Development of a digital corpus of Portuguese didactical texts for language research

Mário Amado Alves
CELGA-ILTEC (Centro de Linguística Geral e Aplicada - Instituto de Linguística Teórica e Computacional)
Faculdade de Letras da Universidade de Coimbra
Portugal
maa@fl.uc.pt

Abstract — **The Corpus de Manuais Escolares consists of 2000 texts from Portuguese schoolbooks, annotated for genre pedagogy research. The extant form of the corpus is a tree of directories and files maintained entirely by hand. We are currently revising this form into one more suitable for computational processing, publication, and evolution. This work has included computational processing itself, with programs written in an incremental way to support the necessary reverse engineering and the envisioned re-engineering. Here we describe the overall process, the programs developed, and some new corpus data already found.**

Keywords - Portuguese language corpus; data discovery; incremental programming; test-driven development

I. INTRODUCTION

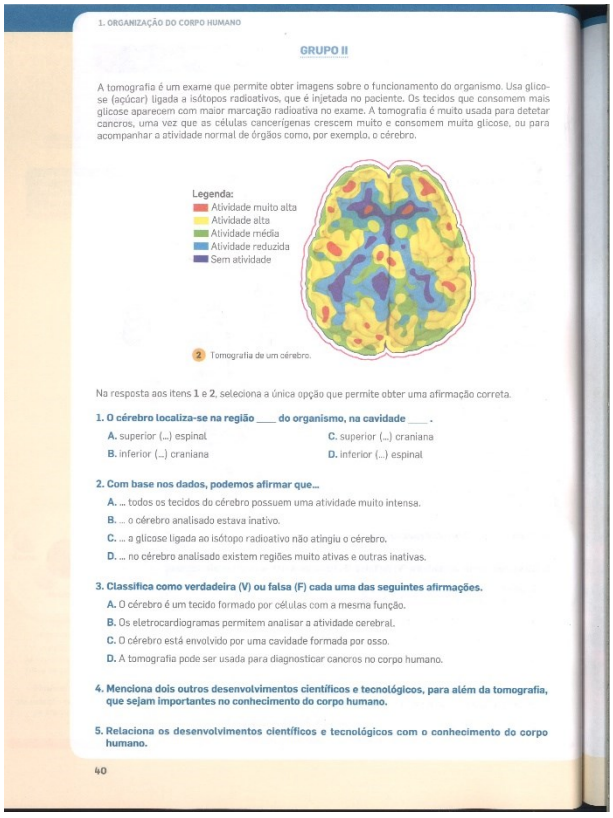
The Corpus de Manuais Escolares consists of approximately 2000 didactical texts in Portuguese, scanned, transcribed, and annotated with genre categories. See Fig. 1 and Fig. 2 for an example text, Table I for the list of categories, section II for details. The corpus is the result of research work carried out since 2017, cf. [1], [2], references thereof, and the Acknowledgements section.

Upon the desire to explore and evolve the corpus with computational processing, in 2021 an effort has been initiated in that direction, counting on the newly joined computer science forces in the person of the first author. The present article describes that work, which has turned out to be one of reverse engineering and transformation from the extant form of the corpus, not suitable for computational processing, to an envisioned form suitable for such processing. In the process, which is incremental, already new data on the corpus have been found which are also reported. We expect such results to be of possible interest for linguists; and the description of the transformation process, of possible interest for engineers involved in similar projects.

The motivation for developing Form Two includes the desire to conduct computational corpus-based studies, like finding quantitatively defined correlations, or simply derive statistics. Another motivation is to facilitate the use of the corpus by other researchers or even the public. Yet another motivation is to allow the evolution of the corpus in a consistent way data-wise.

II. THE CORPUS DE MANUAIS ESCOLARES

The Corpus de Manuais Escolares consists of approximately 2000 texts selected from 64 Portuguese schoolbooks of primary and secondary education (years 1 to 12). All selected texts have been scanned into JPEG files (Fig. 1), and classed with their pedagogical genres (Table I). Of the 2000 texts, approximately 500 have been transcribed, and annotated more delicately, in Word files (Fig. 2).



