Um corpo genológico de textos didáticos

A generic[1] corpus of didactical texts

Mário Amado Alves[1[0000-0003-0924-4872]], Marta Filipe Alexandre[1,2[0000-0003-2898-8762]]

maa@fl.uc.pt, marta.alexandre@ipleiria.pt

[1]CELGA-ILTEC, Universidade de Coimbra, Portugal

[2]Escola Superior de Educação e Ciências Sociais, Politécnico de Leiria, Portugal

Resumo. Apresentamos um corpo genológico de textos didáticos consistindo em aproximadamente 2500 textos didáticos anotados quanto aos seus géneros textuais e quanto à sua estruturação interna. Os textos foram retirados de 64 manuais escolares portugueses publicados. Deste conjunto de textos, uma seleção de aproximadamente 500 encontra-se transcrita e analisada pormenorizadamente, linha a linha. O corpo tem servido de base a inúmeros trabalhos publicados ou praticados, com resultados que demonstram a utilidade — mesmo necessidade — da abordagem genológica para a compreensão, transmissão e produção do texto escolar, quer por professores quer por alunos. No presente artigo descrevemos o corpo — incluindo dados inéditos sobre o mesmo —, e exemplificamos utilizações existentes e possíveis do mesmo. Elaboramos sobre a hipótese de publicação aberta do corpo, nomeadamente de tornar o corpo inteiramente público, disponível gratuitamente online, até à data da conferência. Abordamos o enquadramento legal deste passo, e exortamos a comunidade a conferir as nossas hipóteses.

Palavras-Chave: pedagogia de género, português europeu, lusofonia, corpos.

Abstract. We present a generic corpus of didactical texts consisting of approximately 2500 didactical texts annotated for their textual genres and for their internal structure. The texts were extracted from 64 commercially published Portuguese schoolbooks. Of this set of texts, a selection of approximately 500 were transcribed and delicately analysed, line by line. The corpus has served as basis for countless published works as well as workshops, with results showing the usefulness—even the necessity—of the generic approach for the understanding, transmission, and generation of didactical text, by teachers and students alike. We elaborate on the hypothesis of publishing the corpus in an open access way, making the corpus entirely public, around the time of the conference. We address the issue of the legal framing of such step, and exhort the community to confer our hypotheses.

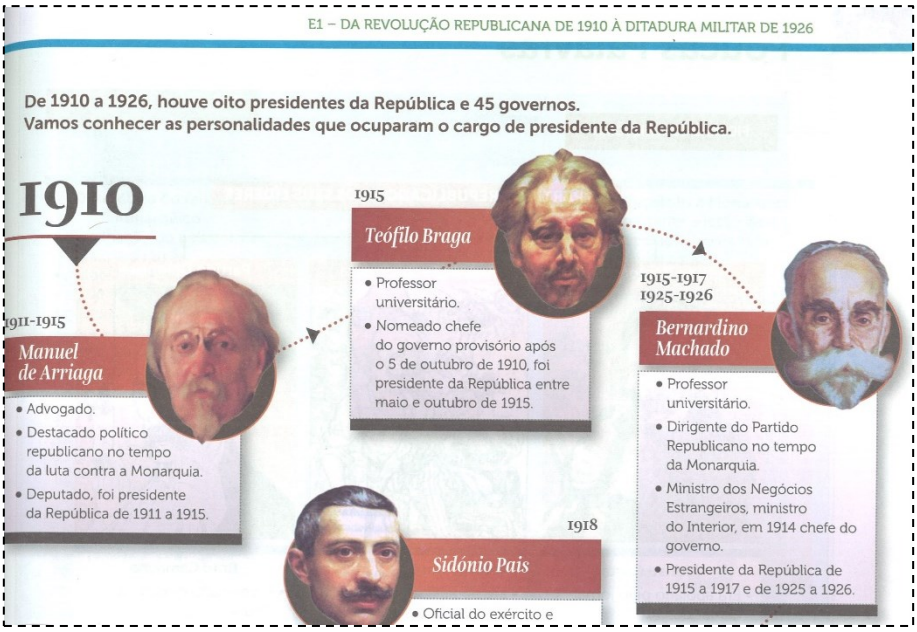
Keywords: genre pedagogy, European Portuguese, Portuguese language at large, corpora.

1 Introdução

Com o intuito de estudar, e divulgar, os *géneros escolares* (Caels, Barbeiro & Gouveia 2020) presentes na literatura didática portuguesa dos níveis básico e secundário, o CELGA-ILTEC tem, desde 2017, recolhido, classificado, e analisado um largo conjunto de textos de manuais escolares dirigidos a esses níveis de ensino (Caels & Quaresma 2019). Atualmente, o repositório consiste em aproximadamente 2500 textos, retirados de 64 manuais escolares portugueses, digitalizados e classificados quanto ao seu género discursivo. Deste conjunto de textos, uma seleção de aproximadamente 500 encontra-se transcrita e analisada delicadamente[2], como exemplificado na Figura 1.

[1] Here and throughout, *generic* refers to genre, not general.

[2] *Delicadeza* é um termo técnico da linguística sistémico-funcional, cf. referências na Secção 2. Mas significa o que parece: detalhe, finura, pormenorização. Ver os exemplos na Figura 1, etc.



Bernardino Machado

- Professor universitário
- Dirigente do Partido Republicano no tempo da Monarquia.
- Ministro dos Negócios Estrangeiros, ministro do Interior, em 1914 chefe do governo.
- Presidente da República de 1915 a 1917 e de 1925 a 1926.

Nível de ensino:
2.º ciclo do EB

Ano:
6.º

Área curricular:
Ciências Sociais

Disciplina:
História e Geografia de Portugal

Domínio:
Portugal do século XX

Subdomínio:
Da revolução republicana de 1910 à ditadura militar de 1926

Manual:
M27

Página:
123

Título	**Bernardino Machado**
Orientação	
	• Professor universitário
Registo de eventos	
dirigente	• Dirigente do Partido Republicano no tempo da Monarquia.
ministro	• Ministro dos Negócios Estrangeiros, ministro do Interior, em 1914 chefe do governo.
presidente	• Presidente da República de 1915 a 1917 e de 1925 a 1926.

Figura 1: Exemplo de texto digitalizado, transcrito, anotado.
Fonte: reprodução do manual M27 e adaptação de Caels & Quaresma (2017), p. 6

Este corpo genológico de textos didáticos tem servido de base a inúmeros trabalhos publicados ou praticados (oficinas), com resultados que demonstram a utilidade — mesmo necessidade — da abordagem genológica para a compreensão, transmissão e produção do texto escolar, quer por professores quer por alunos (cf. publicações supracitadas).

A edição, do repositório como dos artefactos trabalhados, tem sido realizada manualmente, e o acesso ao repositório tem sido interno. Então, em 2021, iniciámos um programa de tratamento informático do repositório, com os seguintes objetivos inter-relacionados:

- permitir o processamento computacional
- facilitar a consulta do corpo
- obter estatísticas totais e correlacionadas
- controlar a qualidade dos dados
- publicar o corpo online em acesso aberto e código aberto (open source)

Esta evolução tomou a forma de transformação do repositório manual, ou Forma Um, numa base de dados normalizada e validada, ou Forma Dois, através de programas de ingestão, validação, geração, e teste, construídos incrementalmente. A Figura 2 ilustra este processo, o qual se encontra descrito de modo mais detalhado, e com foco no melhoramento da qualidade dos dados, em Alves (no prelo).

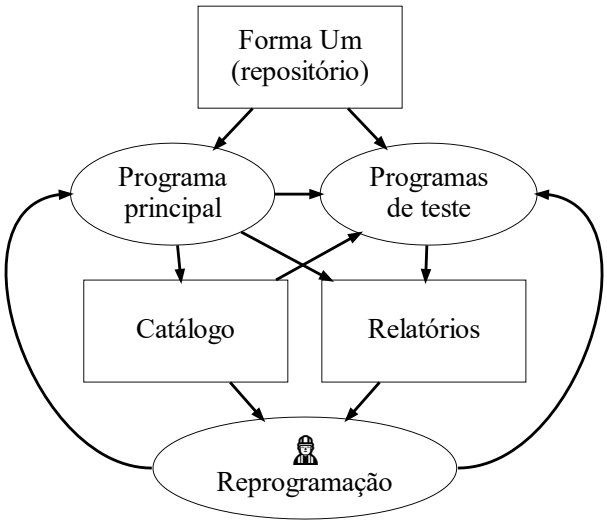


Figura 2: Processo de melhoramento incremental da forma do corpo.
(Retângulos = dados, elipses = processadores, setas = fluxo de informação.)
Fonte: Elaboração própria

A Forma Dois, concretizada no catálogo e outros artefactos, permite a exploração dos dados por qualquer dos seus muitos atributos, e a extração de estatísticas úteis, tais como as quantidades (previamente desconhecidas e mesmo impossíveis de obter) listadas na Tabela 1.

No presente artigo descrevemos o corpo — incluindo dados inéditos sobre o mesmo —, e exemplificamos utilizações existentes e possíveis do mesmo, com os seguintes objetivos:

- divulgar o corpo
- promover a abordagem genológica do texto escolar
- angariar sugestões sobre o futuro do corpo — incluindo a "questão legal"

Documentamos a abordagem genológica na Secção 2. Descrevemos a estrutura de dados do corpo na Secção 3. O corpo tem múltiplas utilizações, pedagógicas, didáticas, linguísticas, e outras, a que já aludimos nesta Introdução; elaboramos sobre um conjunto mais alargado de utilizações na Secção 4.

A publicação aberta do corpo possibilitará a investigação externa, por exemplo, por laboratórios de aprendizagem automática, ou por ciência cidadã, ou em desafios no Kaggle e outras plataformas. Esperamos tornar o corpo inteiramente público, disponível gratuitamente online, até à data da conferência; no presente artigo abordamos igualmente o enquadramento legal deste passo (Secção 5), e nas Conclusões (Secção 6) voltamos a pensar este futuro público do corpo.

