User Manual

APU Writing and Reading Corpus 1979–1988

APU

Table of Contents

Introduction 2
Description 2
Reference line 2
Project team 2
Acknowledgements 3
Structure 3
Corpus makeup 3
Parameters 5
Formats 7
.PDF 7
.XML 8
.TXT plain text 9
.TXT tagged 10
Metadata 10
Editorial conventions 13
Filename system 13
Spelling, punctuation 14
File extent 20
Editing and proof-reading 20
Ethical considerations 20
Online Interface 21
Access 21
Documentation 22
Functionalities 22
Browse 23
Layout with XML and PDF 24
Layout with XML and POS tagging 26
Layout with XML and semantic tagging 26
Search 27
Search data 29
Search hits 33
Download 34
Users' Corner 35
Multidimensional analysis 36
MD approach 36
MD and APU 37

February 2017
The APU team

Introduction

The *APU Writing and Reading Corpus 1979–1988* is a diachronic corpus of British English schoolchildren's data at Year 6-level (primary school). The materials are based on a sample of the *Language Performance Surveys* carried out from 1979 to 1988 by the Assessment of Performance Unit (APU), UK National Foundation for Educational Research (NFER). More specifically, the APU corpus is made up of two components: "School Scripts" from the Writing Surveys and "Basal Readers" from the Reading Surveys. The methodology to compile the corpus builds on a cross-disciplinary approach to literacy development and corpus linguistics. The corpus has been developed in three stages: data selection, data, online interface development.

This is part of a project entitled *'The art of writing English': A corpus of schoolchildren's writings*, funded by Xunta de Galicia, Proxectos Emerxentes (Grant EM2014/028).

Description

Project:	'The art of writing English': A corpus of schoolchildren's writings
Project coordinators:	Nuria Yáñez-Bouza (University of Vigo, Spain) and Victorina González-Díaz (University of Liverpool, UK)
Corpus short name:	APU corpus
Time of compilation:	2014–2016
Number of samples:	522 school scripts written by children (ca. 93,000 words) 21 basal readers written for children (ca. 79,000 words)
Period:	1979, 1988
School level:	Primary, Year 6, 11-year-old pupils
Language:	school scripts: British English basal readers: British English, US English
Project website:	http://apucorpus.webs.uvigo.es
Corpus interface:	http://apucorpus.liverpool.ac.uk
Contact email:	apucorp@liverpool.ac.uk

Reference line

APU Writing and Reading Corpus 1979–1988. Compiled by Nuria Yáñez-Bouza (University of Vigo, Spain) and Victorina González-Díaz (University of Liverpool, UK).

Project team

Coordinators	Nuria Yáñez-Bouza (University of Vigo) Victorina González-Díaz (University of Liverpool)
Other project members	Yolanda Fernández-Pena, Dolores González-Álvarez (University of Vigo)
Research Assistants	Sofía Bemposta-Rivas, Carla Bouzada-Jabois, Evelyn Gandón-Chapela, Carla Seabra-Dacosta (University of Vigo); Roanne M. Ephithite (University of Liverpool/The Reader Organisation)

Collaborators	Annabel Charles, Educational Consultant, assessment of attainment bands Victoria Smith (Weatherhead High School), Karen Rogan (Liverpool Hope University)
IT technical support	Computing Services Department, University of Liverpool; Denys Bondarenko, Prof. David Denison.

Acknowledgements

This project has been generously funded by Xunta de Galicia, Consellería de Cultura, Educación e Ordenación Universitaria. Proxectos Emerxentes, Grant EM2014/028 (2014–2016).

Our gratitude extends to the National Foundation for Educational Research (NFER) and, in particular, its former Deputy Director, Dr Chris Whetton, for agreeing to the use of the APU materials for teaching research purposes. We are also indebted to Dr Greg Brooks, Prof. Bas Aarts, Prof. Dick Hudson, Prof. David Denison, and Dr Anne Qualter for their suggestions, help and advice at different stages of the corpus compilation.

The APU materials have been safeguarded at the University of Liverpool (UK) since 1991 and supervised Dr Victorina González-Díaz since 2007.

Structure

Corpus makeup

The corpus materials are based on the Assessment of Performance Unit (APU) surveys of language performance, carried out by the National Foundation for Educational Research (NFER). The APU **writing surveys** aimed at assessing pupils' performance in different communicative situations, such as editing, describing, reporting, etc. (Gorman et al. 1991: 29). There are scripts by/for primary schoolchildren (Year 6-level, 11-year-olds) as well as by/for secondary schoolchildren (Year 11-level, 15-year-olds). This current version of the corpus focuses on the former age group (**Year 6, 11-year-olds**) and two types of text with a long-standing tradition in UK schools, namely **narration and argumentation**. The importance of genres and the influence of the task on writing performance have been widely acknowledged in previous studies (e.g. Gorman et al. 1991: 30–5, Reppen 1994: 23–32) to the extent that "an essential knowledge of forms of texts is [considered] a prerequisite to full competence in writing" (Kress 1994: xiv). The rationale behind the selection of the younger age-group and these two tasks lies in the observation that children start to be aware of genre differences in writing already at age 8, "using linguistic features to distinguish between narrative tasks and expository tasks", and that at age 11–12 they are "able to control a number of different types of writing tasks", including "a distinct linguistic style for argumentative/persuasive writing" (see Biber et al. 2002: 460, Reppen 1994: 7). Besides, primary education has recently undergone important changes in the curriculum of English grammar teaching (see the National Curriculum statutory programmes). On the practical side, the narration and argumentation tasks can be compared with other children corpora (e.g. *Oxford Children Corpus*; Reppen 1994).

The *APU Writing and Reading Corpus* consists of two major components, namely writings by children – "**School Scripts**" – and writings for children – "**Basal Readers**"–, in line with work by Biber and associates (Reppen 1994, Biber et al. 2002). The former will help us to identify the range of lexical and grammatical features that are (fully or partially) mastered at Year-6 level; the latter will signal what linguistic features this age-group tends to be exposed to and/or presented with as linguistic models.

School Scripts

The selection criteria for the "School Scripts" component are:

a. Tasks that represent the two communicative functions above-mentioned: *Rule* for the argumentative/persuasive function, and *Story* for the narrative/descriptive function.
b. The first and last year of the APU surveys, in order to facilitate diachronic studies: 1979, 1988.
c. Scripts for which both survey years (1979, 1988) and both tasks (Rule, Story) are available.
d. Scripts which are legible, with some exceptional cases of damage affecting a minimum of the running text.
e. Scripts for which the pupil's sex is known.

The "School Scripts" component thus consists of 522 scripts and 92,728 words distributed as shown in Table 1. We aimed at balance inasmuch as possible, but a perfect match across survey years was not possible due to irreparable damage in some of the scripts and the loss of some original materials.

Table 1. Corpus data: School Scripts

Year	Comm. Function	Pupil's Sex	Files	Word Count	Totals
1979	Argumentative-cum-Persuasive	male	65	6,977	**123 files** 12,677 words
		female	58	5,700	
	Narrative-cum-Descriptive	male	65	14,525	**123 files** 28,067 words
		female	58	13,542	
	Total	*male*	*130*	*21,502*	***246 files*** *40,744 words*
		female	*116*	*19,242*	
1988	Argumentative-cum-Persuasive	male	66	6,700	**138 files** 16,058 words
		female	72	9,358	
	Narrative	male	66	15,188	**138 files** 35,926 words
		female	72	20,738	
	Total	*male*	*132*	*21,888*	***276 files*** *51,984 words*
		female	*144*	*30,096*	
Grand Total	Argumentative-cum-Persuasive		261	28,735	**522 files** 92,728 words
	Narrative-cum-descriptive		261	63,993	

Basal Readers

The selection of the "Basal Readers" component includes 13 readers used in the APU Surveys. These are excerpts taken from published children's books dating from 1979, 1982 and 1988. As these sources added to just ca. 15,500 words, in a second compilation stage we supplemented our Basal Reader collection with larger excerpts of texts from some other books used by the APU (original Reading Survey). The rationale behind the book selection had to do with the textual alterations that the original APU reader compilers introduced in the reading materials; in other words, we only included new data samples from those books whose texts had not been adapted in any way by the APU compilers prior to their inclusion in the Reading Surveys. These materials amount to 8 new text samples and ca. 63,800 words. Altogether, this component includes 21 files and 79,306 words. The distribution is displayed in Table 2 below.

All but two basal readers are British sources. The two printed in US are the NFER basal reader "The Flying Machine" and the supplementary source "The Golden Apples of the Sun".

Table 2. Corpus data: Basal Readers

Year	Files		Word Count	
	NFER	Suppl.	NFER	Suppl.
1979	8	4	9,368	32,608
1981	--	1	--	7,811
1982	2	3	1,975	23,379
1988	3	--	4,165	--
Total	*13*	*8*	*15,508*	*63,798*
Grand Total	**21 files**		**79,306 words**	

Parameters

School Scripts

The following parameters have been coded for each script of the "School Scripts" component in the APU corpus.

Table 3. Parameters: School Scripts

Parameter	Description	Notes
Pupil ID	NNNNN	As documented in the original survey. In order to keep anonymity and confidentiality of the participants, the original surveys were documented by numerical ID.
Pupil's sex	male, female	As documented in the original survey. Scripts for which the pupil's sex had not been document have been discarded.
Pupil's date of birth	YYYY-MM-DD	As documented in the original survey. There are some scripts with unknown value.
Script filename	In the format: WYY1tt_NNNNNx	See "Editorial Conventions" for a full description.

