marcado contraste entre os valores apurados nos dois países.

Média de erro por 100 palavras **CUTe**	**Média de erro por 100 palavras** **Lunsford & Lunsford (2008)**
6,20	2,45

Quadro 4: Comparação da taxa de erro entre CUTe e Lunsford & Lunsford (2008)

Considerando em detalhe os diferentes níveis de análise previstos na tipologia (cf. Quadro 5), verifica-se que a *pontuação* e a *sintaxe/semântica* são os níveis com maior

incidência de erros, havendo uma diferença expressiva entre estes dois níveis e os seguintes.

Nível de análise	**Número de erros**	**% do total de erros**
pontuação	379	35,5
sintaxe/semântica	318	29,7
ortografia	151	14,1
semântica	108	10,1
morfossintaxe	51	4,8
sintaxe	38	3,6
morfologia	21	2,0
gralhas	3	0,3

Quadro 5: Ocorrência de erros no CUTe – por nível de análise

Com base numa análise mais fina, por categoria de erro, obtém-se o *top ten* apresentado no Quadro 5.

Categoria de erro	**Número de erros**	**% do total de erros**
Uso da vírgula	357	33,5
Seleção categorial	119	11,2
Seleção semântica	117	11,0
Referência nominal	83	7,8
Outras incompatibilidades semânticas	61	5,7
Maiúsculas/minúsculas	50	4,7
Acentuação	46	4,3
Concordância	42	3,9
Processamento e representação gráfica de fonemas	30	2,8
Coordenação	23	2,2

Quadro 5: Ocorrência de erros no CUTe – por categoria de erro (dez categorias com maior número de ocorrências)

A categoria que ocupa o primeiro lugar do *ranking*, com grande destaque, é a de uso da vírgula, seguindo-se a *seleção categorial* e *seleção semântica*. Estas três categorias em conjunto correspondem a mais de 50% dos erros atestados no corpus.

O uso da vírgula, tradicionalmente interpretado como um problema de domínio de convenções gráficas da representação escrita, pode, na verdade, resultar de problemas ao nível da consciência sintática. Com efeito, as regras de uso de vírgula (como *os constituintes apositivos são delimitados por vírgula* ou *as orações adverbiais antepostas são seguidas de vírgula*) requerem uma capacidade de reflexão sobre a estrutura da frase que muitos alunos revelam não ter.
Por outro lado, os erros de *seleção categorial* e de *seleção semântica* parecem resultar, pelo menos em parte, de uma tentativa de uso de um registo de língua formal/académico, que os alunos não dominam. Tal facto manifesta-se, entre outros aspetos, no uso de léxico culto, cujas propriedades de seleção os alunos desconhecem.

4. Considerações finais

Este artigo apresenta uma caracterização de um projeto seminal na área da escrita académica em português. A partir do trabalho já iniciado, estão lançadas as bases para constituir um *corpus* de aprendizagem de larga escala, que reúna uma amostragem nacional das produções escritas dos alunos do ensino superior português. Esta será uma ferramenta fundamental para o desenvolvimento de investigação na área da escrita académica e para uma intervenção didática empiricamente sustentada e linguisticamente motivada.

Referências

Balsa, Casimiro, Simões, J., Nunes, P., Carmo, R. & Campos, R. (2001). *Perfil dos Estudantes do Ensino Superior. Desigualdades e Diferenciação*. Lisboa: Edições Colibri/CEOS.

Barbeiro, Luís Filipe (1994). *Consciência Metalinguística e Expressão Escrita*. Dissertação de Doutoramento, Universidade do Minho.

Barbeiro, Luís Filipe (1999). *Os Alunos e a Expressão Escrita – Consciência Metalinguística e a Expressão Escrita*. Lisboa: Fundação Calouste Gulbenkian.

Barbeiro, Luís Filipe (2002). O processo de escrita e a relação com a linguagem. In Cristina Mello *et al*. (orgs.) *II Jornadas Científico-Pedagógicas de Português.* Coimbra: Almedina, pp. 101-115.

Baptista, Adriana, Fernanda Leopoldina Viana & Luís Filipe Barbeiro (2011). *O ensino da escrita: dimensões gráfica e ortográfica*. Lisboa: Ministério de Educação.

Cabral, Ana Paula (2003). *Leitura, Compreeensão e Escrita no Ensino Superior e Sucesso Académico*. Dissertação de Doutoramento, Universidade de Aveiro.

Cardoso, Adriana, Maria João Hortas, Encarnação Silva & Tiago Tempera (2012). Competências em Língua Portuguesa à saída da licenciatura: o caso da licenciatura em Educação Básica da ESELx. In *Atas do V Encontro do CIED - Escola e Comunidade*. Lisboa: CIED, pp. 447-460.