1. ORGANIZAÇÃO DO CORPO HUMANO

GRUPO II

A tomografia é um exame que permite obter imagens sobre o funcionamento do organismo. Usa glico-se (açúcar) ligada a isótopos radioativos, que é injetada no paciente. Os tecidos que consomem mais glicose aparecem com maior marcação radioativa no exame. A tomografia é muito usada para detetar cancros, uma vez que as células cancerígenas crescem muito e consomem muita glicose, ou para acompanhar a atividade normal de órgãos como, por exemplo, o cérebro.

Legenda:
Atividade muito alta
Atividade alta
Atividade média
Atividade reduzida
Sem atividade

2 Tomografia de um cérebro.

Na resposta aos itens 1 e 2, seleciona a única opção que permite obter uma afirmação correta.

1. O cérebro localiza-se na região ___ do organismo, na cavidade ___.
A. superior (...) espinal
B. inferior (...) craniana
C. superior (...) craniana
D. inferior (...) espinal

2. Com base nos dados, podemos afirmar que...
A. ... todos os tecidos do cérebro possuem uma atividade muito intensa.
B. ... o cérebro analisado estava inativo.
C. ... a glicose ligada ao isótopo radioativo não atingiu o cérebro.
D. ... no cérebro analisado existem regiões muito ativas e outras inativas.

3. Classifica como verdadeira (V) ou falsa (F) cada uma das seguintes afirmações.
A. O cérebro é um tecido formado por células com a mesma função.
B. Os eletrocardiogramas permitem analisar a atividade cerebral.
C. O cérebro está envolvido por uma cavidade formada por osso.
D. A tomografia pode ser usada para diagnosticar cancros no corpo humano.

4. Menciona dois outros desenvolvimentos científicos e tecnológicos, para além da tomografia, que sejam importantes no conhecimento do corpo humano.

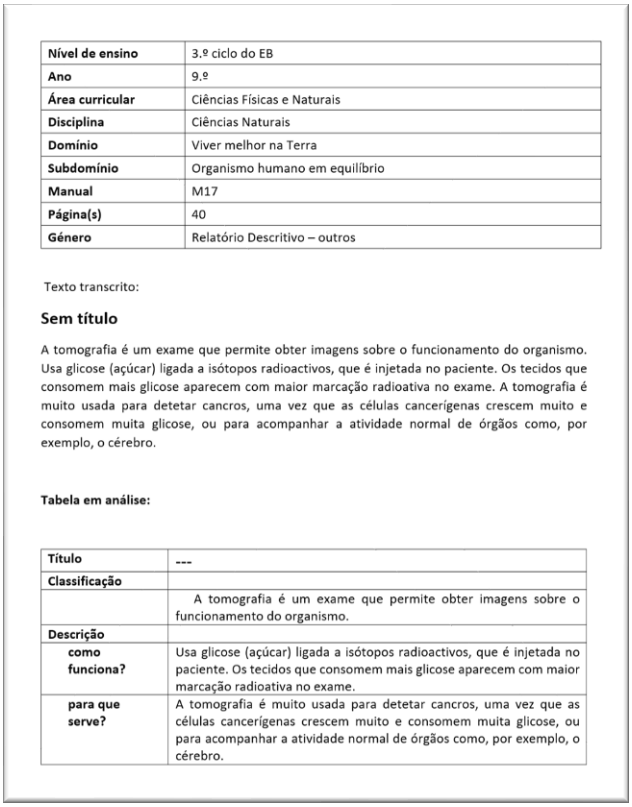
5. Relaciona os desenvolvimentos científicos e tecnológicos com o conhecimento do corpo humano.

40

Figure 1. Example text (M17_40-40) scanned

TABLE I. GÉNEROS DISCURSIVOS

ID	Forma_2
Exp_Cons	Explicação Consequencial
Exp_Fact	Explicação Fatorial
Exp_Hist	Explicação Histórica
Exp_Seq	Explicação Sequencial
Rel_Biog	Relato Biográfico
Rel_Clas	Relatório Classificativo
Rel_Comp	Relatório Composicional
Rel_Desc	Relatório Descritivo
Rel_Func	Relatório Funcional
Rel_Hist	Relato Histórico
Rel_Proc	Relato de Procedimento
RH_EH	Explicação Histórica - Relato Histórico
Tx_Misto	Texto Misto
W_Clas	W CLASSIFICAR
W_Decomp	W DECOMPOR
W_Desc	W DESCREVER
W_Exp	W EXPLICAÇão de procedimentos experimentais… / EXPLICAR
W_Exp_W	W EXPOSITIVO
W_Inst	W INSTRUÇÃO
W_Prot	W PROTOCOLAR / PROTOCOLO
W_Rel	W EXPOSITIVO - relato / RELATAR
W_W_Nul	(sem género atribuído)