Script title	Pupil ID followed by Task	e.g. 12001, Rule
School level	Primary. Year 6, 11-year-olds	
Survey year	1979 1988	
Skill	Writing	
Task	Rule	*Rule*: "Think of a rule which you have to obey", common to 1979 and 1988.
	Story	*Story*: "Short story based on a past experience" in the 1979 surveys; "Story based on a picture" in the 1988 surveys.
Task function	Argumentative-cum-Persuasive	Rule
	Narrative-cum-Descriptive	Story
Attainment band	High Middle Low	Scripts individually assessed by Annabel Charles, Education Consultant,. Features taken into consideration: sentence structure, punctuation, overall structure, paragraphing, selection of detail, vocabulary, awareness of audience.
Length/Extent	Lines Words	See "Editorial Conventions" for a full description.

Basal Readers

The following parameters have been coded for each file of the "Basal Readers" component in the APU corpus.

Table 4. Parameters: Basal Readers

Parameter	Description	Notes
Filename	In the format BRYY1_zzzz	See "Editorial Conventions" for a full description.
Short title		Indeterminate length, e.g. Whales1
Function	Narrative-cum-Descriptive	
Publication year	1979, 1982, 1988	Unbalanced due to scarcity of materials
Bibliographic reference	Full title, publisher, publication place	
Author's name		
Author's sex	Male, female	
Contents	Chapters, page numbers	
Length/Extent	Lines Words	See "Editorial Conventions" for a full description.

Formats

The APU corpus exists in different versions to suit different users and uses:

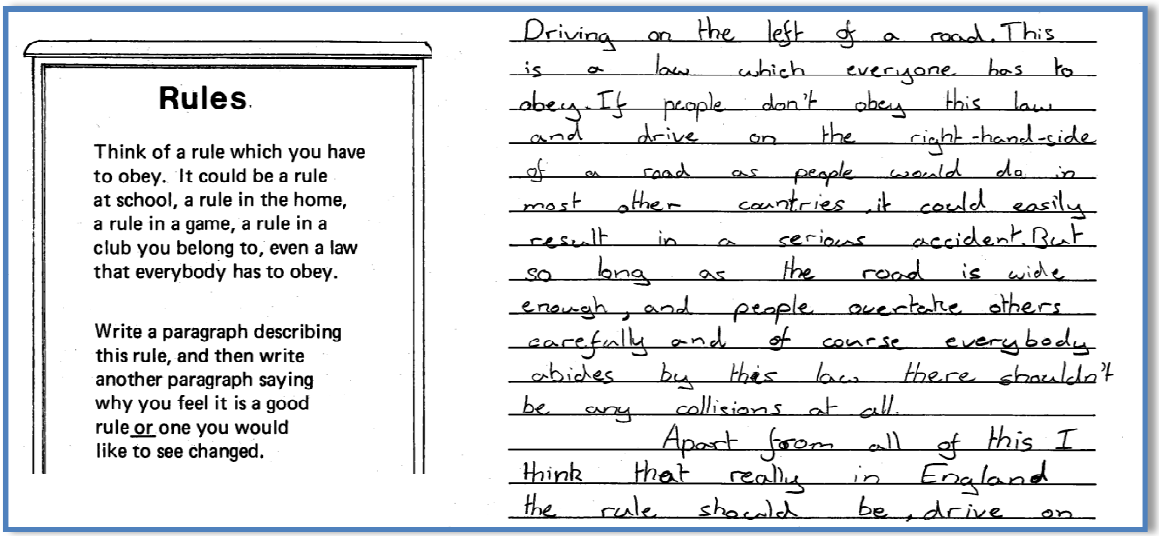
- digitised images in .PDF format
- transliteration in .XML format with TEI-Lite mark-up and metadata, without linguistic tagging
- transliteration in .TXT format, plain text and original spelling
- .TXT format, plain text and normalised spelling
- .TXT format with part-of-speech tagging (CLAWS7)
- .TXT format with semantic tagging (USAS)

.PDF

The corpus materials from the APU surveys exist in paper format. They have been safeguarded at the University of Liverpool (UK) since 1991, and supervised by Dr Victorina González-Díaz since 2007.

Due to their linguistic and cultural value, it was thought appropriate to digitise the original scripts and basal readers for a better preservation and for ease of study. The materials were scanned as images in .PDF format, on a two-side-page landscape layout, as show in Figure 1. The task was carried out during October-November 2014 at the Department of English, School of the Arts, University of Liverpool.

Figure 1. Format: .PDF digitised image (sample: W881ru_23005m)



Rules

Think of a rule which you have to obey. It could be a rule at school, a rule in the home, a rule in a game, a rule in a club you belong to, even a law that everybody has to obey.

Write a paragraph describing this rule, and then write another paragraph saying why you feel it is a good rule or one you would like to see changed.

Driving on the left of a road. This is a law which everyone has to obey. If people don't obey this law and drive on the right-hand-side of a road as people would do in most other countries it could easily result in a serious accident. But so long as the road is wide enough, and people overtake others carefully and of course everybody abides by this law there shouldn't be any collisions at all.

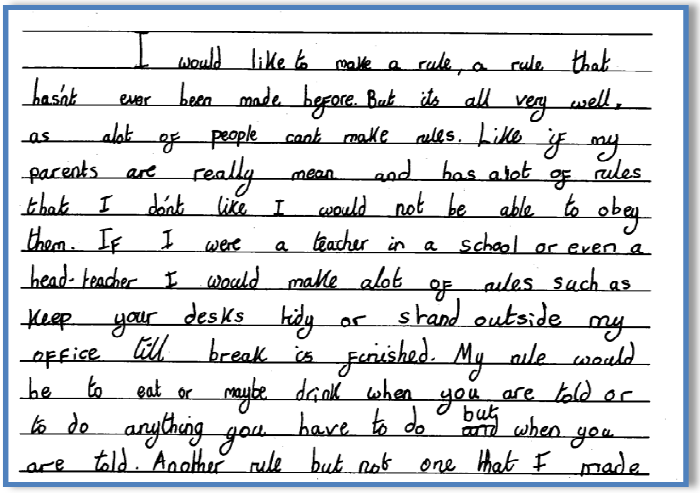
Apart from all of this I think that really in England the rule should be, drive on

.XML

The master-copy of the corpus is XML-compliant with TEI-Lite mark-up and headers (TEI P5). The XML files have been compiled in UTF-8 character set with the XML editor oXygen (2015). The online interface displays the running text of the original script with mouse-over effects to flag words or a string of words around which there is an XML tag. The XML tagset for the "School Scripts" and for the "Basal Readers" components is provided in the online interface. The interface allows users to search files by text or by XML tag.

Figure 2. Format: XML and TEI-Lite version (sample: W881ru_23113f)

Figure 2-a. PDF image



I would like to make a rule, a rule that hasn't ever been made before. But its all very well, as alot of people cant make rules. Like if my parents are really mean and has a lot of rules that I dont like I would not be able to obey them. If I were a teacher in a school or even a head-teacher I would make alot of rules such as keep your desks tidy or stand outside my office till break is finished. My rule would be to eat or maybe drink when you are told or to do anything you have to do ~~and~~ but when you are told. Another rule but not one that I made

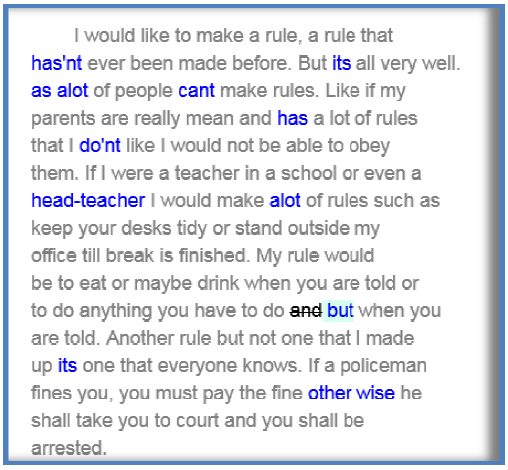
Figure 2-b. XML format in oXygen

```
<text xml:id="W881ru_23113f">

<body>

<pb n="1"/>
<lb/>
<lb/><hi rend="indent">I would like to make a rule, a rule that</hi>
<lb/><sic corr="hasn't" type="apostrophe">has'nt</sic> ever been made before. But <sic corr="it's" type="apostrophe">its</sic> all very well.
<lb/><sic corr="As" type="punctuation period">as</sic> <sic corr="a lot" type="spelling join-split">alot</sic> of people <sic corr="can't"
type="apostrophe">cant</sic> make rules. Like if <note resp="#NYB" comment="morphosyntax concord SV">my
<lb/>parents are really mean and <sic corr="have" type="morphosyntax">has</sic></note> a lot of rules
<lb/>that I <sic corr="don't" type="apostrophe">do'nt</sic> like I would not be able to obey
<lb/>them. <note resp="#NYB" comment="morphosyntax subjunctive">If I were</note> a teacher in a school or even a
<lb/><sic corr="headteacher" type="spelling join">head-teacher</sic> I would make <sic corr="a lot" type="spelling join-split">alot</sic> of rules such as
<lb/>keep your desks tidy or stand outside my
<lb/>office till break is finished. My rule would
<lb/>be to eat or maybe drink when you are told or
<lb/>to do anything you have to do <subst><del rend="crossed out">and</del><add place="above">but</add></subst> when you
<lb/>are told. Another rule but not one that I made
```