Cardoso, Adriana, Manuel Luís Costa & Susana Pereira (2002). Para uma tipologia de erros. *Ler Educação*, 2 (2), pp. 5-25.

Cardoso, Adriana, Catarina Magro, João Braz & Teresa Nunes (2012). Errare humanum est. Comunicação apresentada no congresso *Former à l'écrit universitaire : un terrain pour la linguistique.* Paris, Université Paris Ouest Nanterre La Défense e Université de Chicago. Novembro 2012.

Cardoso, Adriana & Catarina Magro (2013). Problemas de coesão referencial na escrita académica em português. Comunicação apresentada na *3ª Conferência Internacional em Gramática e Texto - GRATO 2013*. Lisboa, Faculdade de Ciências Sociais e Humanas, Universidade Nova de Lisboa. Dezembro 2013.

Carvalho, José António Brandão (1999). *O Ensino da Escrita - Da Teoria às Práticas Pedagógicas*. Braga: IEP, Universidade do Minho.

Carvalho, José António Brandão & Jorge Rocha Pimenta (2005). Escrever para aprender: escrever para exprimir o aprendido. In Bento Silva e Leandro Almeida (orgs.) *Actas do Congresso Galaico-Português de Psicopedagogia (8).* Braga: Universidade do Minho, Instituto de Educação e Psicologia, pp. 1877-1885.

Costa, João (2007). Conhecimento gramatical à saída do Ensino Secundário: estado actual e consequências na relação com leitura, escrita e oralidade. In Carlos Reis (org.) *Actas - Conferência Internacional sobre o Ensino do Português*. Lisboa: Ministério de Educação, pp. 149-165.

Costa, Ana (2010). *Estruturas Contrastivas: Desenvolvimento do Conhecimento Explícito e da Competência de Escrita*. Dissertação de Doutoramento, Universidade de Lisboa.

Díaz-Negrillo, Ana & Jesús Fernández-Domínguez (2006). Error tagging systems for learner corpora. *Revista Española de Linguistica Aplicada* 19, pp. 83-102.

Duarte, Inês (2003). Estrutura da frase simples e tipos de frase. In Maria Helena Mira Mateus et al. (orgs.) *Gramática da Língua Portuguesa*. Lisboa: Editorial Caminho, pp. 433-506.

Gilquin, Gaëtanelle, Sylviane Granger & Magali Paquot (2007). Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes* 6 (4), pp. 319‑335.

Granger, Sylviane (1998). The computer learner corpus: a versatile new source of data for SLA research. In Sylviane Granger (ed.) *Learner English on Computer*. Londres/Nova York: Addison Wesley Longman, pp. 3-18.

Granger, Sylviane (2004). Computer learner corpus research: current status and future prospects. In Ulla Connor e Thomas A. Upton (eds.) *Language and Computers, Applied Corpus Linguistics. A Multidimensional Perspective*. Amesterdão: Rodopi, pp. 123-145.

Granger, Sylviane, Joseph Hung & Stephanie Petch-Tyson (eds) (2002). *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching.* Amesterdão/Filadélfia: John Benjamins.

Granger, Sylviane, Gaëtanelle Gilquin & Fanny Meunier (eds.) (2013). *Twenty Years of Learner Corpus Research. Looking Back, Moving Ahead*. Lovaina: Presses universitaires de Louvain.

Lunsford, Andrea & Karen Lunsford (2008). Mistakes Are a Fact of Life: A National Comparative Study. *College Composition and Communication* 59 (4), pp. 781-806.

Nesselhauf, Nadja (2004). Learner corpora and their potential for language teaching. In John Sinclair (org.) *How to Use Corpora in Language Teaching*. Amesterdão: John Benjamins, pp. 125-152.

Rodrigues, Leila Calil Saade, & Luísa Alves Pereira (2008). Dificuldades de Síntese da Informação Escrita: a pertinência de uma didáctica do escrito no Ensino Superior. *Palavras 33*, pp. 27-35.

Rodrigues, Leila Calil Saade (2010). *Dificuldades de síntese na escrita de alunos do Ensino Superior Politécnico*. Dissertação de Doutoramento, Universidade de Aveiro.

Rodrigues, Sónia Valente & Purificação Silvano (2009). O desenvolvimento da competência linguística como factor de qualificação no processo de escrita. Um estudo no âmbito do projecto IELP. In

Alexandra Fiéis e Maria Antónia Coutinho (orgs.) *Textos Seleccionados. XXIV Encontro Nacional da Associação Portuguesa de Linguística*. Lisboa: APL, Colibri, pp. 437-451.

Vasconcelos, Rosa Maria, Sílvia Monteiro & Magda Pinheiro (2007). Competências de escrita em Alunos Universitários. *World Congress on Communication and Arts. 2007*. São Paulo: WCCA, pp.75-78.