Tabela 1: Extensão do corpo.

Quantidade de textos	2476	
Quantidade de textos transcritos	402	
Quantidade de parágrafos transcritos	2677	
Quantidade de ocorrências de palavra	70645	
Quantidade de ocorrências de pontuação	13373	
Quantidade de ocorrências de espaço branco	67725	
Quantidade de carateres	453399	
Quantidade de manuais	64	
Quantidade de volumes	94	
Extensão média (páginas por volume)	200	(estimativa)
Extensão total (volumes × extensão média)	18800	(100.00%)
Extensão digitalizada (páginas)	2917	(15.51%)
Extensão transcrita (páginas)	455	(2.42%)

Fonte: Elaboração própria.

2 O conceito de género escolar

O conceito de género que preside ao presente trabalho filia-se no pensamento e prática da Escola de Sydney, uma corrente enquadrada na Linguística Sistémico-Funcional. Conforme é explicitado por Caels, Barbeiro e Gouveia (2020, p. 15), a Escola de Sydney rege-se por dois objetivos principais:

- identificar e descrever os géneros associados a diferentes contextos socioculturais;
- desenvolver estratégias e intervenções pedagógicas informadas pela noção de género.

Neste âmbito particular, a motivação para o estudo dos géneros é essencialmente pedagógica e faz parte de um programa mais vasto de literacia que visa a igualdade social no acesso à educação e ao conhecimento.

Para uma história da pedagogia de base genológica em Portugal, leia-se Gouveia, Alexandre e Caels (no prelo). Uma pequena nota sobre a origem deste conceito e a diversidade de enquadramentos teóricos em que se tem desenvolvido o trabalho dos investigadores do grupo Discurso e Práticas Discursivas Académicas do CELGA-ILTEC pode ser encontrada no Portal dos Géneros Escolares & Académicos (cf. https://sites.ipleiria.pt/pge/recursos/definicoes-genero/).

Os géneros são entendidos como tipos de texto que se definem de acordo com a sua função e de acordo com a sua estrutura. A função refere-se ao papel social e comunicativo que os textos concretizam num determinado contexto. A estrutura, por seu turno, diz respeito à organização discursiva colocada ao serviço da realização dessas funções. Subjacente ao conceito de género encontra-se, pois, uma visão necessariamente contextualizada dos textos.

Segundo se lê em Martin e Rose (2008, p.6), o género corresponde a uma categoria que permite agrupar textos de acordo com as suas semelhanças e é definido como "uma configuração recorrente de significados que representa uma prática social de uma dada cultura". Destaca-se, pois, a sua natureza estruturada (um género corresponde a uma determinada configuração) e a sua dimensão funcional (um género representa uma determinada prática no contexto). Cabe ainda notar que o conceito de género tem sido aplicado a configurações recorrentes de significados não só em textos verbais (seja no discurso falado seja no escrito), como ainda em outros tipos de recursos semióticos como a imagem, a música ou a ação, por exemplo.

Dentro da diversidade de géneros que circulam numa dada cultura ou comunidade, os géneros escolares reportam às configurações discursivas recorrentes que são produzidas num contexto circunscrito: o contexto da escola. Assim, sob o escopo dos géneros escolares encontram-se os vários tipos de textos, orais e escritos, que circulam no sistema educativo, desde o ensino básico ao ensino secundário. Em contrapartida, para apontar para os tipos de textos respeitantes ao ensino superior usa-se o termo géneros académicos (cf. Silva e Santos, 2018).

Para além do interesse natural que a descrição dos géneros escolares terá para um conhecimento mais aprofundado da língua em uso, o foco assumido sobre o género decorre de uma motivação essencialmente pedagógica. Na verdade, o género é tomado como elemento basilar para a ação pedagógica e isso dá-se por uma razão fundamental: "os géneros escolares configuram um conhecimento

de literacia especializado, frequentemente implícito ou oculto no currículo, mas, ainda assim, objeto de avaliação por parte do sistema" (Caels, Barbeiro e Gouveia, 2020, p. 16). Dito de outra forma, o domínio dos géneros constitui, de facto, uma condição para o sucesso escolar.

No caso concreto do corpo de textos que é objeto do presente trabalho, foca-se um subconjunto dos géneros escolares: os textos didáticos que circulam sob a forma escrita nos manuais. Diz-se que são textos didáticos porque se consideram apenas os textos produzidos pelos autores dos manuais. Para uma visão geral das motivações e das opções metodológicas que presidiram à seleção, organização e extração dos textos, leia-se Caels e Quaresma (2019).

3 Estrutura de dados do corpo

Todos dados exceto a informação visual são alfanuméricos, isto é, os seus valores são constituídos por caracteres (letras, dígitos, pontuação, símbolos, espaços), no padrão Unicode.

3.1 Classe Texto

A principal classe de dados do corpo é o *texto.* O corpo é um conjunto de textos com as seguintes características.

Cada texto é naturalmente identificado pela sua *localização:*

- identificação do (volume do) manual escolar (cf. secção Manuais antes das Referências)
- número da(s) página(s), anexo ou destacável

Cada texto é classificado quanto ao(s) seu(s) *género(s):*[3]

1. *Descrição de propriedades*
2. *Explicação Consequencial*
3. *Explicação de eventos*
4. *Explicação Fatorial*
5. *Explicação Histórica*
6. *Explicação Sequencial*
7. *Instrução*
8. *Protocolo*
9. *Relato Biográfico*
10. *Relato de eventos*
11. *Relato de Procedimento*
12. *Relato Histórico*
13. *Relatório Classificativo*
14. *Relatório Composicional*
15. *Relatório Descritivo*

Informação adicional sobre o texto, derivada a partir da informação anterior:

- ano de escolaridade a que refere o texto (o ano está registado na descrição do manual)
- disciplina (registada na respetiva parte do manual)
- família genológica (cada género pertence a uma determinada família)

Informação adicional contingente:

- curso (nível secundário)
- área
- domínio
- subdomínio
- tópico

Informação visual: digitalização, em formato JPEG, da(s) página(s) impressas do local do texto. (Esta é a única informação que não é alfanumérica. Todas restantes informações são alfanuméricas.)

[3] Lista "documentada" ainda em validação, cf. Alves (no prelo).

3.2 Subclasse Texto Transcrito

Esta classe é uma importante extensão da anterior, contendo os seguintes atributos:

- *transcrição* do texto, na forma duma sequência de parágrafos; neste contexto, parágrafo incluí títulos, subtítulos e itens de lista, isto é, estes objetos de texto são também denominados parágrafos
- *anotação:* conjunto(s) de etiquetas associadas a segmentos do texto; um segmento pode ser um parágrafo, ou abranger múltiplos parágrafos, ou todo o texto, ou ser uma parte dum parágrafo; cada etiqueta está associada a um só segmento; mais sobre as etiquetas na subsecção dedicada

3.3 Classe Anotação

Uma anotação genológica é uma ou mais árvores de etiquetas genológicas associadas a segmentos de texto.

A raiz de cada árvore é a designação do género, um conjunto fechado de etiquetas.

Os ramos principais são *etapas* do género, um conjunto igualmente fechado de etiquetas, dependente do género.

Os sub-ramos e folhas são fases, fatores, ou outras categorias, dependentes da etapa, e formam um conjunto aberto de etiquetas. Estas etiquetas terminais criadas pelo anotador refletem diretamente o conteúdo do segmento anotado, ou a sua relação com a superestrutura (ou ambas as coisas).

3.4 Implementação

A Forma Um consiste num repositório de ficheiros de imagem (JPEG) e documentos Word. Os documentos Word contêm a transcrição e a análise, ou anotação, conforme exemplificado na Figura 1. O género e outros metadados encontram-se codificados também no nome dos ficheiros e das pastas.

Há também documentos publicados, também preparados manualmente, como o já citado Caels & Quaresma (2017), que contêm vários textos transcritos (alguns inéditos em relação ao repositório).