Nível de ensino	3.º ciclo do EB
Ano	9.º
Área curricular	Ciências Físicas e Naturais
Disciplina	Ciências Naturais
Domínio	Viver melhor na Terra
Subdomínio	Organismo humano em equilíbrio
Manual	M17
Página(s)	40
Género	Relatório Descritivo – outros

Texto transcrito:

Sem título

A tomografia é um exame que permite obter imagens sobre o funcionamento do organismo. Usa glicose (açúcar) ligada a isótopos radioactivos, que é injetada no paciente. Os tecidos que consomem mais glicose aparecem com maior marcação radioativa no exame. A tomografia é muito usada para detetar cancros, uma vez que as células cancerígenas crescem muito e consomem muita glicose, ou para acompanhar a atividade normal de órgãos como, por exemplo, o cérebro.

Tabela em análise:

Título	---
Classificação	
	A tomografia é um exame que permite obter imagens sobre o funcionamento do organismo.
Descrição	
como funciona?	Usa glicose (açúcar) ligada a isótopos radioactivos, que é injetada no paciente. Os tecidos que consomem mais glicose aparecem com maior marcação radioativa no exame.
para que serve?	A tomografia é muito usada para detetar cancros, uma vez que as células cancerígenas crescem muito e consomem muita glicose, ou para acompanhar a atividade normal de órgãos como, por exemplo, o cérebro.

Figure 2. Example text (M17_40-40) transcribed (middle) and analyzed (bottom)

All selections—of the 64 books, of the 2000 texts from the books, and of the 500 from the 2000—feature, by design, an approximately uniform distribution over variables such as: book, education level, curricular area, pedagogical genre.

The conceptual data scheme of the corpus is as follows. The main concepts, or classes, are: book, text, page, scan, transcription, analysis, genre, curricular area.

A text belongs to a book. A text belongs to one curricular area. Certain books cover multiple curricular areas. The curricular areas are organized in a complicated taxonomy of themselves.

A text is in a scan. A scan is on one of more image files, depending on the number of pages. A text is situated either exactly or partially on the (possibly unary) sequence of pages, that is, a text may be (and often is) situated in a proper subset of the scan.

A text may have a transcription. A transcription may have one or more analyses. An analysis refers to one or more genres. Genres are organized in a complicated taxonomy of themselves. Superclasses of genres are called families. Subcategories and subsubcategories are called levels of delicacy in the vernacular. An analysis may have up to two delicacy levels. The analysis in Fig. 2 has two levels.

III. FORM ONE. REPOSITORY

Form One, extant, consists of all JPEG and Word files in a directory tree, with class information encoded, by hand, in the names of the files and directories, plus transcriptions and annotations in Word files.

In the construction of Form One, computer science support was practically null. Only the most common computer tools, notably the Windows file manager and Microsoft Word, were used directly by the linguists. Identification and classification of texts was done in the name of files, without any automated control or check. The data schema was only partially documented, and not enforced at all. This has resulted in errors in the codification (uncovered in the construction of Form Two). Together with a fragile schema, this has made it impossible to make automated selections or statistics by field value, or any computational processing in general.

Fig. 3 portrays the file tree of Form One (folders only), exposing its internal heterogeneity: most nodes are polymorphic (in the object-oriented sense).

The information coded in the names of the folder nodes are either transient, or redundant with the information coded in the names of the terminal file nodes. Or so it seems. In any case, this is a useful assumption to start the reverse engineering process, as it allows to use only the names of the files to extract the information. The documentation of Form One supports this assumption, by prescribing a scheme for the file name, and only for the file name, as reproduced in Fig. 4.