Figure 2-c. XML format online



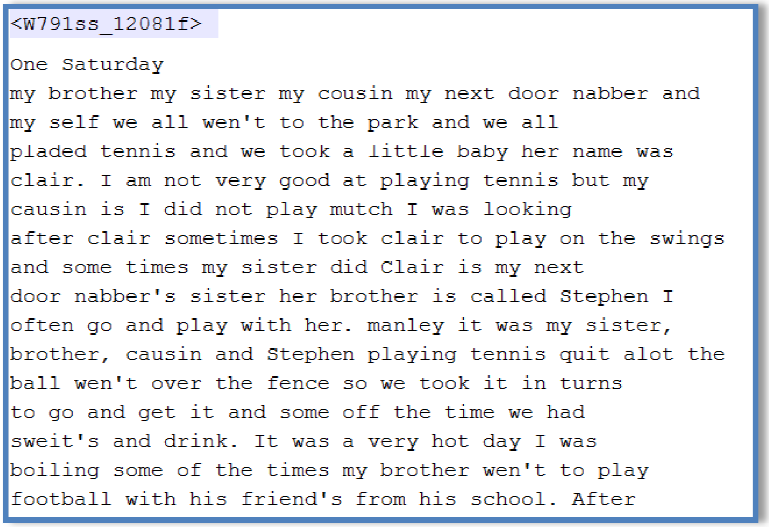
I would like to make a rule, a rule that has'nt ever been made before. But its all very well. as alot of people cant make rules. Like if my parents are really mean and has a lot of rules that I do'nt like I would not be able to obey them. If I were a teacher in a school or even a head-teacher I would make alot of rules such as keep your desks tidy or stand outside my office till break is finished. My rule would be to eat or maybe drink when you are told or to do anything you have to do ~~and~~ but when you are told. Another rule but not one that I made up its one that everyone knows. If a policeman fines you, you must pay the fine other wise he shall take you to court and you shall be arrested.

.TXT plain text

A .TXT version has been produced in plain text (Latin-1) with original spelling, punctuation and lineation. This has been derived from the XML master version, by stripping the text off all tags and textual annotations, except for the Text ID enclosed in caret brackets at the start of each file. Files can be opened with any text editor such as Notepad++.

The 'official' word counts for the corpus are calculated from this untagged, plain text version with original spelling. Individual word counts for each file can be consulted in the "Frequency List" tool available on the APU online interface.

Figure 3. Format: untagged, plain .TXT (sample: W791ss_12081f)



```
<W791ss_12081f>
One Saturday
my brother my sister my cousin my next door nabber and
my self we all wen't to the park and we all
pladed tennis and we took a little baby her name was
clair. I am not very good at playing tennis but my
causin is I did not play mutch I was looking
after clair sometimes I took clair to play on the swings
and some times my sister did Clair is my next
door nabber's sister her brother is called Stephen I
often go and play with her. manley it was my sister,
brother, causin and Stephen playing tennis quit alot the
ball wen't over the fence so we took it in turns
to go and get it and some off the time we had
sweit's and drink. It was a very hot day I was
boiling some of the times my brother wen't to play
football with his friend's from his school. After
```

.TXT tagged

In order to facilitate the linguistic analysis of the materials, each file in the corpus has been tagged morphologically for part-of-speech and semantically. This has been produced with W-Matrix, a software tool for corpus analysis developed at UCREL, University of Lancaster (http://ucrel.lancs.ac.uk/wmatrix/).

The output from W-Matrix is provided in two formats:

- .TXT format with part-of-speech (POS) tagging, CLAWS7
- .TXT format with semantic tagging, USAS

Both the CLAWS7 and the USAS tagset are provided in the APU online interface, and can also be consulted on the W-Matrix website. W-Matrix provides frequency lists for words with POS tags and semantic tags; the lists can be sorted alphabetically or by frequency. These are available in the APU online interface.

The input for W-Matrix comes from the .TXT untagged version. A normalised spelling version was needed as an intermediate stage, given that CLAWS7 and USAS take present-day standard English as reference for automatic tagging. We are indebted to Prof. David Denison (University of Manchester) for producing the systematic routine in order to normalise the original files.

Figure 4. Format: POS tagged text (sample: W791ss_12001m)

Figure 5. Format: USAS tagged text (sample: W791ss_12001m)

Metadata

The corpus metadata have been stored in full detail in a MS Access relational database. A selection of the metadata information has been coded in XML files with TEI-headers attending to the four major TEI elements:

- *file description*: filename, full bibliographic information, source description, sample extent, funding, compilers, publication statement;
- *encoding description*: project compilation and coding description, for instance spelling normalisation, text alignment, hyphenation, etc.;
- *profile description*: domain and language involved;
- *revision history*: version production and revision documentation.

The parameters displayed in the online interface section METADATA DESCRIPTION are listed in the tables below and are displayed in the BROWSE layout, as shown in the screenshots that follow. (See section on "Parameters".)

Table 5. Metadata: School Scripts

Pupil's details	Script reference	Script description
ID number	Domain	Level
Sex	Survey date	Skill
Date of birth	Filename	Function
	Script title	Task
		Attainment band
		Length (lines, words)

Figure 6. Metadata online: XML headers (sample: W791ss_12001m)

```
<TEI>

<teiHeader>


<fileDesc>
    <titleStmt>
        <idno type="filename">W791ss_12001m</idno>
        <idno type="database ID">644</idno>
        <title>12.001, Short story based on a past experience</title>
        <author>
            <name>12.001</name>
            <sex value="1">male</sex>
            <date>1968 March 4</date>
            <p>Creator role: pupil</p>
        </author>
    </titleStmt>
```

Figure 7. Metadata online: School Scripts (sample: W791ru_12010f)

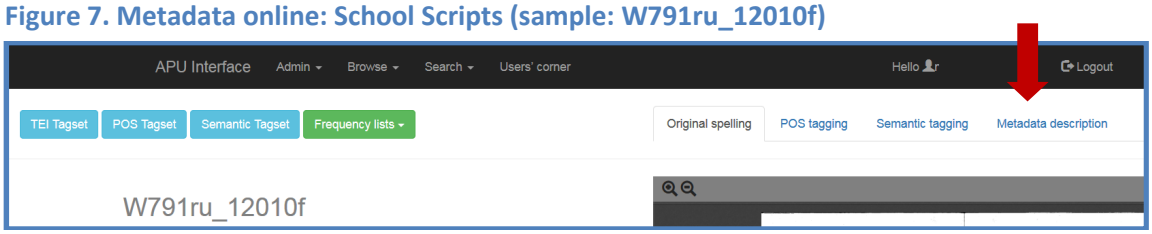


Figure 8. Metadata online: School Scripts (sample: W791ru_12010f)

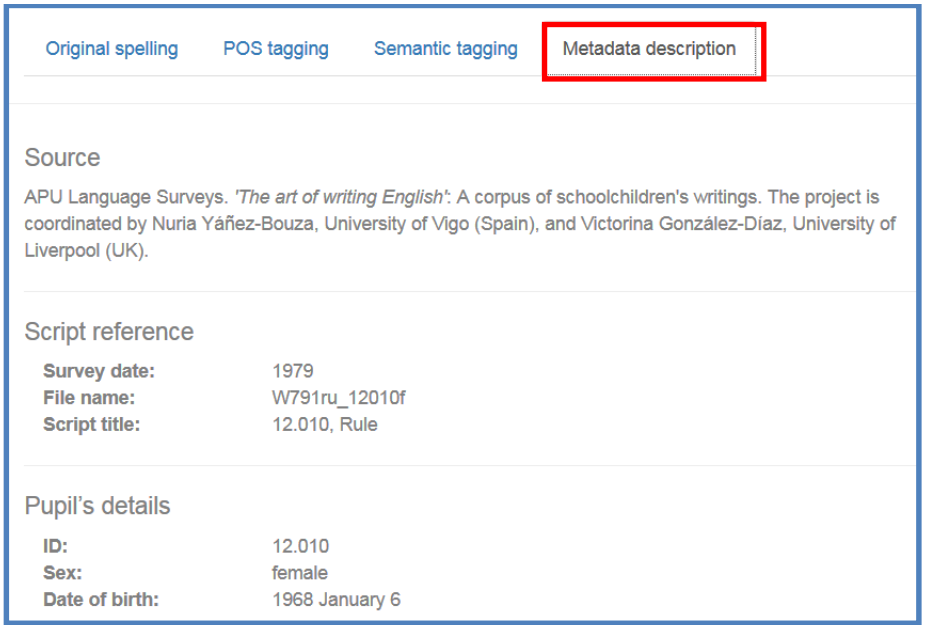


Table 6. Metadata: Basal Readers

Basal Reader reference	Basal Reader description
Domain	Function
Filename	Author
Short title	Author's sex
	Publication year
	Bibliographic reference
	Contents
	Length (lines, words)

Figure 9. Metadata online: Basal Readers (sample: BR79_karthur3)