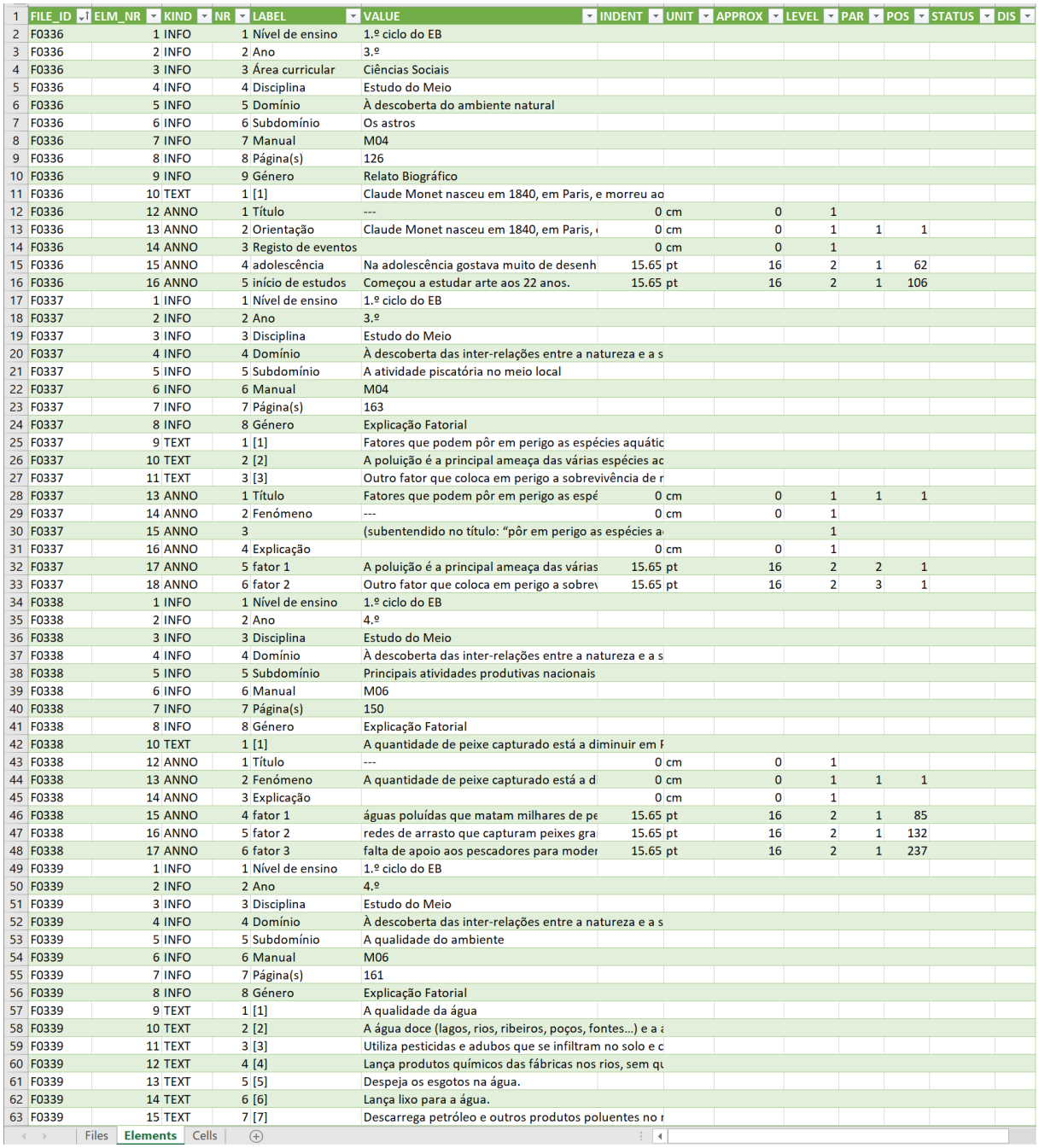
Estes dados são processados pelos programas criados para o efeito (Figura 2), criando uma base de dados normalizada de acordo com as classes acima descritas, e implementada nas entidades seguidamente descritas.

3.5 Catálogo

(Os artefactos deste projeto encontram-se identificados e descritos numa mistura de inglês e português, devido à ligação a tecnologias anglófonas.)

O catálogo da Forma Um é implementado num livro Excel *catalog.xlsx* com as folhas, ou tabelas de dados, *Files, Elements, Cells,* descritas infra. A Figura 3 dá uma visão do catálogo.

Este catálogo contém já alguns elementos da Forma Dois: identificadores uniformes; identificadores corrigidos; associação exata entre anotação e texto transcrito.



	FILE_ID	ELM_NR	KIND	NR	LABEL	VALUE	INDENT	UNIT	APPROX	LEVEL	PAR	POS	STATUS	DIS
2	F0336	1	INFO	1	Nível de ensino	1.º ciclo do EB								
3	F0336	2	INFO	2	Ano	3.º								
4	F0336	3	INFO	3	Área curricular	Ciências Sociais								
5	F0336	4	INFO	4	Disciplina	Estudo do Meio								
6	F0336	5	INFO	5	Domínio	À descoberta do ambiente natural								
7	F0336	6	INFO	6	Subdomínio	Os astros								
8	F0336	7	INFO	7	Manual	M04								
9	F0336	8	INFO	8	Página(s)	126								
10	F0336	9	INFO	9	Género	Relato Biográfico								
11	F0336	10	TEXT	1	[1]	Claude Monet nasceu em 1840, em Paris, e morreu ao								
12	F0336	12	ANNO	1	Título	---	0	cm	0	1				
13	F0336	13	ANNO	2	Orientação	Claude Monet nasceu em 1840, em Paris,	0	cm	0	1	1	1		
14	F0336	14	ANNO	3	Registo de eventos		0	cm	0	1				
15	F0336	15	ANNO	4	adolescência	Na adolescência gostava muito de desenh	15.65	pt	16	2	1	62		
16	F0336	16	ANNO	5	início de estudos	Começou a estudar arte aos 22 anos.	15.65	pt	16	2	1	106		
17	F0337	1	INFO	1	Nível de ensino	1.º ciclo do EB								
18	F0337	2	INFO	2	Ano	3.º								
19	F0337	3	INFO	3	Disciplina	Estudo do Meio								
20	F0337	4	INFO	4	Domínio	À descoberta das inter-relações entre a natureza e a s								
21	F0337	5	INFO	5	Subdomínio	A atividade piscatória no meio local								
22	F0337	6	INFO	6	Manual	M04								
23	F0337	7	INFO	7	Página(s)	163								
24	F0337	8	INFO	8	Género	Explicação Fatorial								
25	F0337	9	TEXT	1	[1]	Fatores que podem pôr em perigo as espécies aquátic								
26	F0337	10	TEXT	2	[2]	A poluição é a principal ameaça das várias espécies ac								
27	F0337	11	TEXT	3	[3]	Outro fator que coloca em perigo a sobrevivência de r								
28	F0337	13	ANNO	1	Título	Fatores que podem pôr em perigo as espé	0	cm	0	1	1	1		
29	F0337	14	ANNO	2	Fenómeno	---	0	cm	0	1				
30	F0337	15	ANNO	3		(subentendido no título: “pôr em perigo as espécies a				1				
31	F0337	16	ANNO	4	Explicação		0	cm	0	1				
32	F0337	17	ANNO	5	fator 1	A poluição é a principal ameaça das várias	15.65	pt	16	2	2	1		
33	F0337	18	ANNO	6	fator 2	Outro fator que coloca em perigo a sobrev	15.65	pt	16	2	3	1		
34	F0338	1	INFO	1	Nível de ensino	1.º ciclo do EB								
35	F0338	2	INFO	2	Ano	4.º								
36	F0338	3	INFO	3	Disciplina	Estudo do Meio								
37	F0338	4	INFO	4	Domínio	À descoberta das inter-relações entre a natureza e a s								
38	F0338	5	INFO	5	Subdomínio	Principais atividades produtivas nacionais								
39	F0338	6	INFO	6	Manual	M06								
40	F0338	7	INFO	7	Página(s)	150								
41	F0338	8	INFO	8	Género	Explicação Fatorial								
42	F0338	10	TEXT	1	[1]	A quantidade de peixe capturado está a diminuir em F								
43	F0338	12	ANNO	1	Título	---	0	cm	0	1				
44	F0338	13	ANNO	2	Fenómeno	A quantidade de peixe capturado está a d	0	cm	0	1	1	1		
45	F0338	14	ANNO	3	Explicação		0	cm	0	1				
46	F0338	15	ANNO	4	fator 1	águas poluídas que matam milhares de pe	15.65	pt	16	2	1	85		
47	F0338	16	ANNO	5	fator 2	redes de arrasto que capturam peixes gra	15.65	pt	16	2	1	132		
48	F0338	17	ANNO	6	fator 3	falta de apoio aos pescadores para moder	15.65	pt	16	2	1	237		
49	F0339	1	INFO	1	Nível de ensino	1.º ciclo do EB								
50	F0339	2	INFO	2	Ano	4.º								
51	F0339	3	INFO	3	Disciplina	Estudo do Meio								
52	F0339	4	INFO	4	Domínio	À descoberta das inter-relações entre a natureza e a s								
53	F0339	5	INFO	5	Subdomínio	A qualidade do ambiente								
54	F0339	6	INFO	6	Manual	M06								
55	F0339	7	INFO	7	Página(s)	161								
56	F0339	8	INFO	8	Género	Explicação Fatorial								
57	F0339	9	TEXT	1	[1]	A qualidade da água								
58	F0339	10	TEXT	2	[2]	A água doce (lagos, rios, ribeiros, poços, fontes...) e a a								
59	F0339	11	TEXT	3	[3]	Utiliza pesticidas e adubos que se infiltram no solo e c								
60	F0339	12	TEXT	4	[4]	Lança produtos químicos das fábricas nos rios, sem qu								
61	F0339	13	TEXT	5	[5]	Despeja os esgotos na água.								
62	F0339	14	TEXT	6	[6]	Lança lixo para a água.								
63	F0339	15	TEXT	7	[7]	Descarrega petróleo e outros produtos poluentes no r								

Files | Elements | Cells

Figura 3: Foto de ecrã do catálogo.
Fonte: Elaboração própria

3.6 Tabela *Files*

Todos ficheiros do repositório, quer *jpg*, quer *docx.* Esquema na tabela 2.