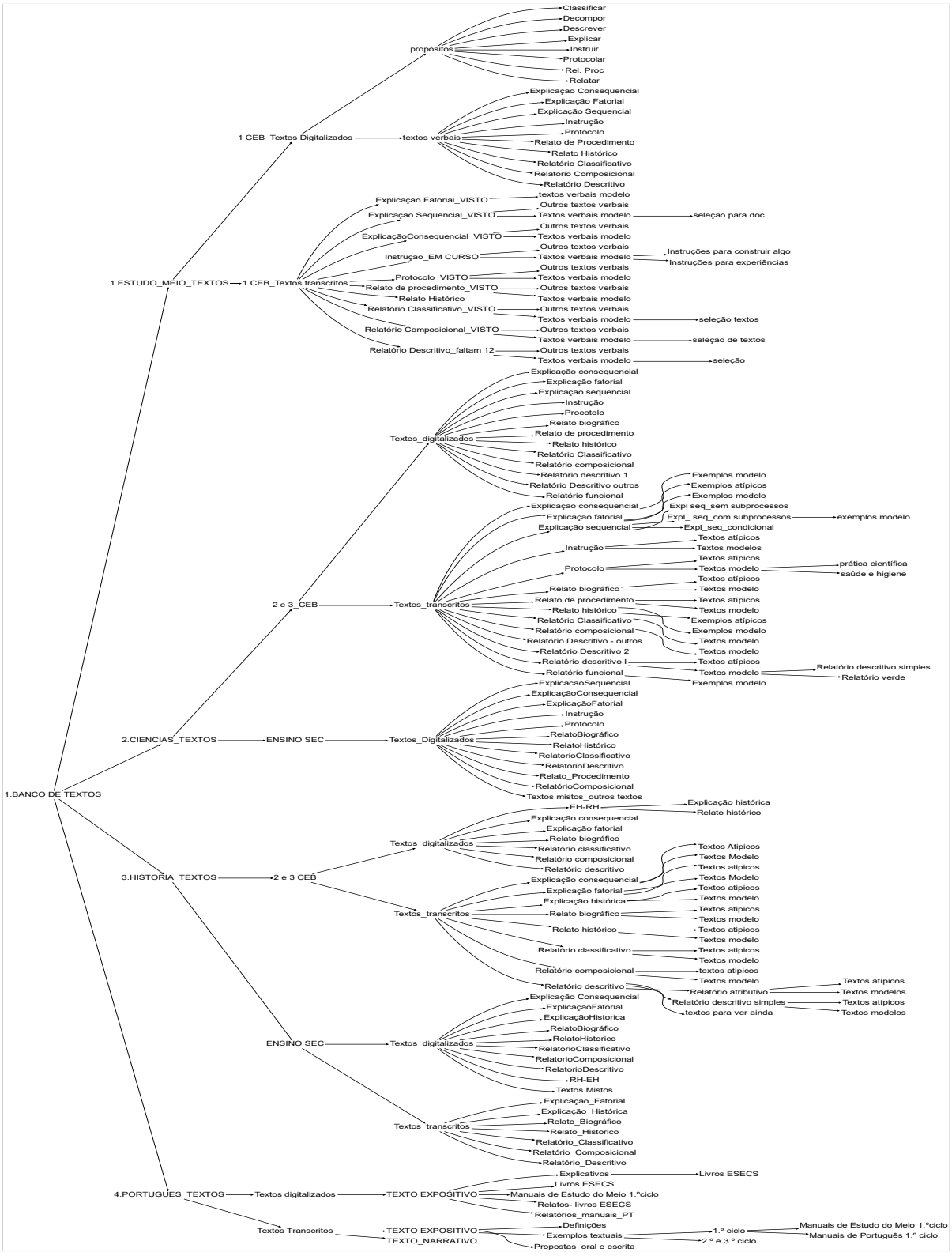


Figure 3. Form One directory tree

CICLO DE ENSINO – ANO de escolaridade – ÁREA – DISCIPLINA - MANUAIS (M) - PAGINAS (P) – GÉNERO

Exemplo:

EB23-05-CMFN-CN-M09-P21-PROTOCOLO

Figure 4. Definition of file name format in Form One documentation

IV. Form Two. Database

Form Two, developing, contains all information of Form One, plus some pieces of knowledge mostly from sources contained or referenced in the encircling project *Portal dos Géneros Académicos* [3]. Form Two is the result of transposing Form One into a database in normal form.