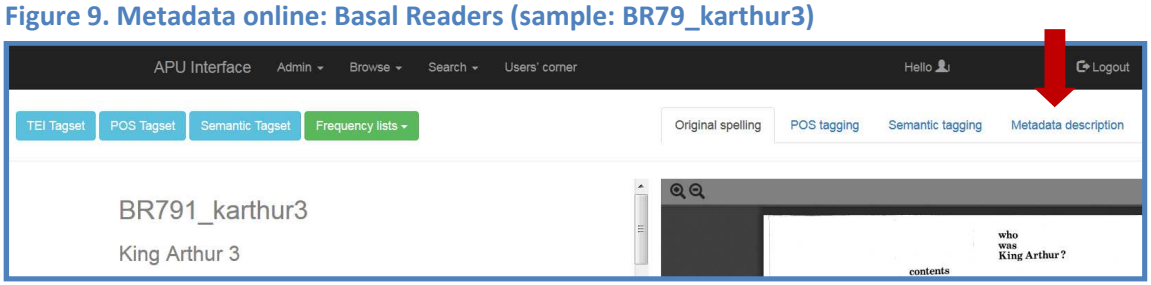
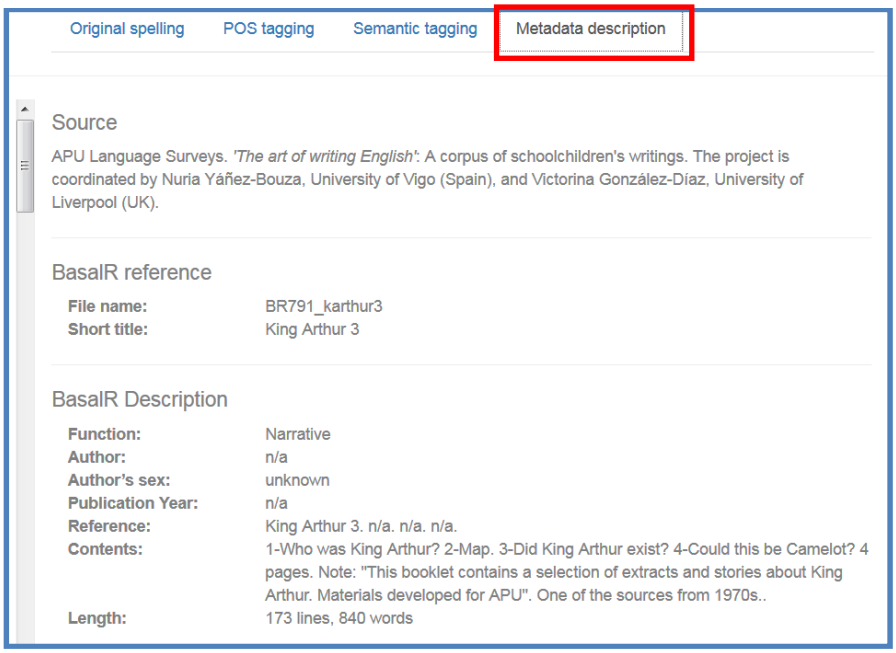


Figure 10. Metadata online: Basal Readers (sample: BR79_karthur3)



Editorial conventions

Filename system

All filenames in the corpus follow a predefined formula for the sake of consistency and ease of identification.

School Scripts

WYY1tt_NNNNNx, where

W	Writing domain
YY	Survey year (79=1979, 88=1988)
1	Primary level
tt	Task abbreviation (ru=Rule, ss=Story Ending, sp=Story based on a picture)
NNNNN	Pupil ID as in the original surveys
x	Pupil's sex (m=male, f=female)

For instance: W791ru_12001m

This is a script from the APU Writing survey, from the 1979 surveys, from Primary school level (11-year-olds), and from the Rule task. It is written by a pupil with the ID number 12001, who is a male student.

Basal Readers

BRYY1_zzzz, where

BR =	Basal Reader, Reading domain
YY =	Survey year (e.g. 79=1979, 82=1982, 88=1988)
1 =	Primary level
zzzz =	Abbreviated title, indeterminate length

For instance, BR881_klion1
This is a basal reader, from the 1988 surveys, from Primary school level (11-year-olds), with the title abbreviation “klion1”.

We recommend that individual citations from the APU corpus include the text identifier (filename), e.g. “W791ru_12001m” for School Scripts, “BR881_klion1” for Basal Readers.

Spelling, punctuation

School Scripts
Transliterations of the school scripts retain the original spelling, including orthographic and grammatical deviations from written standard English, as well as the original punctuation (or lack thereof). These deviations are displayed as such in the online interface, and the appropriate XML tag has been added in order to indicate the normalised form; for instance, the word *answer* is misspelled as *aswer* in W791ss_12001m.

Figure 11. School Scripts: spelling deviations (sample: W791ss_12001m)

Figure 11-a.

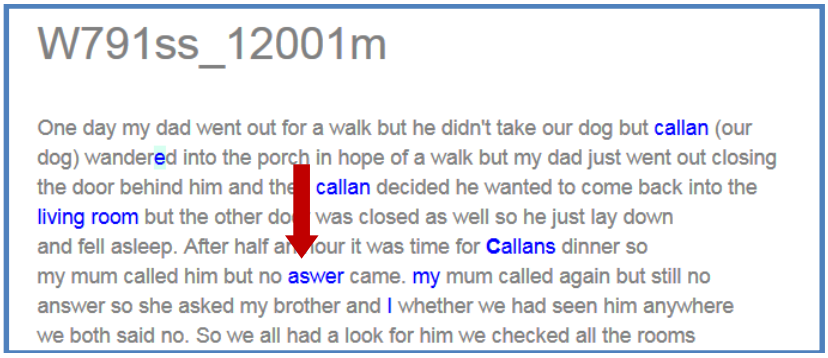


Figure 11-b.

Words that are split across two lines are indicated by the symbol used in the original script, if any, including a hyphen (-) or a double hyphen symbol (=) at the end of the first line, and sometimes repeated at the start of the next. These broken words are displayed split in the XML version. For instance, the word *crashes* is split with a hyphen as *crash-es* in W791ru_12028m, the word *infants* is split without hyphenation but logically as *in/fants* in W791ru_12043m, while the word *well* is oddly split as *we/ll* in W791ru_12067f (see Figure 12).

Figure 12. School Scripts: broken words across two lines (sample: W791ru_12067f)

Figure 12-a. PDF snapshot

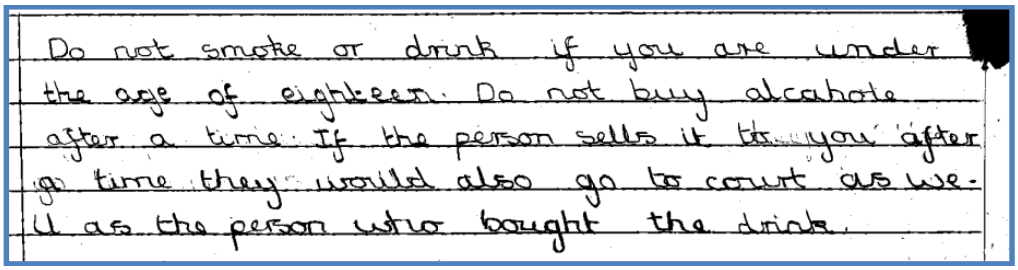


Figure 12-b. Transliteration online

Do not smoke or drink if you are under the age of eighteen. Do not buy alcahole after a time. If the person sells it to you after a time they would also go to court as we-ll as the person who bought the drink.

Spelling variation is sometimes found in cases in which contemporaneous orthographical conventions appear to differ from present-day practice. A case in point is the spelling of compounds. The original spelling in the transliteration has in all cases been retained (see Figure 13: *care-taker* in W881ru_23026f and *caretaker* in W881ru_23165m). For its part, in the XML files we have tagged instances in which practices seem to have varied over time (cases where compounds that are nowadays written as a single lexeme appear in the school scripts hyphenated or as separate words, like *playground* in W881ru_23026f), in order to mark the change in conventions. The decision to tag these instances is based on a scrutiny, on a case-by-case basis, of the spellings provided by three main sources: the *Oxford English Dictionary*, the *British National Corpus* (1960-1975 dataset) and *Googlebooks* (British English, 1960-1990 dataset). To give an example: for the word *playground*, the OED only records the fused spelling of the word (i.e. *playground*). The BNC features both the separate (*play ground*) and the hyphenated (i.e. *play-ground*) form, although the differences in frequency are telling (*play ground* 6.78 instances per million words and *play-ground* 0.16 tokens per million words). *Googlebooks* records the prevalence of the fused form *playground* over any other spelling variants of the word (see Figure 14). On the basis of this evidence, we tagged the separate (*play ground*) and the hyphenated (*play-ground*) variants of *playground* as options deviating from the spelling conventions of written standard English.