Tabela 2: Esquema da tabela *Files*

FILE_NR	Numeração sequencial dos ficheiros. A ordem é arbitrária.
FILE_ID	Identificador do ficheiro, baseado no FILE_NR.
PATH	Caminho completo do ficheiro no repositório.
NAME	Nome e extensão do ficheiro.
BASE	Nome do ficheiro, sem a extensão.
EXT	Extensão do nome do ficheiro (*jpg* ou *docx*).
PREFIX	Prefixo do nome do ficheiro com a informação NIVEL, ANO, AREA, DISCIPLINA.
FIXED	Forma corrigida automaticamente de PREFIX?
NIVEL	Nível de ensino codificado em PREFIX.
ANO	Ano de escolaridade codificado em PREFIX.
AREA	Área curricular codificada em PREFIX.
DISCIPLINA	Disciplina codificada em PREFIX.
MANUAL	Código do Manual codificado no nome do ficheiro.
PAGINA	Número de página sem letra, a primeira ou a única, codificado no nome do ficheiro.
Ímpar	PAGINA é ímpar. (Coluna acrescentada manualmente na folha. Pode não existir em certas versões.)
PAGINA_2	Última página. Pode ser igual a PAGINA (no caso de página única).
QT_PAG	Quantidade de páginas.
PAGINA_3	Número de página com letra codificado no nome do ficheiro. Todas páginas com letras são únicas.
PAGINA_4	Descritores não numéricos de páginas. Todas estas são únicas.
PAGINA_5	Designação de anexo.
LETRA	Letra da PAGINA_3.
GENERO	Descrição do género no nome do ficheiro.
GENERO_ID	Identificação uniforme do GENERO.
GENERO_F2	Designação extensa do GENERO_ID.
TRANSCRITO	O ficheiro refere um texto transcrito?
TOPICO	Designação em certos documentos externos, cf. Melhoramentos.
ID_TEXTO	Identificador do texto baseado em MANUAL e no intervalo de páginas.
MATCHED_1	Todas partes do nome do ficheiro foram reconhecidas com sucesso.
DOCUMENT	Ficheiro é um docx?
DOC_PAR_CNT	Número de parágrafos do docx.
DOC_TBL_CNT	Número de tabelas do docx.
DOC_LABELS	Etiquetas da primeira tabela do docx (conteúdo da coluna 1 da tabela).
DOC_LBL_NIV	Nível de ensino nessa tabela (conteúdo da coluna 2).
DOC_LBL_ANO	Ano de escolaridade na tabela.
DOC_LBL_DIS	Disciplina na tabela.
DOC_LBL_DOM	Domínio na tabela.
DOC_LBL_SUB	Subdomínio na tabela.
DOC_LBL_MAN	Manual na tabela.
DOC_LBL_PAG	Página(s) na tabela.
DOC_LBL_GEN	Género na tabela.
DOC_LBL_AREA	Área curricular na tabela.
DOC_PAR_1	Primeiro parágrafo da transcrição.
DOC_MORE_PARS	Restantes parágrafos.
DOC_ANN_LBL_1	Etiquetas não indentadas (segunda tabela, coluna 1).
DOC_ANN_LBL_2	Etiquetas indentadas (idem).
GEN_GEN	Enumeração GENERO, DOC_LBL_GEN
GEN_GEN_GEN	Enumeração GENERO, DOC_LBL_GEN, GENERO_ID
ID_TEXTO_GEN	Enumeração ID_TEXTO, GENERO_ID

Fonte: Elaboração própria.

3.7 **Tabela *Elements***

Elementos dos docx regulares, isto é, documentos que têm a forma exemplificada na Figura 4. Assim, a tabela *Elements* representa os textos transcritos. A tabela é polimórfica: o valor de certos campos depende do tipo de item, INFO, TEXT, ANNO, identificado em KIND. Esquema na Tabela 3.

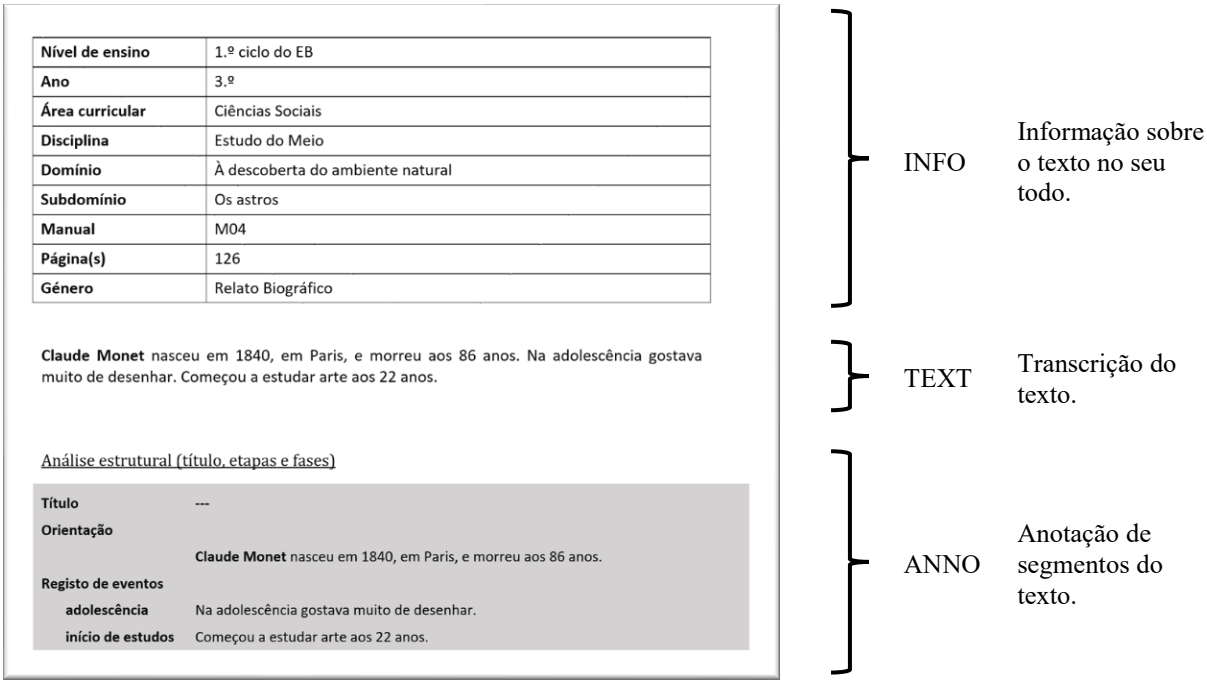


Figura 4: Forma regular de documento de transcrição e anotação.
Fonte: Elaboração própria

Tabela 3: Esquema da tabela *Elements*

FILE_ID	Identificação do ficheiro docx. Congruente com folha *Files.*		
ELM_NR	Número sequencial do elemento em cada ficheiro.		
KIND	Tipo de elemento (INFO, TEXT, ANNO)		
NR	Número sequencial do elemento em cada tipo.		
	KIND = INFO	KIND = TEXT	KIND = ANNO
LABEL	Nome da informação (conteúdo da coluna 1)	Número sequencial do parágrafo transcrito (= NR)	Etiqueta de anotação (conteúdo da coluna 1)
VALUE	Valor da informação (conteúdo da coluna 2)	Texto do parágrafo transcrito	Segmento anotado (conteúdo da coluna 2)
	KIND = ANNO		
INDENT	Valor da indentação (alinhamento à esquerda).		
UNIT	Unidade da indentação.		
APPROX	Valor aproximado da indentação.		
LEVEL	Nível da anotação na hierarquia.		
PAR	Número do parágrafo transcrito a que pertence o segmento.		
POS	Posição de início, no parágrafo transcrito, do segmento anotado. (Em bytes?)		
STATUS	(?)		
DIS	Distância de Levenshtein(?) entre segmento e transcrição.		

Fonte: Elaboração própria.

3.8 Tabela *Cells*

Todos elementos dos docx, regulares ou não. Maioritariamente são células de tabelas, donde o nome. Inclui todos parágrafos e todas células de todas tabelas de todos docx. É útil para examinar docx irregulares, que não estão representados em *Elements*, mas que mesmo assim podem conter material transcrito. Os dados *Elements* são derivados a partir de *Cells*.

Esta tabela é ligeiramente polimórfica, no número de tabela TBL: 0 = não tabela = parágrafo; caso em que ROW e COL não têm significado (contêm o valor 0 meramente técnico). Esquema na Tabela 4.

Tabela 4: Esquema da tabela *Cells*

FILE_ID	Identificação do ficheiro docx. Congruente com tabela *Files.*
ITEM_NR	Número sequencial do item (parágrafo ou célula) em cada ficheiro.
TBL	Número sequencial da tabela em cada ficheiro. 0 = não tabela = parágrafo
ROW	Número sequencial da linha em cada tabela. 0 = não aplicável
COL	Número sequencial da coluna em cada linha de tabela. 0 = não aplicável
ALIGN	Valor do alinhamento à esquerda (indentação).
NR	Só o número de ALIGN.
UNIT	Unidade de ALIGN (pontos, centímetros...)
VALUE	Conteúdo da célula ou texto do parágrafo.

Fonte: Elaboração própria.

3.9 Forma Dois

A Forma Dois realiza plenamente a centralidade da classe *texto* — que na Forma Um se encontra "emaranhada" em ficheiros, tipos de ficheiro, e pastas. Novos programas estão sendo criados que, a partir do Catálogo, geram artefactos completamente orientados pelas "verdadeiras" classes de dados: texto, género, manual, anotação, etc. Seguem-se alguns exemplos de tais artefactos.

3.10 Hipertexto

O hipertexto do corpo contém índices pré-coordenados, cruzando ou empilhando categorias úteis, e dados estatísticos totais ou cruzados. As Figuras 5-7 mostram a página principal, o índice genológico, e um texto transcrito.

Este hipertexto ainda não está disponível na web pública, cf. Secção 5.