Form Two consists of a set of master tables in CSV format (Comma-Separated Values) that index the texts: the catalog. Table I is an example of such a table. The texts themselves are stored in plain text files identified in the catalog. The annotations are stored in the catalog, or in extra dedicated CSV files (with a meta-schema called *annotation by identification* in preparation by the first author).

Development of Form Two has started mid 2021 and is ongoing at the time of writing. A complete and validated public version of Form Two is expected late 2022.

The construction of Form Two has required the reverse engineering of Form One.

V. Discovering Form One

We have done the reverse engineering, or discovery, of Form One, in an incremental and test-driven way as follows.

We have written programs that process the file tree in order to extract all information possible, including identifying and fixing errors in the file names.

The programs are written incrementally, in a test-driven development cycle. Each program ingests the file tree data, analyses them, and generates a report. The report is examined for errors and inspiration for program increments. The cycle repeats.

Program errors, or bugs, are fixed first. Then data errors. All data errors are fixed by the program. Some error detection is also implemented.

Multiple main programs are created, to deal with different aspects of the data. One program simply enumerates the distinct file extensions in the repository, and the number of files per each, with result:

docx => 584
jpg => 1987

Other programs generate the catalog, i.e. reports that are the tables of the catalog. All programs are unified conceptually, and also formally, as refactoring takes place upon the whole. Also, common modules are created by design. For example, all programs that access class information extracted from the file name share the same pattern matching entities.

VI. Extracting class information

Class information about each text is encoded in the file name, in the directory path, and in the transcription if it exists. The first form, file name, is preferred for extraction, because it is the only one documented (Fig. 4), and also because the others are redundant (section II) and multiform (Fig. 3).

We started by programming the documented format (Fig. 4), and generating a report of match successes and failures over all files. The report exposes deviant occurrences. We add the deviant patterns to the program. The cycle repeats until all occurrences are matched correctly.

Table II shows the final set of patterns as formulated in the source code of the program in Spitbol [5] and Ada [6]. The expressions ... ** Val(*x*) denote assignment, of the matched material ..., to variable Val(*x*); the ampersand (&) is concatenation; *or* means alternation (choice of alternatives).

Note the many alternatives (keyword *or*) of the pattern for pages in Name_Pat. The documented simple format P*n* is in fourth position for the composite pattern to work, as the deviant forms are more complicated.

Also note how a big part of the file name, here called simply the *prefix*, is defined separately, for complexity management. Prefix errors turned out to be few, and so are dealt with by individual fixing, with the function in Table III, before matching.

TABLE II. File name patterns.

```
Name_Pat: Pattern :=
  Arb ** Val(Prefix) &
  "-" & ("M" & Number & Opt(LC_Letter)) **
        Val(Manual) &
  "-" &
  (
     "P" & Number ** Val(Pagina) & "e" & Number **
        Val(Pagina_2)
  or
     "P" & (Number & LC_Letter) ** Val(Pagina_3)
  or
     "P" & Number ** Val(Pagina) & "-" & Number **
        Val(Pagina_2)
  or
     "P" & Number ** Val(Pagina)
  or
     (("destacavel" or "destacável") &
        Opt(LC_Letter)) * Val(Pagina_4)
  or
     "P" &
     ("anexo" & Opt(Number & Opt(LC_Letter))) *
        Val(Pagina_5)
  )
  & "-" & Rest * Val(Genero);

Prefix_Pat: Pattern :=
  Break("-") * Val(Nivel) &
  "-" & Number * Val(Ano) &
  "-" & Break("-") * Val(Area) &
  "-" & Rest * Val(Disciplina);
```