Figure 13. School Scripts: compound words and (lack of) hyphenation

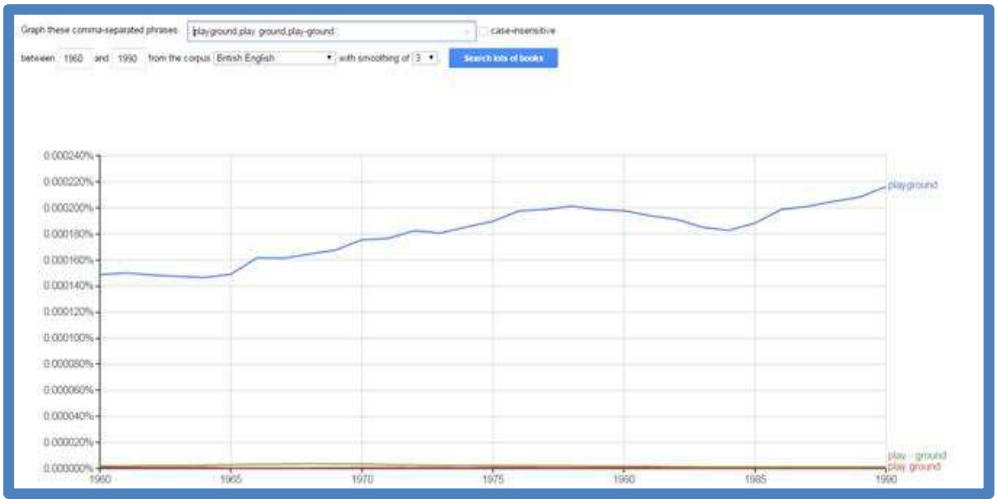
Figure 13-a. W881ru_23026f: *play ground* but *care-taker*

It is quite a good rule because we can get bad teeth plus gum disease. Another reason is that we may litter the play ground with sweet papers and the care-taker might get sick of tidying the play-ground up. Another one can be feeling sick, and the teachers would be running around after us.

Figure 13-b. W881ru_23165m: *playground* and *caretaker*

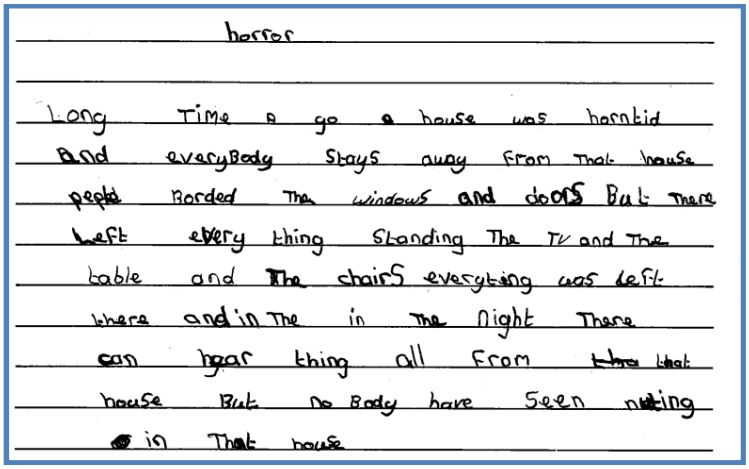
At school children are'nt allowed to throw litter in the playground. It [redacted] is one of the rules because it makes the school look filthy and also the caretaker has to pickup all the litter. It is very unfair to the caretaker and some of the chil-dren too. Sometimes birds and other unfortunate creatures get chewing gum stuck on their throat. During the year visitors come to our school and if they NB

Figure 14. *Googlebooks* dataset snapshot for the spelling variants of the word *playground*



Spelling variants of the type *-ise/-ize* (*realise/realize*) and *-our/-or* (*colour/color*) have been retained as in the original and have not been normalised; for instance, the title of the story in W881sp_23040m (see Figure 15).

Figure 15. School Scripts: spelling variation *-our/-or* (sample: W881sp_23040m)



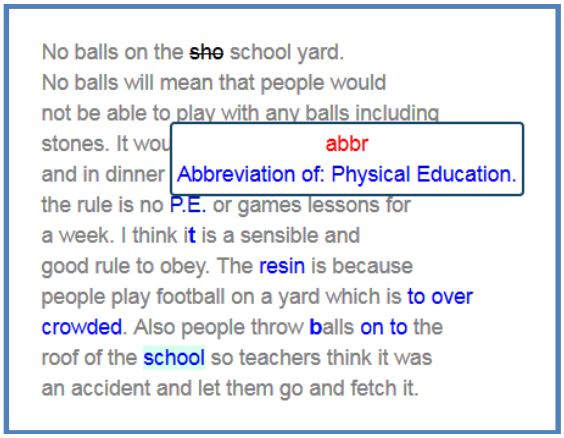
Abbreviations have been retained in their short form as written in the original script. The expanded form has been coded in the XML version; for instance, *P.E.* for the school subject *Physical Education* in W791ru_12142m (see Figure 16).

Figure 16. School Scripts: abbreviations expanded (sample: W791ru_12142m)

Figure 16-a. XML tagging

```
<lb/>No balls on the <del rend="crossed out">sho</del> school yard.
<lb/>No balls will mean that people would
<lb/>not be able to play with any balls including
<lb/>stones. It would be in progress only in break
<lb/>and in dinner time. The punishment for breaking
<lb/>the rule is no <abbr expan="Physical Education">P.E.</abbr> or games lessons for
<lb/>a week. I think i<hi rend="overwritten">t</hi> is a sensible and
<lb/>good rule to obey. The <sic corr="reason" type="spelling">resin</sic> is because
```

Figure 16-b. Display online

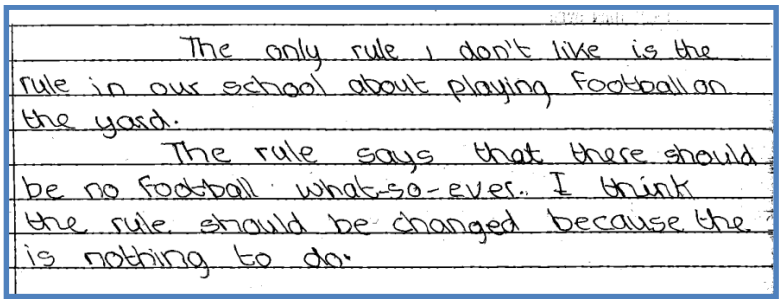


All quotation marks are retained in the text and are represented by appropriate Unicode characters.

Text alignment, lineation and paragraph indentation have been preserved. For instance, pupils usually write the title of their task with centre alignment or with indentation, as in W791ru_12132m (see Figure 17).

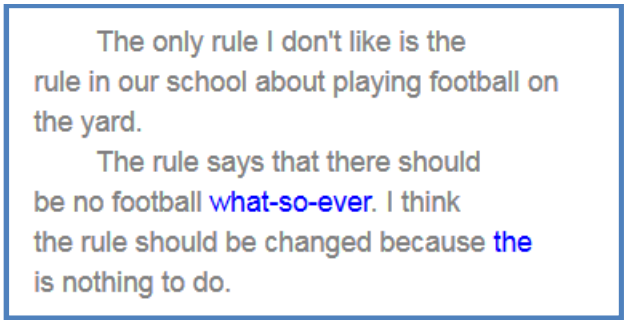
Figure 17. School Scripts: text alignment preserved (sample: W791ru_12132m)

Figure 17-a. PDF image


The only rule I don't like is the rule in our school about playing football on the yard.

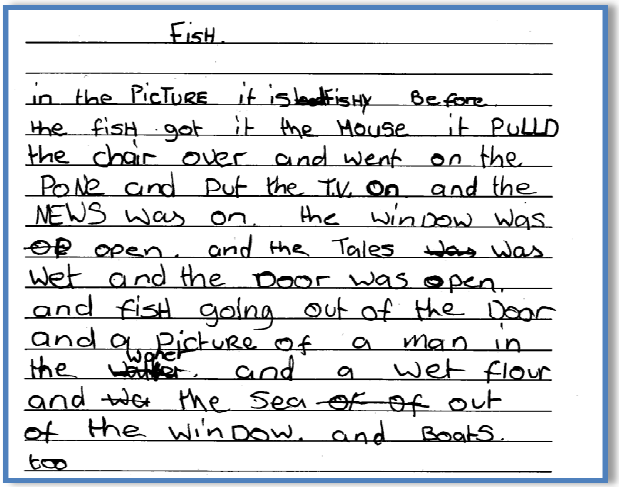
The rule says that there should be no football whatso-ever. I think the rule should be changed because the is nothing to do.

Figure 17-b. Display online



Lettering size changes have not been documented; that is, scripts with (seemingly) upper-case letters other than in word-initial position have been normalised to lower case. The reason for this is the difficulty to decide objectively on whether the letter form is intended to be an upper-case letter (misused) or whether the pupil makes use of the same letter form regardless of the place in which it appears. To give an example, in the file W881sp_23018f there is a clear inconsistent use of upper-case and lower-case letter forms in word-initial, medial and even final position (see Figure 18). The exception to this practice of spelling normalisation is when capital letters are used for emphasis or when they are used consistently in word-initial position of nouns such as *School*. Users interested in handwriting can explore the original scripts in the PDF images provided in the online interface.