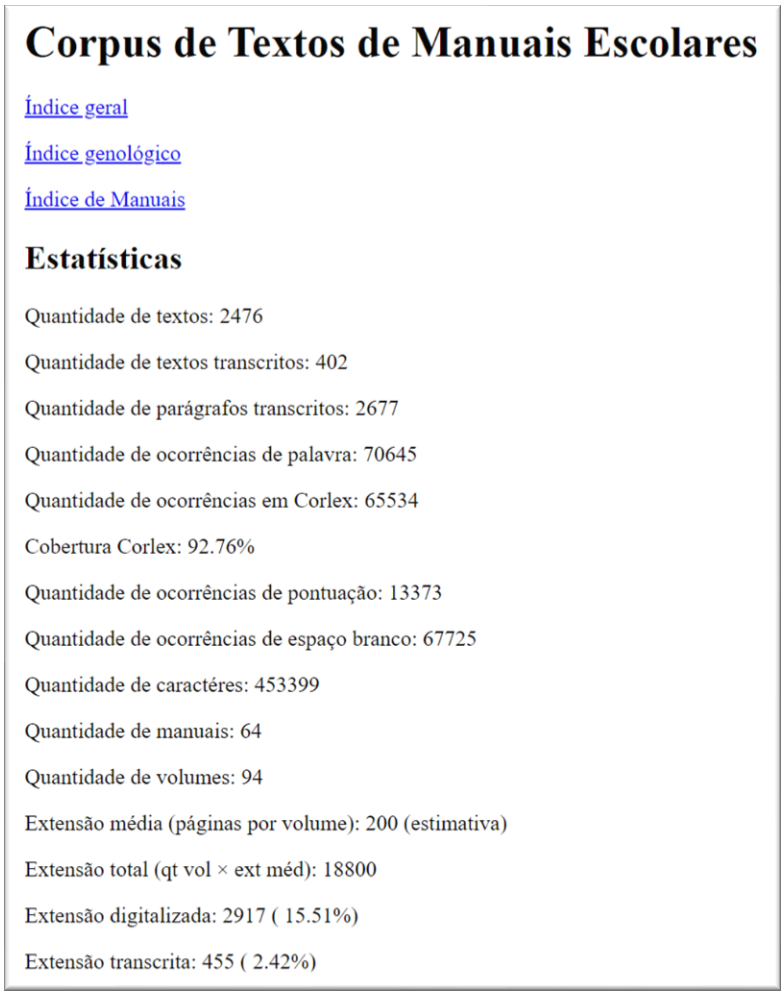


Figura 5: Hipertexto. Página principal.
Fonte: Elaboração própria

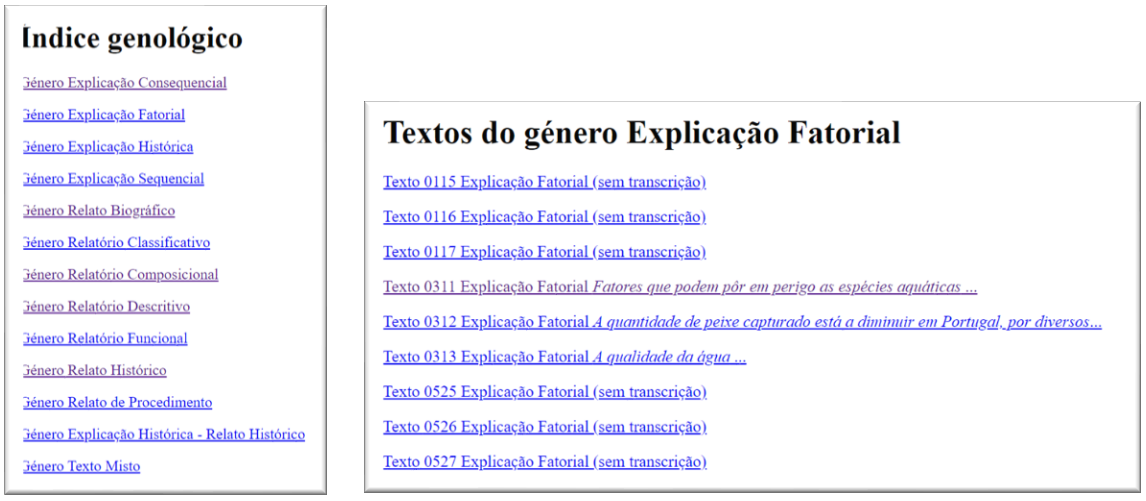


Figura 6: Hipertexto. Índice genológico, e uma das respetivas páginas.
Fonte: Elaboração própria

Figura 7: Hipertexto. Um texto transcrito.
Fonte: Elaboração própria

3.11 **Protótipo *Interactive Synchronized Highlighting (ISH)***

A visualização de anotações linguísticas é uma questão complexa — e um tópico quente, cf. Aralikatte, R. and Søgaard, A. (2020), Kaplan, D. and Iida, R. and Tokunaga, T. (2010), Neves, M. and Ševa, J. (2021), Yang, J., Zhang, Y., Li, L. (2018).

Perante a inadequação, à nossa estrutura de dados, de todas abordagens existentes, desenvolvemos a seguinte interface baseada em hipertexto, designada *Interactive Synchronized Highlighting (ISH)* (destaque interativo sincronizado). Trata-se dum protótipo, ainda não automaticamente integrado com a base de dados do corpo.

As Figuras 8 e 9 ilustram estados de utilização desta interface. Trata-se dum texto transcrito real do corpo. As seleções fazem-se por mero posicionamento do ponteiro do rato sobre o texto, ou sobre as etiquetas: as partes envolvidas do texto ou da anotação destacam-se automática e sicronizadamente. Convidamos o leitor a experimentar a interface "viva" em:

`https://marius-linguist.github.io/ISH_Prototype/`

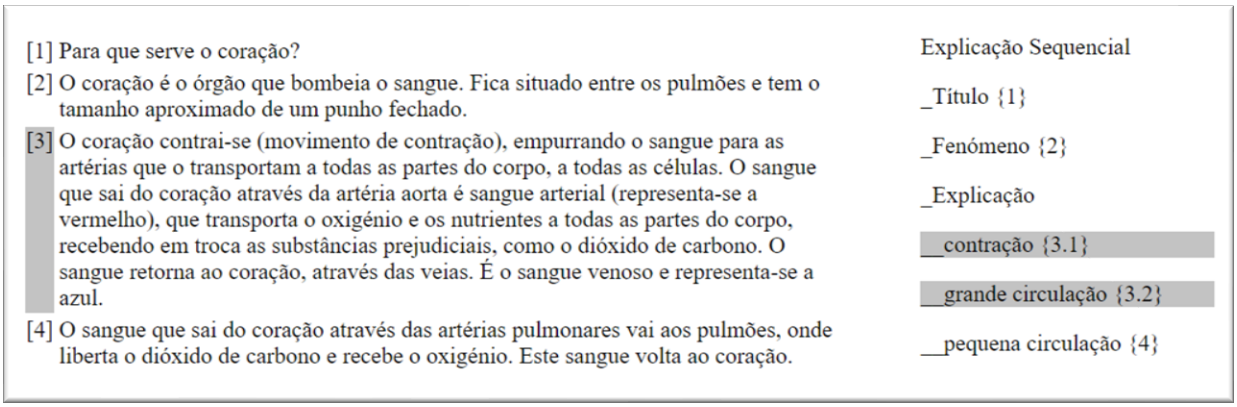


Figura 8: Parágrafo selecionado à esquerda, automaticamente destacando as respetivas etiquetas à direita.
Fonte: Elaboração própria

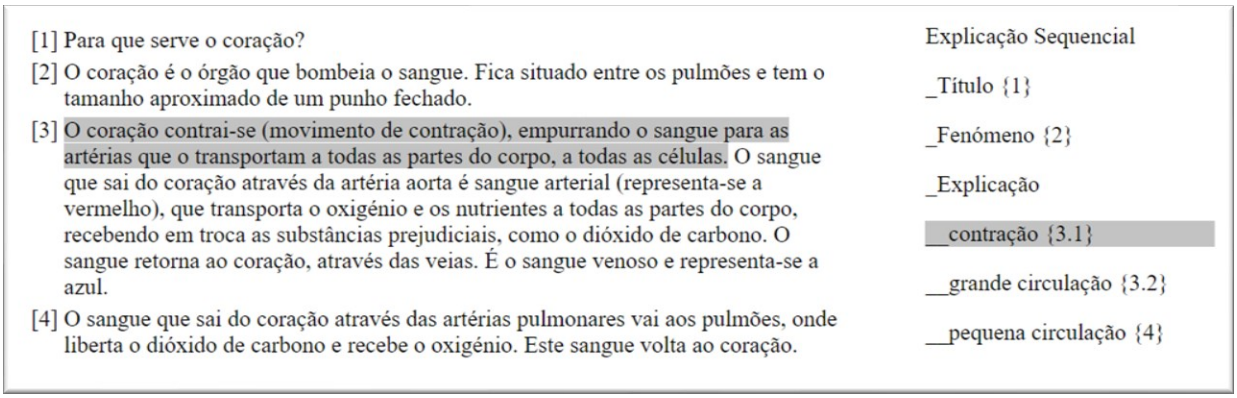


Figura 9: Segmento selecionado, automaticamente destacando a respetiva etiqueta.
Fonte: Elaboração própria

4 Exemplos de utilização

Não há uma única forma de uso do corpus que possa ser dada como a mais frequente ou representativa junto dos investigadores do grupo. Pelo contrário, ao recensear os vários exemplos de utilização dos textos, fica evidente que as motivações analíticas e pedagógicas têm levado, desde 2017 até à data, não apenas à exploração de diferentes aspetos dos textos, mas também ao uso e apresentação diversificada dos próprios textos.

Com o intuito de documentar os géneros escolares de três áreas-chave (Português, Ciências Naturais e História) foi feito um recenseamento dos géneros. A partir das digitalizações das páginas dos manuais, fez-se as transcrições manuais dos textos verbais e, com o intuito de apresentar um conjunto de textos modelo, que pudessem ser tomados como exemplos completos e bem estruturados dos géneros, desenvolveu-se trabalho sobre essas mesmas transcrições. O uso das transcrições serviu de base a sistematizações complementares como, por exemplo, a listagem dos assuntos (e.g. Caels & Quaresma, 2018, p. 5) ou a representação diagramática da distribuição dos conteúdos ao longo da estruturação dos textos (e.g. Caels & Quaresma, 2018, p. 10).

Significativamente, o conjunto de publicações especificamente focadas nesta linha de trabalho, entendidas como recursos pedagógicos e designadas como 'brochuras', incluem a reprodução de transcrições textuais (onde se podem ler os textos verbais sob uma forma gráfica diferente), a par de tabelas e diagramas (onde se podem ler dados relacionados com os textos verbais transcritos), bem como reproduções de segmentos multimodais das páginas (onde se podem ler fragmentos coloridos das páginas propriamente ditas).