TABLE III. CORRECTING KNOWN INDIVIDUAL ERRORS

```
function Correct_Prefix (X: String) return String is
  (if    X = "EB1-01-EM"   then "EB1-1-W_EM-EM"
   elsif X = "EB1-02-EM"   then "EB1-2-W_EM-EM"
   elsif X = "EB1-03-EM"   then "EB1-3-W_EM-EM"
   elsif X = "EB1-04-EM"   then "EB1-4-W_EM-EM"
   elsif X = "SEC-10-BIO"  then "SEC-10-CMFN-BIO"
   elsif X = "SEC-10-HIST" then "SEC-10-CS-HIST"
   elsif X = "SEC-11-BIO"  then "SEC-11-CMFN-BIO"
   elsif X = "SEC-11-HIST" then "SEC-11-CS-HIST"
   elsif X = "SEC-12-BIO"  then "SEC-10-CMFM-BIO"
   elsif X = "SEC-12-HIST" then "SEC-12-CS-HIST"
   elsif X = "EB1-02EM"    then "EB1-2-W_EM-EM"
   elsif X = "EB23-06 -CMFN-CN" then "EB23-06-CMFN-CN"
   else X);
```

VII. FIXING GENRE INFORMATION

One discovery program generates the list of all distinct genres as encoded in file names (Table V). We find that there are 66 distinct forms, in high contrast with the 15 documented (Table IV).

Attentive inspection reveals multiple causes for this discrepancy, including typos and non-documented genres.

The fixing mechanism is to manually associate each item with a unique ID (Table V). Most typos are resolved, and the other cases, while still there is doubt, are signalled with a W identifier, for *working*, or *waiting for an expert* (or *wary, weird, what?*). The thus annotated table is then used by the program to perform automatic correction. This trick allows continuation of the discovery without being blocked for lack of complete information on this data aspect.

This mechanism also generates Table I, of all the genres effectively at work in Form Two.

TABLE IV. DOCUMENTED GENRES (ADAPTED)

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*

TABLE V. GENRES AS ENCODED IN FILE NAMES (FORMA_1)

ID	*Forma_1*
Rel_Desc	*RELATÓRIODESCRITIVO---*
W_Clas	*CLASSIFICAR*
W_Decomp	*DECOMPOR*
W_Desc	*DESCREVER*
W_Exp	*EXPLICAÇão de procedimentos experimentais...*
Exp_Hist	*EXPLICAÇÃO HISTORICA*
RH_EH	*EXPLICAÇÃO HISTORICA-RELATO HISTÓRICO*
Exp_Cons	*EXPLICAÇÃOCONEQUENCIAL*
Exp_Cons	*EXPLICACAOCONSEQUENCIAL*
Exp_Cons	*EXPLICAÇÃOCONSEQUENCIAL*
Exp_Fact	*EXPLICACAOFATORIAL*
Exp_Fact	*EXPLICAÇÃOFATORIAL*
Exp_Hist	*EXPLICACAOHISTORICA*
Exp_Hist	*EXPLICAÇÃOHISTÓRICA*
Exp_Hist	*EXPLICAÇÃOHISTÓRICA (2)*
Exp_Seq	*EXPLICAÇÃOSEQUENCIAL*
Exp_Seq	*EXPLICAÇÃOSEQUENCIAL (1)*
W_Exp	*EXPLICAR*
W_Exp_W	*EXPOSITIVO*
Exp_Cons	*EXPOSITIVO - explicação consequencial*
Exp_Fact	*EXPOSITIVO - explicação fatorial*
Exp_Hist	*EXPOSITIVO - explicação histórica*
Exp_Seq	*EXPOSITIVO - explicação sequencial*
W_Rel	*EXPOSITIVO - relato*
Rel_Hist	*EXPOSITIVO - relato histórico*
Rel_Clas	*EXPOSITIVO - relatório classificativo*
Rel_Desc	*EXPOSITIVO - relatório descritivo*
Exp_Seq	*EXPOSITIVO_Explicação sequencial*
Exp_Seq	*EXPOSITIVO_Explicação sequencial*
Rel_Hist	*EXPOSITIVO_Relato histórico*
Rel_Clas	*EXPOSITIVO_Relatório classificativo*
Rel_Comp	*EXPOSITIVO_Relatório composicional*
Rel_Desc	*EXPOSITIVO_Relatório descritivo*
W_Inst	*INSTRUCAO*
W_Inst	*INSTRUÇÃO*
W_Inst	*INSTRUÇÃO_continuar aqui*
W_Inst	*INSTRUIR*
W_Prot	*PROTOCOLAR*
W_Prot	*PROTOCOLO*
W_Prot	*PROTOCOLO (2)*
Rel_Proc	*REL.PROC*
Rel_Proc	*REL.PROC.*
W_Rel	*RELATAR*
Rel_Hist	*RELATO HISTÓRICO*
Rel_Proc	*RELATO_PROCEDIMENTO*
Rel_Biog	*RELATOBIOGRAFICO*
Rel_Biog	*RELATOBIOGRÁFICO*
Rel_Proc	*RELATODEPROCEDIMENTO*
Rel_Desc	*RELATODESCRITIVO*
Rel_Hist	*RELATOHISTORICO*
Rel_Hist	*RELATOHISTÓRICO*
Rel_Clas	*RELATORIOCLASSIFICATIVO*
Rel_Clas	*RELATÓRIOCLASSIFICATIVO*
Rel_Comp	*RELATORIOCOMPOSICIONAL*
Rel_Comp	*RELATÓRIOCOMPOSICIONAL*
Rel_Comp	*RELATORIOCOMPOSICONAL*
Rel_Desc	*RELATORIODESCRITIVO*
Rel_Desc	*RELATÓRIODESCRITIVO*
Rel_Desc	*RELATÓRIODESCRITIVO---*
Rel_Desc	*RELATÓRIODESCRITIVO1*
Rel_Desc	*RELATÓRIODESCRITIVO-OUTROS*
Rel_Func	*RELATÓRIOFUNCIONAL*
RH_EH	*RH-EH*
Tx_Misto	*TEXTO MISTO*
Tx_Misto	*TEXTOMISTO*
W_W_Nul	