Figure 18. School script with inconsistent use of upper-case letter forms (sample: W881sp_23018f)



Basal Readers

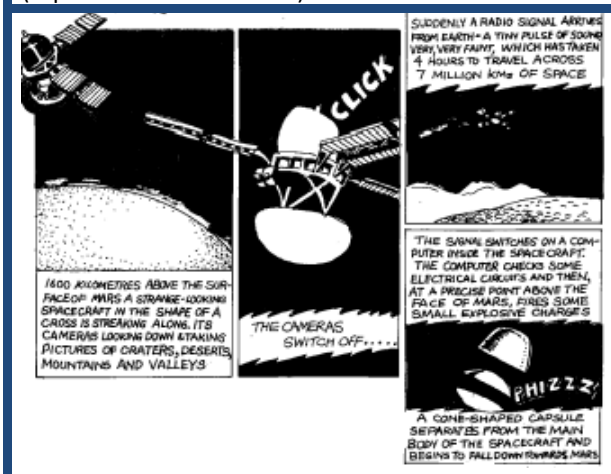
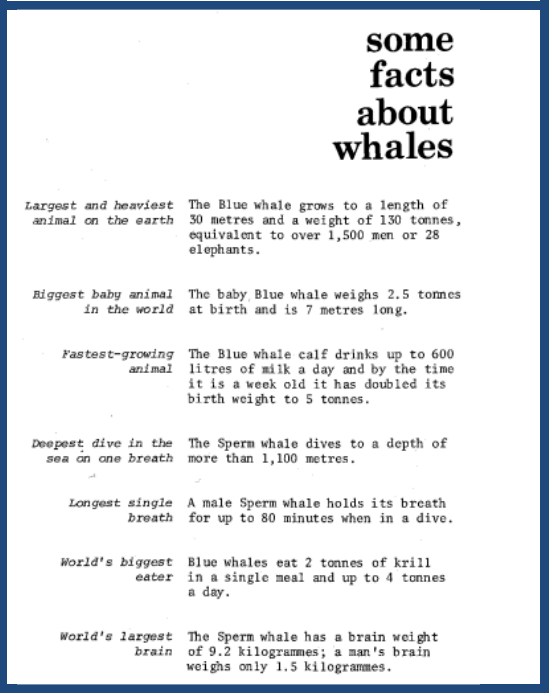
Transliterations of the basal readers retain the original spelling, including typos, should there be any.

Pictures have not been included, but their omission has been duly annotated in the XML version and they are displayed in the PDF images.

Tables and vignettes have been coded in XML format and are also displayed in the PDF original images. The transliteration is based on what we call a 'logical reading' protocol; that is, rather than adopting a common transliteration principle for all tables/vignettes (for instance, either top-down or left-right throughout), we selected one (top-down) or the other (left-right)

depending on which of the two would provide a more coherent reading of the table/vignette for a user who would (hypothetically) not have access to the original PDF documents. For instance, the text of the vignettes in BR791_spacestarts8 displayed on the left-hand column in Figure 19 has been transliterated in a top-down fashion, whereas the text of the table in BR791_Whales1 displayed on the right-hand column in Figure 19 follows a left-to-right transliteration practice.

Figure 19. Transliteration of tables and vignettes

Table from Basal Reader BR791_spacestarts8, p.5 (top-down transliteration)	Table from Basal Reader BR791_Whales1, p. 1 (left-right transliteration)
	

Words that are split across two lines are indicated by the symbol used in the original, if any, including a hyphen (-) or a double hyphen symbol (=) at the end of the first line, and sometimes repeated at the start of the next.

Abbreviations have been retained in their short form as written in the original script. The expanded form has been coded in the XML version.

Lettering size changes have been documented in the XML version only. Users interested in handwriting can explore the PDF images provided in the online interface.

Text alignment, lineation and paragraph indentation have been preserved.

All quotation marks are retained in the text and are represented by appropriate Unicode characters.

File extent

The corpus metadata documents the total **word count** and the total **line count** by school script and basal reader. The complete list is provided in an Excel file in the online interface.

Lines are counted from the first line with running text to the last line with running text, including intermediate blank lines. This applies to both components of the corpus. Counts have been calculated from the XML version with the XML editor oXygen.

The Perl-script used to count '**words'** was kindly produced by Prof. David Denison (University of Manchester), February 2016. The word count is an accurate word count of the actual text; that is, of everything in the file which is not enclosed in <…> brackets used for XML tags. Hyphenated words are counted as one item, as are all items other than punctuation surrounded by white space. Note that word counts have been calculated from the version with original spelling, deviations included.

In the "School Scripts" component, the Rule scripts tend to be shorter than the Story scripts. Rules usually extend over 100 words, while Stories usually go over 200 or even 300 words. In the "Basal Reader" component, the length of the samples varies considerably. Basal Readers used in the NFER surveys range from ca. 700 words to ca. 2,000 words; those compiled as supplement range from ca. 6,000 words to ca. 10,000 words, based on a 30-page sample.

The BROWSE tool in the APU online interface provides a variety of "frequency lists" with word counts for the untagged file with original spelling, the original file with normalised spelling, the POS-tagged file (POS tags alone and POS-word combinations), and the semantically tagged file (semantic tags alone and semantic-word combinations).

Editing and proof-reading

All texts in the "School Scripts" component of the APU corpus were keyed-in by hand from the digitised image of the original scripts. Most texts in the "Basal Readers" component were scanned with an Optical Character Recognition (OCR) program and then checked against the original digitised images; others were keyed in and then checked against the original digitised images.

In all cases, the initial transliteration was by the project's research assistants. Each research assistant was allocated a specific batch of texts (i.e. 1979 rule, 1979 story, 1988 rule, 1988 story, basal readers) in order to keep the transcription hands consistent across batches. All texts in both components were annotated in line with the project guidelines for mark-up, and then proofread in three rounds: by the project coordinators, by another research assistant, by a native speaker.

Ethical considerations

As indicated in the "Copyright Statement", the project members have formally agreed to observe the original privacy undertakings given to the participating children's parents and schools by ensuring that no child who participated in the surveys can be identified in any publication arising from the digitalisation of or research based on the materials. Ensuring that no child, classmates or relatives can be identified implies, as appropriate, anonymising names or citations in the .XML version and in the tagged versions, as well as blanking out written names on the digitised .PDF images. See for instance the anonymised passage in file W791ru_12005m.

Figure 20. Anonymised School Scripts (W791ru_12005m)

Figure 20-a. PDF image

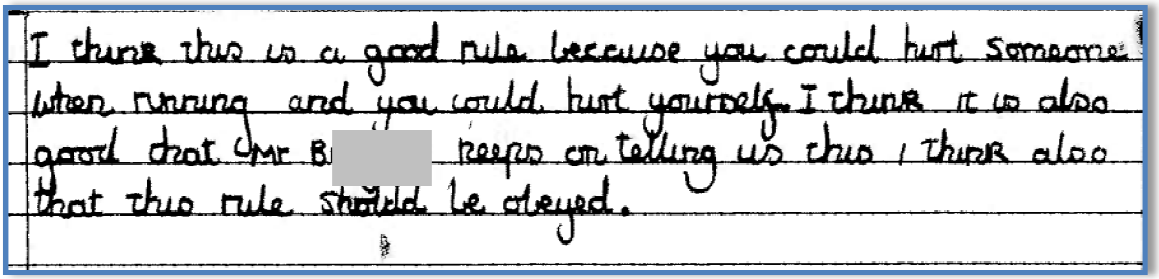


Figure 20-b. XML format

```
<lb/>I think this is a good rule because you could hurt someone
<lb/>when running and you could hurt yourself. I think it is also
<lb/>good that <name>Mr B</name> keeps on telling us this I think also
<lb/>that this rule should be obeyed.
```

Figure 20-c. Display online

I think this is a good rule because you could hurt someone when running and you could hurt yourself. I think it is also good that Mr B keeps on telling us this I think also that this rule should be obeyed.

Online Interface

Access

The APU corpus is freely available online at http://apucorpus.liverpool.ac.uk.

Access will be granted to interested users upon receipt of the APU User Agreement, whereby they will agree formally to the conditions of use. It is available on the website.

The APU transcriptions shall only be used for non-profit teaching and research. Extracts may be quoted under normal conditions of fair use and must acknowledge the source. The material drawn from the APU corpus, whether printed, in electronic, or any other form, is intended for the said registered user only and may not be distributed, or transferred to a third party.

The copyright statements for the APU materials are as follows:

School Scripts

Copyright rests with ©The University of Liverpool 2015 and based on the rights passed to us by NFER. The project members agree to observe the original privacy undertakings given to the participating children's parents and schools by ensuring that no child who participated in the surveys can be identified in any publication arising from the digitalisation of or research based on the materials, and agrees further to obtain equivalent written undertakings from any colleague involved in those processes. Ensuring that no child can be identified implies, as appropriate, anonymising quotations, blanking out written names or bleeping out spoken names.

The APU transcriptions shall only be used for non-profit teaching and research. Extracts may be quoted under normal conditions of fair use and must acknowledge the source.

Basal Readers

Copyright rests with ©The University of Liverpool 2015 and based on the rights passed to us by NFER. Reproduction of the images from the original supplementary materials has been kindly granted by the publishers, 2016.

The APU transcriptions shall only be used for non-profit teaching and research. Extracts may be quoted under normal conditions of fair use and must acknowledge the source.