A natureza multimodal dos textos dos manuais escolares foi, aliás, objeto de análise por si só, conforme se pode ler em Quaresma e Caels (2020). Neste caso, tomam-se as imagens das páginas digitalizadas para se poder descrever, analisar e comparar a organização e disposição dos elementos

verbais e não verbais no espaço das páginas de manuais de diferentes disciplinas (História vs. Ciências Naturais). Assim, o exemplo de utilização são os ficheiros de imagem e o interesse despertado por aprofundar esta linha de análise justifica a procura de possibilidades de tratamento e anotação dos ficheiros das imagens.

Num artigo recente, Alexandre e Caels (no prelo) apresentam o ponto de situação relativamente à linha de pesquisa sobre a didática da História, deixando entrever outras abordagens, algumas ainda exploratórias, sempre aplicadas sobre uma pequena seleção de textos. Esta estratégia tem ajudado a contornar as dificuldades de tomar o corpus como um todo, decorrente da necessidade de corrigir e concluir as anotações e da dimensão dos dados.

O tratamento quantitativo de dados extraídos a partir do corpus encontra-se concretizado em Barbeiro, Caels e Quaresma (2017) e a análise encetada em Caels, Quaresma e Barbeiro (2017) incluiu, numa fase posterior, ferramentas estatísticas. O trabalho resultante encontra-se em processo de revisão científica.

As novas formas do corpo, especialmente o Catálogo, já sustentaram alguns novos estudos, por exemplo Alves et al. (no prelo), ao permitir localizar os respetivos textos, e facilitar a sua análise.

4.1 Dimensões dos textos

A Forma Dois também permite conhecer várias dimensões dos textos, dando resposta a questões de investigação tais como:

- os textos de anos mais avançados são mais longos?
- têm vocabulário mais rico?
- textos de humanidades têm mais marcadores discursivos?
- têm géneros mais delicados?
- etc.

As variáveis disponíveis incluem:

- quantidade de parágrafos
- quantidade de carateres
- quantidade de palavras
- quantidade de sinais de pontuação
- quantidade de marcadores discursivos
- quantidade de níveis de delicadeza
- *type/token ratio*
- etc.

Algumas variáveis podem ser cruzadas entre si, por exemplo carateres por parágrafo. Todas variáveis resultantes podem ser agregadas: quer por texto; quer por subconjunto de textos, por exemplo todos textos de cada categoria. As categorias incluem, como já vimos (Secção 3.1):

- ano de escolaridade
- nível de ensino
- área
- género
- família genológica

Este é um trabalho que está a decorrer. Resultados preliminares incluem a observação de que os textos dos 11º e 12º anos têm menos parágrafos (Figura10), mas estes são mais longos (Figura 11), que os textos do 10º ano.

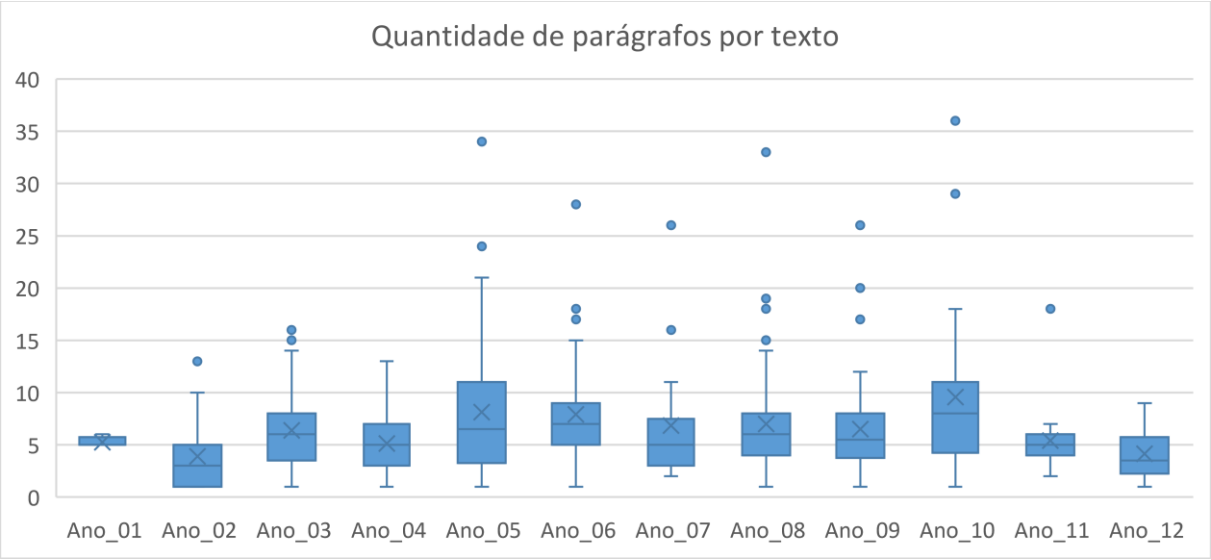


Figura 10. Exemplo de dimensão dos textos: quantidade de parágrafos por texto, por ano de escolaridade.
Fonte: Elaboração própria.

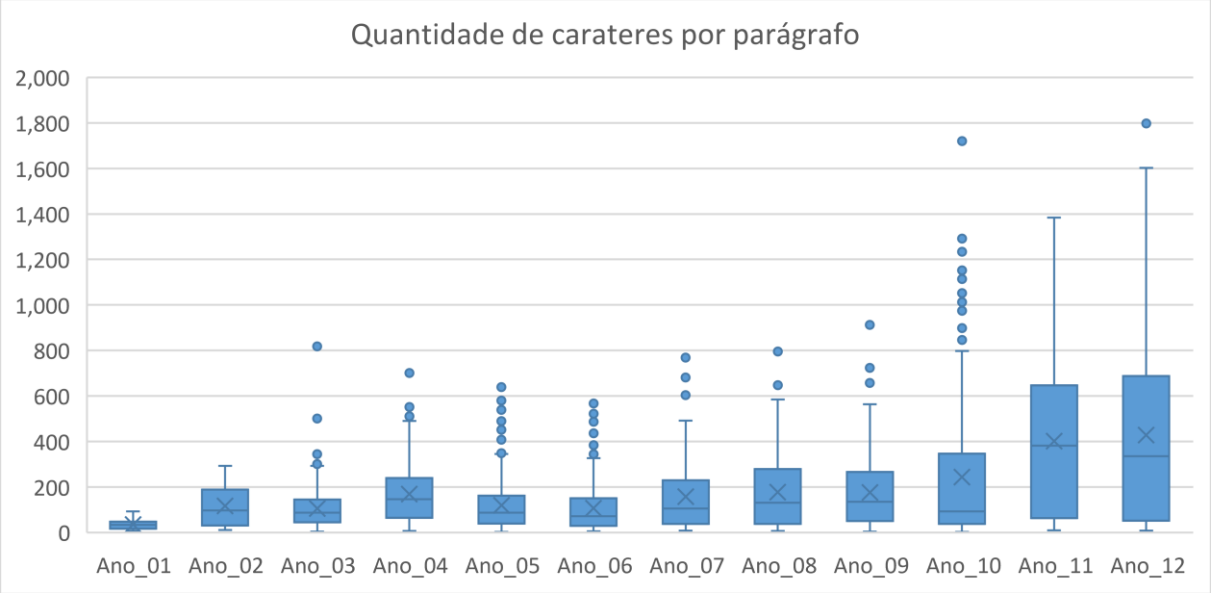


Figura 11. Exemplo de dimensão dos textos: quantidade de carateres por parágrafo, por ano de escolaridade.
Fonte: Elaboração própria.

4.2 Qualidade dos dados

A Forma Dois também permite cruzar dados para verificação dos mesmos. Cruzamentos úteis incluem:

- Quantidade de textos por categoria.
- Anos por nível.
- Etc.

Por exemplo, o cruzamento de anos por nível permitiu detectar o erro na Forma Um de classificar um texto do 4º ano como pertencendo ao nível EB23 (2º e 3º ciclos do Ensino Básico) (Figura 12).

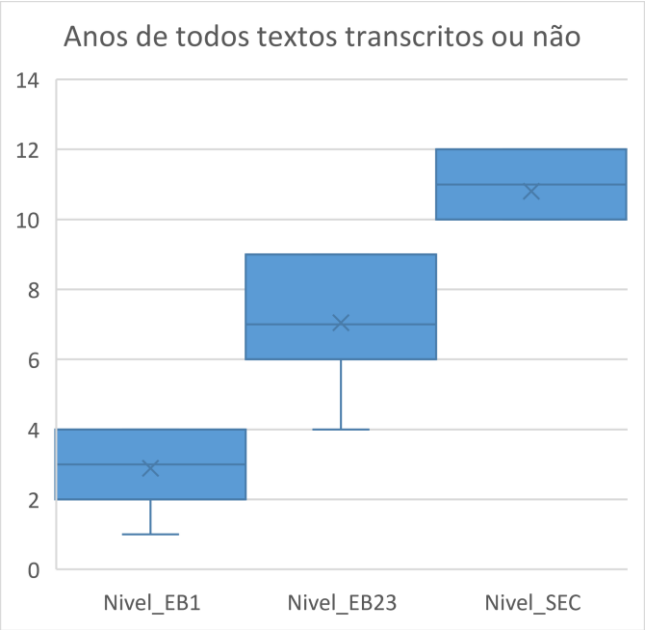


Figura 12. Exemplo de cruzamento de dados para verificação: anos por nível.
Fonte: Elaboração própria.

Todos erros detetados, são, obviamente, corrigidos, pelo processo incremental exposto na Secção 1. Tal abordagem é indissociável do nosso objetivo de manter e fornecer um corpo de alta qualidade e usabilidade, representativo e fiável.