VIII. Some Data Results

With the above apparatus in place, albeit still developing, already a host of statistical information about the corpus can be retrieved. For example, counts by any field, and their distribution over the number of files, or any other field. The fields being, to wit: *Path, Name, Base, Ext, Prefix, Fixed, Nivel, Ano, Area, Disciplina, Manual, Pagina, Pagina_2, Pagina_3, Pagina_4, Pagina_5, Genero, Genero_ID, Genero_F2, Transcrito, Topico, ID_Texto*.

A choice of results follow (Fig. 5, Fig. 6).

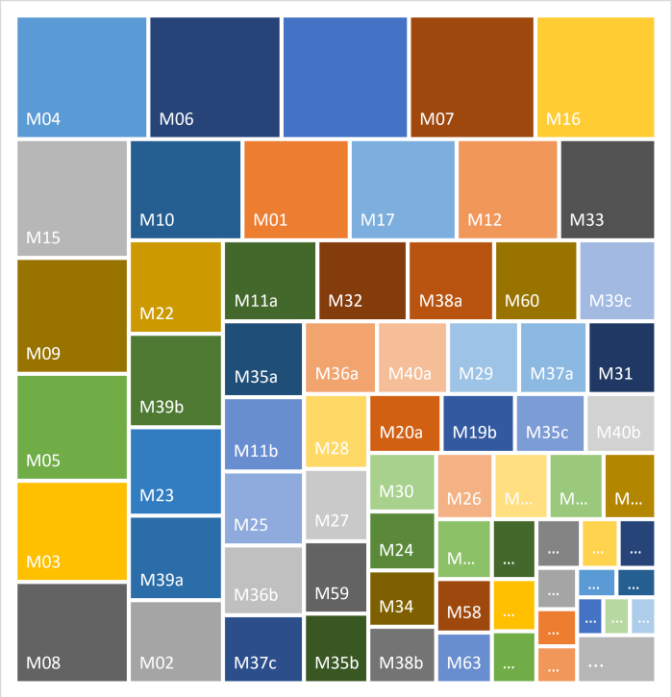


Figure 5. Distribution of files per book

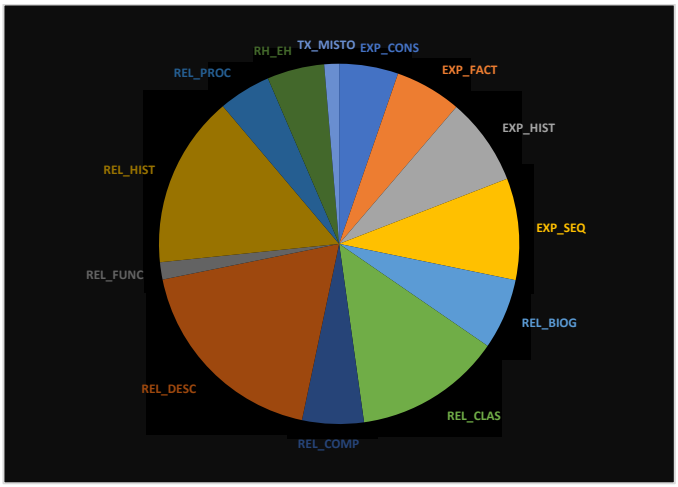


Figure 6. Distribution of valid genres (not W)

IX. Schema Evolution

The data schema of Form One is not normalized (in the database design sense). Namely, it violates normal form (cf. [4]), by not acknowledging the functional dependencies that exist among certain fields, for example:

```
ano :> nível

disciplina, ano(nível?) :> área

manual :> tudo
```

(Again, using Gio Wiederhold's terminology and even notation [4].)

By ignoring the dependencies, the fields have been treated as equals, or as equally dependent on... what? And put side by side in constructs like the file name. This accounts for a number of errors in the data and other otherwise unnecessary complications.

The schema must evolve into a fully normalized one in Form Two. This aspect of the evolution is still under work, notably in an effort to distill the complicated taxonomies of curricular areas and genre categories.

X. Conclusion

The programmatic approach to the discovery of Form One is crucial to obtain reliable data, and to enable the evolution of the corpus towards computational processing and publication. With a size in the thousands of items, a purely manual maintenance of the data is bound to fail in keeping its consistence, as shown in the issues revealed by discovering Form One.