Documentation

The following documents are available to registered users via the APU online interface:

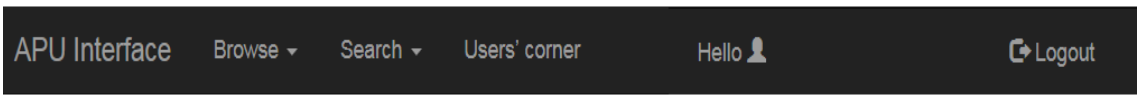
- Common to all files
  - XML tagset (TEI P5)
  - POS tagset (CLAWS7)
  - Semantic tagset (USAS)
- For each individual file
  - Frequency list, words in .TXT file (original and normalised)
  - Frequency list, POS tags
  - Frequency list, words and POS tags
  - Frequency list, semantic tags
  - Frequency list, words and semantic tags
  - Word counts per file

Functionalities

Two layouts have been built in the web-based application: BROWSE and SEARCH. It is also possible to DOWNLOAD the search hits into a CSV file.

The USERS' CORNER is an optional tool for users to document their work with APU, be it in form of publication or teaching materials.

Figure 21. Dashboard



Browse

The BROWSE tool allows users to read and explore individual files. Users can select the relevant School Script or Basal Reader from the BROWSE list in each section.

Figure 22. BROWSE tool: School Scripts file selection

School scripts files list

File ID	Year	Type of text	Pupil's sex	Attainment Band
W791ru_12001m	1979	Argumentative, persuasive	male	High
W791ru_12002m	1979	Argumentative, persuasive	male	Low
W791ru_12003m	1979	Argumentative, persuasive	male	Low
W791ru_12004m	1979	Argumentative, persuasive	male	Low

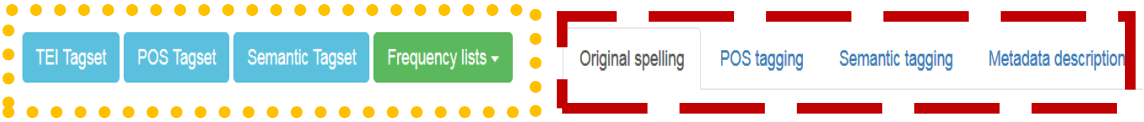
Figure 23. BROWSE tool: Basal Reader file selection

Basal readers files list

File ID	Year	Type of text
BR791_apples	1953	Narrative
BR791_dragondanger	1976	Narrative
BR791_dragons5	n/a	Narrative
BR791_flymachine9	1953	Narrative

Users can then explore the selected file in three different layouts, described and illustrated below in turn.

Figure 24. Layouts in the BROWSE tool



One can move from layout to layout by clicking on the tabs on the right-panel (red square with straight discontinuous lines) in Figure 24: "Original spelling", "POS tagging", "Semantic tagging". In addition, there is the tab "Metadata description" for bibliographic information of the selected file.

The tabs on the left-panel (yellowish square with dotted lines) in Figure 24 are displayed in all layouts. These are:

- "TEI tagset", related to the layout "Original spelling", XML file;
- "POS tagset", related to the layout "POS tagging";
- "Semantic tagset", related to the layout "Semantic tagging";
- "Frequency lists", which includes the list of words in (a) the untagged file with original spelling, (b) the original file with normalised spelling, (c) the POS-tagged file (POS tags alone and POS-word combinations), and (d) the semantically tagged file (semantic tags alone and semantic-word combinations).

Figure 25. BROWSE tool: POS tagset view

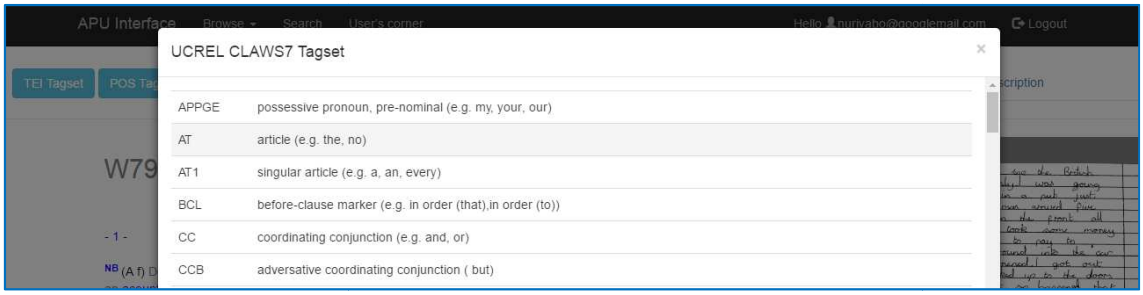
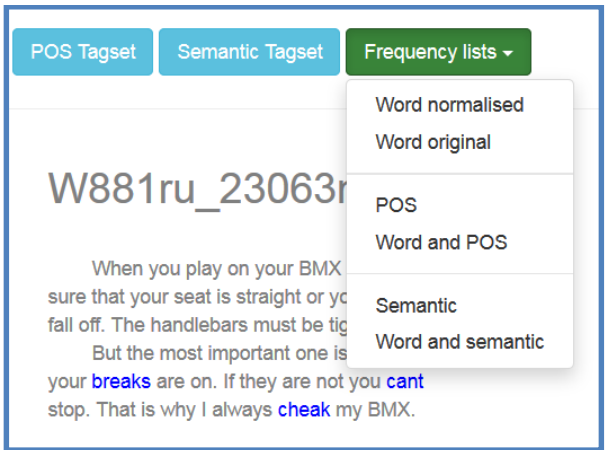
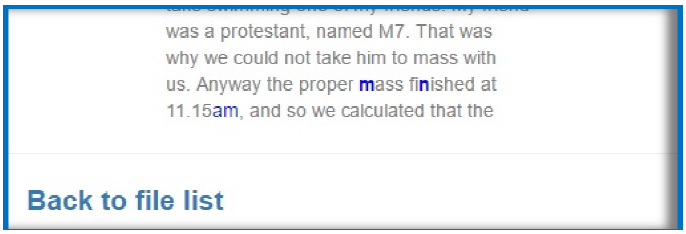


Figure 26. BROWSE tool: Frequency lists drop-down menu



After browsing the relevant file, users can return to the main file list by clicking "Back to file list" at the bottom of the page.

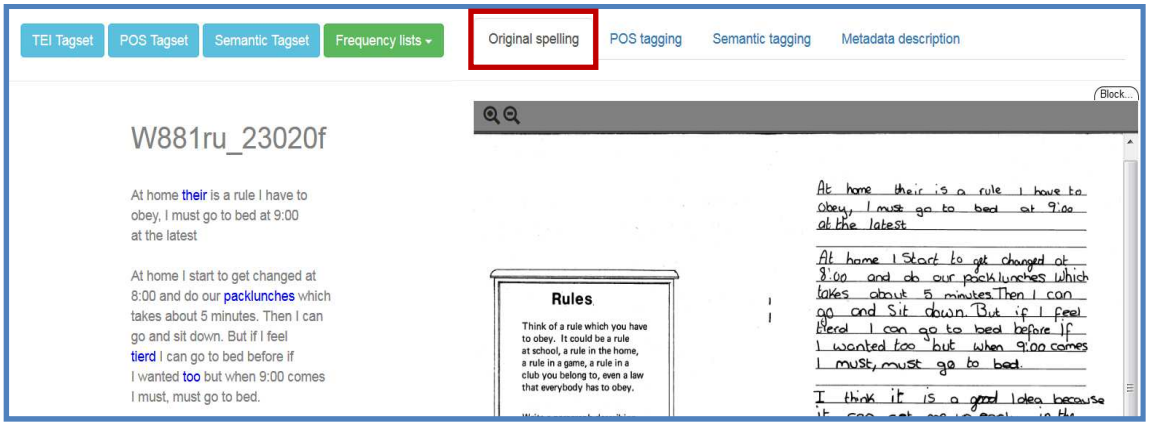
Figure 27. BROWSE tool: Return to the BROWSE list of files



Layout with XML and PDF

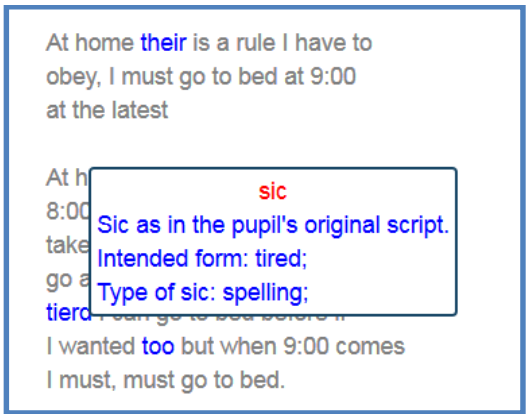
On clicking on a file, users are taken to the default layout in which the School Script/Basal Reader is shown in **running text with original spelling and XML tags side by side the digitised PDF image** – see Figure 28.