5 Visão de publicação integral

Pretendemos publicar todo o corpo genológico de textos didáticos, em acesso aberto, gratuito, e online, a semelhança do que que já fizemos com outros corpos, cf. `http://celga-iltec.uc.pt/` (Recursos / Corpora). Dois objetivos imediatos presidem a esta visão:

- estender a utilização direta do corpo, a todo o público (mas especialmente a professores e alunos)
- facilitar a investigação externa, por exemplo, por laboratórios de inteligência artificial, linguística computacional, ou ciência cidadã

Esta visão de publicação foi e continua a ser uma das motivações para o desenvolvimento da Forma Dois, ao consideramos esta forma um requisito técnico indispensável para a qualidade da mesma hipotética publicação.

5.1 Enquadramento legal

Resolvidas como estão as questões técnicas, encaramos uma última dificuldade a transpor, de carácter legal. Uma vez que o corpo contém material de 64 manuais escolares comercializados, é necessário assegurar a legalidade da publicação do corpo relativamente ao direito de autor e normas conexas.

Em verdade, uma das nossas expetativas ao publicar o presente artigo, é que o mesmo possa ter o efeito de gerar ideias, na comunidade, de soluções para esta dificuldade, baseadas na experiência, ou mesmo em conhecimento jurídico especializado. Ou confirmar, ou infirmar, a seguinte hipótese.

5.2 Hipótese do artigo científico

A publicação online do corpo assemelhar-se-ia a um volumoso artigo científico em acesso aberto. Assim, beneficiaria de igual quadro legal, nomeadamente a permissão de publicação de extratos para efeitos didáticos ou de investigação, sem necessidade de autorização explícita.

Na verdade, o corpo já conta com a publicação em acesso aberto online, de aproximadamente 120 textos transcritos (sem a imagem digitalizada), a maior parte na coleção *Textos Modelo* (Caels & Quaresma, 2017). Tal publicação representa cerca de 30% de todas transcrições do corpo (que por sua vez representam somente 2,5% da extensão total dos manuais do corpo).

6 Conclusões

O corpo genológico de textos didáticos aqui apresentado foi, e continua a ser, um recurso útil, mesmo indispensável, à pedagogia genológica portuguesa. Tanto quanto sabemos, é um corpo único na lusofonia, se não no mundo. Uma possibilidade de evolução é a sua expansão em direção à lusofonia propriamente dita, ao português pluricêntrico, por agregação de novos dados oriundos das variedades desta língua.

Outra hipotética via de expansão seria em direção ao texto académico em geral, incluindo o nível superior, talvez por integração com corpos já existentes também geridos por nós, por exemplo, o *Corpus de Português Académico*[4].

A publicação integral do corpo em acesso aberto alavancaria estas e outras evoluções. O presente trabalho de construção da Forma Dois foi grandemente motivado por tal potencialidade. Possa o presente artigo despertar igual interesse na comunidade, e originar a necessária crítica, incluindo ideias de como circular no labirinto legal do direito de autor português, lusófono, e mundial.

Agradecimentos

Toda a substância do corpo genológico de textos didáticos existe já na sua Forma Um, um corpo constituído no seio do grupo de trabalho DPDA do CELGA-ILTEC, decorrente da colaboração entre os vários membros, em especial: Ângela Quaresma, Carlos A. M. Gouveia, Fausto Caels, Joana Vieira Santos, Helga Arnauth, Luís Filipe Barbeiro, Paulo Nunes da Silva, para além do segundo autor. É uma grande honra e gosto poder aqui expressar gratidão aos nossos bons colegas.

O primeiro autor do presente artigo é também devedor à Adacore (adacore.com) e ao seu programa GAP (GNAT Academic Program), pelo formidável compilador de Ada, e apoio. Uma nota de gratidão especial à memória de Robert Dewar, criador da indispensável biblioteca GNAT.Spitbol

O presente trabalho foi realizado na unidade de investigação CELGA-ILTEC, FCT UID 4887.

Manuais escolares abrangidos pelo corpo

Carvalho, M. J. (2016) *Todos Juntos – Estudo do Meio – 1º Ano – Manual*. Lisboa: Santillana.	M01
Letra, C. & Afreixo, A. M. (2011) *Mundo da Carochinha – Estudo do Meio – 2º ano – Manual*. Alfragide: Gailivro.	M02
Lima, E., Barrigão, N., Pedroso, N. & da Rocha, V. (2016) *Alfa – Estudo do Meio – 2.º Ano – Manual*. Porto: Porto Editora.	M03
Guimarães, D. & Alves, S. (2012) *Desafios – Estudo do Meio – 3º ano – Manual*. Lisboa: Santillana.	M04
Letra, C. & Afreixo, A. M. (2012) *Carochinha – Estudo do Meio – 3º ano – Manual*. Alfragide: Gailivro.	M05
Neto, F. P. (2013) *Despertar – Estudo do Meio – 4º ano – Manual*. Maia: Edições Livro Directo.	M06
Pires, P. & Gonçalves, H. (2013) *A Grande Aventura – Estudo do Meio – 4º ano – Manual*. Alfragide: Texto Editora.	M07
Lopes, A., Brandão, D., Mendes, J. & Vaz, S. (2016) *100% Vida – Ciências Naturais – 5º Ano – Manual*. Alfragide: Texto Editora.	M08

[4] `http://celga-iltec.uc.pt/` (Recursos / Corpora / CPA). Este corpo encontra-se também em atualização, com novos dados de 2019-2021.

Valente, B., Feio, M. Pacheco, I. Pereira, P. & Gomes, J. (2016) *Biosfera – Ciências Naturais – 5º Ano – Manual*. Vila Nova de Gaia: Edições Asa.	M09
de Sales, A., Portugal, I. & Morim, J. A. (2011) *Clube da Terra – Ciências Naturais – 6º Ano – Manual*. Alfragide: Texto Editora.	M10
Marcelino, S., Magalhães, V. & Morais-Pequeno, R. (2011) *Fazer Ciência – Ciências Naturais – 6º ano – Manual*. Alfragide: Edições Sebenta.	M11a M11b
Carrajola, C., Martin, L. & Hilário, T. (2014) *Desafios – Ciências Naturais – 7º ano – Manual*. Lisboa: Santillana.	M12
Ribeiro, E., Silva, J. C. & Oliveira, O. (2014) *Ciência & Vida – Ciências Naturais – 7º ano – Manual*. Vila Nova de Gaia: Edições Asa.	M13
Delgado, Z., Canha. P. & Trinca, C. B. (2014) *À Descoberta da Vida – Ciências Naturais – 8º ano – Manual*. Alfragide: Texto Editora.	M14
Oliveira, O., Ribeiro, E. & Silva, J. C. (2014) *Ciência & Vida – Ciências Naturais – 8º Ano – Manual*. Vila Nova de Gaia: Edições Asa.	M15
Campos, C. & Dias, M. (2015) *Terra CN – Ciências Naturais – 9º Ano – Manual*. Alfragide: Texto Editora.	M16
Oliveira, O., Ribeiro, E., & Silva, J. C. (2015) *Ciência & Vida – Ciências Naturais – 9º Ano – Manual*. Vila Nova de Gaia: Edições Asa.	M17
da Silva, A. D., Santos, M. E., Gramaxo, F., Mesquita, A. F., Baldaia, L. & Félix, J. M. (2016). *Terra, Universo de Vida – Biologia e Geologia – 10.º Ano – Manual*. Porto: Porto Editora.	M18a M19b
Ferreira, J. (2007) *Planeta com Vida – Biologia E Geologia (N) – 10º ano – Manual*. Lisboa: Santillana.	M19a M19b
Ferreira, J. & Carrajola, C. (2008) *Planeta com Vida – Biologia E Geologia (N) – 11º ano – Manual*. Lisboa: Santillana.	M20a M20b
Matias, O., Martins, P., Dias, A. G., Guimarães, P. & Rocha, P. (2016) *Biologia e Geologia 11 – 11.º Ano – Manual*. Porto: Areal Editores.	M21a M21b
da Silva, A. D., Santos, M. E., Mesquita, A. F., Baldaia, L. & Félix, J. M. (2016) *Terra, Universo de Vida – Biologia – 12.º Ano*. Porto: Porto Editora.	M22
Ribeiro, E., Silva, J. C. & Oliveira, O. (2009) *Manual Biodesafios – Biologia 12º ano*. Vila Nova de Gaia: Edições Asa.	M23
Carrajola, C., Castro, M. & Hilário, T. (2009) *Planeta com vida – Biologia – 12º ano.* Lisboa. Santillana	M24
Matias, A., Oliveira, A. R. & Cantanhede, F. (2016) *Novo HGP 5 – História e Geografia de Portugal – 5º Ano – Manual*. Alfragide: Texto Editora.	M25
Sousa, L., Soares, L. & Albino, M. (2016) *Máquina do Tempo – História e Geografia de Portugal – 5º Ano – Manual*. Vila Nova de Gaia: Edições Asa.	M26
Barreira, A., Moreira, G., Moreira, M. & Rodrigues, T. (2011) *HistGeo – História E Geografia De Portugal – 6º ano – Manual*. Vila Nova de Gaia: Edições Asa.	M27
Oliveira, A. R. & Cantanhede, F. (2015) *Novo HGP – História e Geografia de Portugal – 6º Ano – Manual*. Alfragide: Texto Editora.	M28
Oliveira, A. R., Cantanhede, Catarino, I. F., Gago, M. & Torrão, P. (2014) *O Fio Da História – História – 7º ano – Manual*. Alfragide: Texto Editora.	M29
Santos, L. A., Santos, L. A., Neto, J. & Neto, H. (2014) *Desafios – História – 7º ano – Manual*. Lisboa: Santillana.	M30
Barreira, A. & Moreira, M. (2014) *Páginas Da História – História – 8º ano – Manual*. Vila Nova de Gaia: Edições Asa.	M31
Oliveira, A. R., Catarino, I., Cantanhede, F., Gago, M. & Torrão, P. (2014) *O Fio Da História – História – 8º ano – Manual*. Alfragide: Texto Editora.	M32
Barreira, A., Moreira, M. & Rodrigues, T. (2015) *Páginas da História – História – 9º Ano – Manual*. Vila Nova de Gaia: Edições Asa.	M33
Neto, H., Santos, L. A., Cruz, T., Santos, L. A. & Neto, J. (2015) *Desafios – História – 9º Ano – Manual*. Lisboa: Santillana.	M34
Veríssimo, H., Lagarto, M. & Barros, M. (2013) *História em Perspetiva – História A – 10º ano – Manual*. Vila Nova de Gaia: Edições Asa.	M35a M35b M35c
Costa, A., Gago, M. & Marinho, P. (2013) *O Horizonte Da História – História A – 10º ano – Manual*. Alfragide: Texto Editora.	M36a M36b M36c
Costa, A., Gago, M., & Marinho, P. (2014) *O Horizonte da História – História A – 11º ano – Manual*. Alfragide: Texto Editora.	M37a M37b