We note that the programs developed for discovery can also support the continual verification, correction, and exploration of Form One, in the corpus-based research for which the Corpus de Manuais Escolares has been created.

But the vision of Form Two—the Database—is surely to gain predominance, as we have now reached the point of the necessary technical confidence, backed by the programs already developed and tested.

Acknowledgments

This research was done in University of Coimbra R&D unit CELGA-ILTEC, FCT UID: 4887.

The monumental work that is Form One of the Corpus de Manuais Escolares is the result of research carried out since 2017 by Ângela Quaresma, Carlos A. M. Gouveia, Fausto Caels, Joana Vieira Santos, Helga Arnauth, Luís Filipe Barbeiro, Marta Filipe Alexandre, Paulo Nunes da Silva.

I am indebted to Adacore (adacore.com) and their GNAT Academic Program for the excellent Ada compiler, programming studio, and support.

REFERENCES

[1] Mário Amado Alves and Marta F. Alexandre and Fausto Caels, "A avaliatividade nos manuais de história: análise exploratória", 2.º Encontro Nacional sobre Discurso Académico, CELGA-ILTEC/UC e CLUP, evento online, 9 e 10 de setembro.

[2] Marta F. Alexandre and Fausto Caels, "Los géneros escolares en Portugal: resultados de un proyecto de investigación sobre libros de texto de portugués, ciencias e historia", Jornada de Enseñanza de la Lengua – Programa de Formación Docente en Lengua y Literatura (PRODELL), Instituto del Desarrollo Humano de la Universidad Nacional de General Sarmiento (UNGS), Provincia de Buenos Aires, Argentina, 2021, novembro 20.

[3] DPDA, Portal dos Géneros Escolares & Académicos, https://sites.ipleiria.pt/pge/, consulted 2022.

[4] Gio Wiederhold, *Database Design,* Mcgraw Hill Computer Science Series, 1983.

[5] AdaCore, *GNAT Reference Manual, GNAT, The GNU Ada Development Environment,* GNAT Pro Edition, Version 23.0w, Date: Apr 08, 2022, AdaCore, https://www.adacore.com/documentation

[6] Ada Conformity Assessment Authority, *Ada Reference Manual, 2012 Edition with 2016 corrections, Language and Standard Libraries,* http://www.ada-auth.org/