	M37c
Veríssimo, H., Lagarto, M. & Barros, M. (2014) *História Em Perspetiva – História A – 11º ano – Manual*. Vila Nova de Gaia: Edições Asa.	M38a M38b M38c
Fortes, A., Gomes, F. F. & Fortes, J. (2016) *Linhas da História 12 – História A – 12.º Ano*. Porto: Areal Editores.	M39a M39b M39c
Veríssimo, H., Lagarto, M. & Barros, M. (2008) *Nova Construção da História – História 12º – Manual*. Vila Nova de Gaia: Edições Asa.	M40a M40b M40c
Lopes, C. G. & Lima, M. C. S. (2016) *Os Fantásticos – Português – 1º Ano – Manual*. Alfragide: Gailivro.	M41
Marques, M. J. & Gonçalves, C. (2014) *Desafios – Português – 2º ano – Manual*. Lisboa: Santillana.	M42
Melo, P. & Costa, M. (2014) *A Grande Aventura - Português - 2º ano - Manual.* Alfragide: Texto Editores.	M43
Neto, F. P. (2011) *Os Tagarelas - Português – 2º ano – Manual*. Edições livro direto.	M44
Letra, C. & Borges, M. (2012) *Língua Portuguesa – 3º ano – Manual.* Alfragide: Gailivro.	M45
Marques, M. J. & Gonçalves, C. (2012) *Desafios – Português – 3º ano – Manual*. Lisboa: Santillana.	M46
Letra, C. & Borges, M. (2013) *Português – 4º ano – Manual.* Alfragide: Gailivro.	M47
Melo, P. & Costa, M. (2013) *A Grande Aventura – Português – 4º ano – Manual*. Alfragide: Texto Editora.	M48
Costa, M. J. & Traça, M. E. (2008) *Passa Palavra – 5º ano, Língua Portuguesa.* Porto: Porto Editora.	M49
Lopes, M. C. & Rola, D. N. (2010) *Novo Português em Linha, Língua Portuguesa – 5º ano – Manual.* Lisboa: Plátano Editora.	M50
Santiago, A. & Paixão, S. (2014) *P6 – Português – 6º ano – Manual*. Alfragide: Texto Editora.	M51
Variz, C., Romão, S. & Dias, L. S. (2014) *Desafios – Português – 6º ano – Manual*. Lisboa: Santillana.	M52
Ferreira, I. L., Delgado, I., Mendes, R. & Lopes, M. F. (2011) *Desafios – Português – 7º ano – Manual*. Lisboa: Santillana.	M53
Villas-Boas, A. & Vieira, M. (2011) *Entre Palavras – Português – 7º Ano – Manual*. Alfragide: Edições Sebenta.	M54
Santiago, A. & Paixão, S. (2014) *P8 – Português – 8º ano – Manual*. Alfragide: Texto Editora.	M55
Silva, I. & Marques, C. (2014) *Contos & Recontos – Português – 8º ano – Manual*. Vila Nova de Gaia: Edições Asa.	M56
Ferreira, I. L., Delgado, I. & Mendes, R. (2013) *Desafios – Português – 9º ano – Manual*. Lisboa: Santillana.	M57
Marques, C. & Silva, I. (2013) *Letras & Companhia – Português – 9º ano – Manual*. Vila Nova de Gaia: Edições Asa.	M58
Freitas, E. M., Ferreira, I. G. & Barbosa, M. L. (2015) *O Caminho das Palavras – Português – 10º ano.* Porto: Areal Editores.	M59
Pinto, E., Fonseca, P. & Baptista, V. (2015) *Novo Plural – Português – 10º ano.* Lisboa: Raiz Editora.	M60
Magalhães, O. & Costa, F. (2013) *Entre Margens – Português – 10º ano.* Porto: Porto Editora.	M61
Silva, P., Cardoso, E. & Moreira, M. C. (2014) *Expressões – Português – 11º ano.* Porto: Porto Editora.	M62
Martins, F. & Moura, G. B. (2012) *Página Seguinte – Português – 12º ano – Manual*. Alfragide: Texto Editora.	M63
Peixoto, M. J., Fonseca, C. & Cardoso, A. M. (2012) *Com Textos – Português – 12º ano – Manual*. Vila Nova de Gaia: Edições Asa.	M64

Referências

Alexandre, M. F., & Caels, F. (no prelo). Investigação sobre o discurso da História em Portugal: um ponto de situação. In 2º Encontro Nacional sobre Discurso Académico, 2021. Volume em preparação.

Alves, M. A. (no prelo). Development of a digital corpus of Portuguese didactical texts for language research. In CISTI'2022 - 17th Iberian Conference on Information Systems and Technologies, Technical University of Madrid (UPM), Madrid, Spain. To appear in IEEExplore.

Alves, M. A.; Alexandre, M. F.; Caels, F. (no prelo). A avaliatividade nos manuais de história: análise exploratória. In 2º Encontro Nacional sobre Discurso Académico, 2021. Volume em preparação.

Aralikatte, R. and Søgaard, A. (2020). Model-based annotation of coreference. In Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 74–79, Marseille, 11–16 May 2020, European Language Resources Association (ELRA)

Barbeiro, L. F., Caels, F., & Quaresma, A. (2020). Géneros textuais e interdisciplinaridade nas Aprendizagens Essenciais. . In D. Alves, H. G. Pinto, I. S. Dias, M.ª O. Abreu & R. Gillain (Orgs.), Atas da IX Conferência Internacional Investigação, Práticas e Contextos em Educação, pp. 82-9. ESECS-Politécnico de Leiria.

Caels, F. & Quaresma, A. (2017). Exemplos Textuais dos géneros de História 2.º e 3.º ciclos do Ensino Básico. Dezembro de 2017. https://sites.ipleiria.pt/pge/ (Portal dos Géneros Escolares & Académicos / Recursos Pedagógicos / Textos Modelo)

Caels, F., & Quaresma, A. (2018). Os géneros das Ciências Naturais, 2.º e 3.º ciclos do Ensino Básico: Explicação Sequencial. CELGA-ILTEC. ISBN: 978-989-20-9133-4

Caels, F. & Quaresma, A. (2019). Caracterização dos géneros do Ensino Básico e Secundário. In F. Caels, L. F. Barbeiro & J. V. Santos (orgs.), *Discurso académico: Uma área disciplinar em construção* (pp. 108-133). CELGA-ILTEC da Universidade de Coimbra, ESECS-Politécnico de Leiria.

Caels, F., Barbeiro, L. F. & Gouveia, C. A. M. (2020). Géneros escolares segundo a Escola de Sydney: Propósitos, estruturas e realizações textuais. *Indagatio Didactica 12(2*): 13-32.

Caels, F., Quaresma, A. & Barbeiro, L. F. (2017, 27-29 setembro). Sobre o papel da língua nas taxonomias científicas. (comunicação oral). XXXIII Encontro Nacional da Associação Portuguesa de Linguística. Universidade de Évora, Évora.

Gouveia, C. A. M., Alexandre, M. F. & Caels, F. (no prelo). Learning to use R2L in Portugal. In Rose, D. & Acevedo, C. (eds.), *Reading to learn, reading the world: How genre-based literacy pedagogy is democratizing education*. Equinox.

Kaplan, D. and Iida, R. and Tokunaga, T. (2010). Annotation Process Management Revisited. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), 2010 may 19-21, Valletta, Malta, European Language Resources Association (ELRA)

Martin, J. R. & Rose, D. (2008). *Genre Relations: Mapping Culture*. Equinox.

Neves, M. and Ševa, J. (2021). An extensive review of tools for manual annotation of documents. Briefings in Bioinformatics, Volume 22, Issue 1, January 2021, Pages 146–163, https://doi.org/10.1093/bib/bbz130

Quaresma, A., & Caels, F. (2020). Elementos verbais e visuais em manuais de História e de Ciências Naturais. In D. Alves, H. G. Pinto, I. S. Dias, M.ª O. Abreu & R. Gillain (Orgs.), Atas da IX Conferência Internacional Investigação, Práticas e Contextos em Educação, p. 296. ESECS-Politécnico de Leiria.

Santos, D., Bick, E., Wlodek, M. (2020). Avaliando entidades mencionadas na coleção ELTeC-por = Assessing named entities in the ELTeC-por collection. Linguamática, Vol. 12, Nr. 2, 2020, pp. 29–49.

Silva, P. N. & Santos, J. V. (2018). Do saber ao poder. Estruturas retóricas e planos de texto em teses de doutoramento. In Z. Aquino, P. R., Gonçalves-Segundo & M. A. Pinto (Orgs.). O poder do discurso e o discurso do poder. Vol. 2, pp. 178-196. Editora Paulistana. ISBN 978-85-5336001-7.

Yang, J., Zhang, Y., Li, L. (2018). YEDDA: A Lightweight Collaborative Text Span Annotation Tool. arXiv:1711.03759v3 [cs.CL] 25 May 2018