Figure 28. Layout I: text with original spelling side by side PDF image (sample: W881ru_23020f)



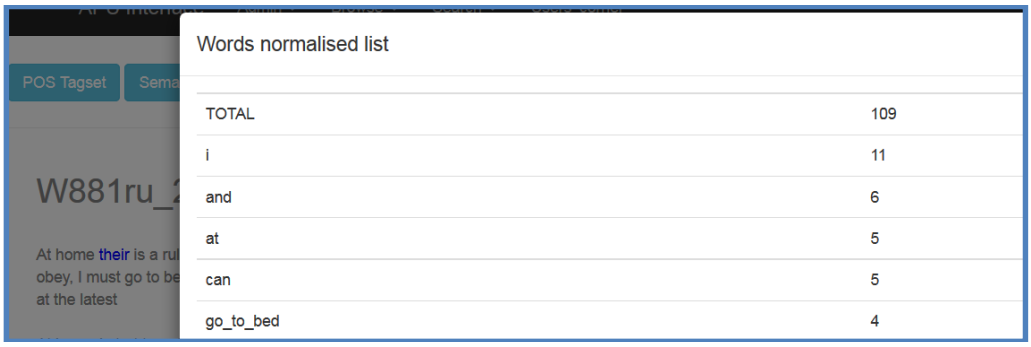
To the left of the screen is the original text from the .XML file annotated with a subset of TEI tags. Words or strings of words to which a tag has been added are flagged in blue colour. Mousing over the relevant word(s) will display a **pop-up bubble with the annotation in it**; see, for instance, the sic tag for the misspelled word in Figure 29. For further details about the TEI subset of tags used in the APU corpus and how they are displayed, see the XML tagset document available in the online interface.

Figure 29. Layout I: Pop-up bubbles with tag description (sample: W881ru_23020f)



Users can consult the **word frequency lists** by clicking on the green button above the transliteration, for both the text with original spelling and with normalised spelling (see Figure 26 above and Figure 30 below).

Figure 30. Layout I: Word frequency list (sample: W881ru_23020f)



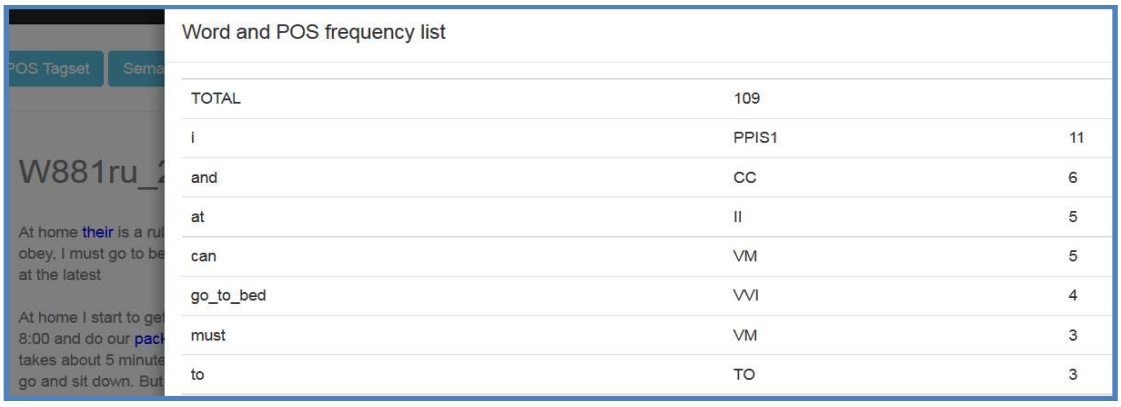
To the right of the screen is the **digitised image** in .PDF format. Users can zoom in or out as convenient. Due to copyright restrictions, images cannot and must not be downloaded, printed or reproduced otherwise. As explained in the section on "Ethical Considerations", personal names have been blanked out in order to comply with anonymity and confidentiality obligations.

Layout with XML and POS tagging

A second layout displays the **text with original spelling with XML tags side by side the text with part-of-speech tagging**, as illustrated in Figure 31. The output from W-Matrix displays the POS tag attached to the word with underscore. For a better visualisation, different tags are displayed in different colours. Users can consult the POS tagset (CLAWS7) and the POS frequency list from the top-left banner (Figure 32).

Figure 31. Layout II: text with original spelling side by side text with POS tagging (sample: W881ru_23020f)

Figure 32. Layout II: POS and word frequency list (sample: W881ru_23020f)



Layout with XML and semantic tagging

The third layout works identical to the second layout, this time displaying the **text with original spelling and XML tags side by side the text with semantic tagging**, as illustrated in Figure 33. The output from W-Matrix displays the semantic tag attached to the word with underscore. For a better visualisation, different tags are displayed in different colours. Users can consult the semantic tagset (USAS) and the semantic tag frequency list from the top-left banner (Figure 34).

Figure 33. Layout III: text with original spelling side by side text with semantic tagging (sample: W881ru_23020f)

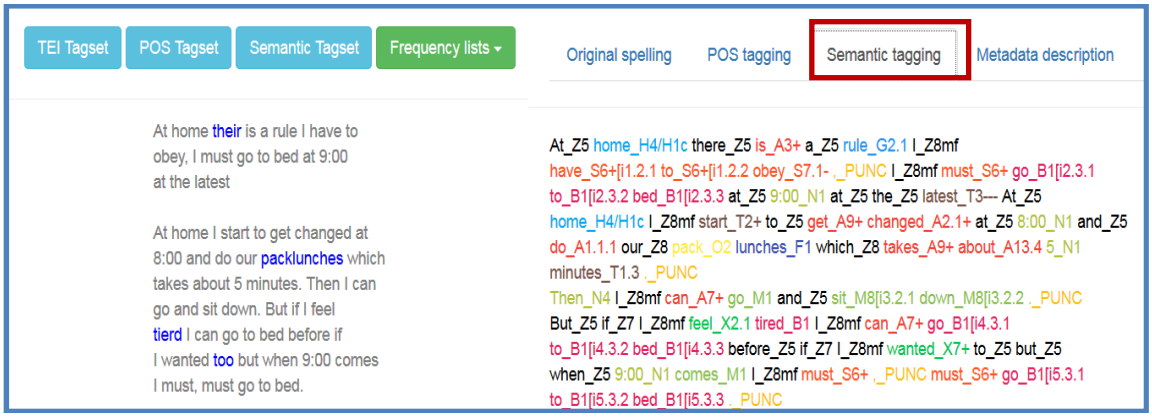
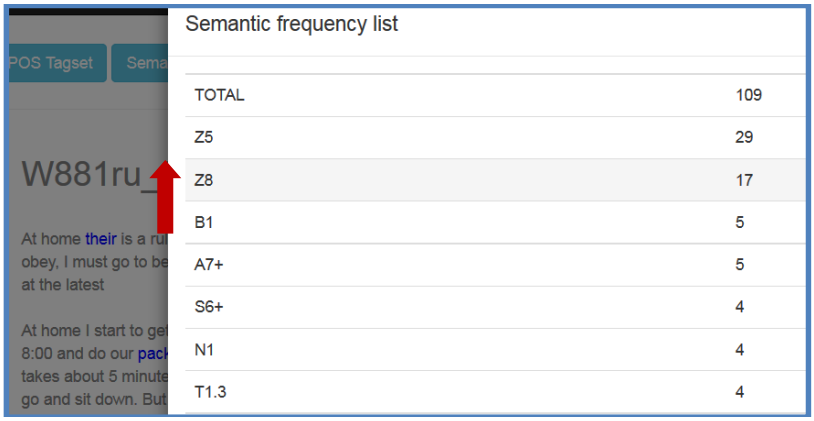


Figure 34. Layout III: Semantic frequency list (sample: W881ru_23020f)



Search

The SEARCH tool allows users to search the corpus in various ways attending to users' interests, for both the "School Scripts" and "Basal Readers" components.

There are two main sections: search Metadata, and search Data. Click on the button "Search" to retrieve the search results, and on "Clear search" to clear all selected parameters at once. (Users should avoid pressing "Enter".)

Figure 35. SEARCH tool: School Scripts

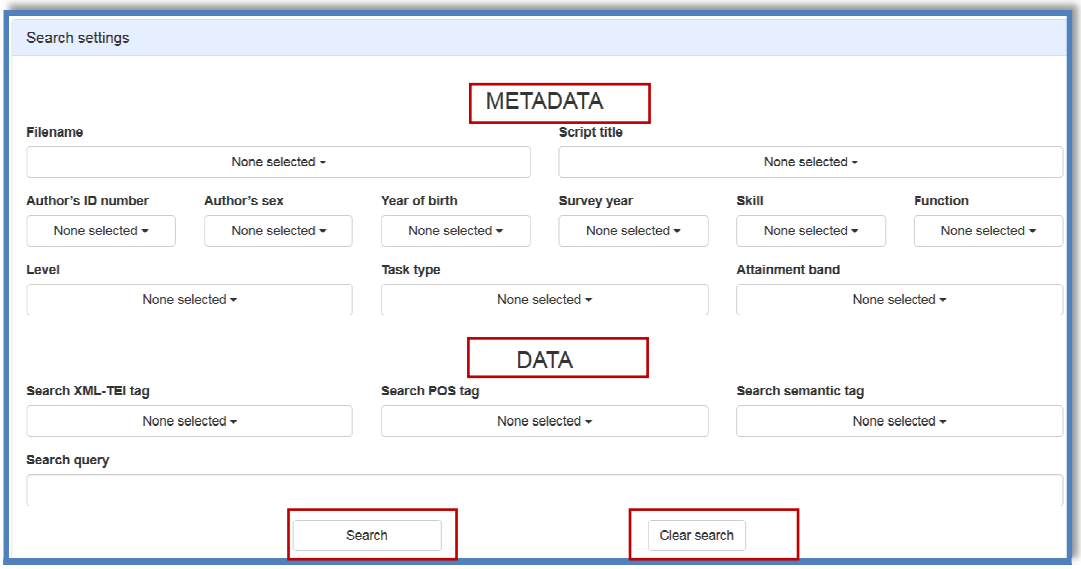
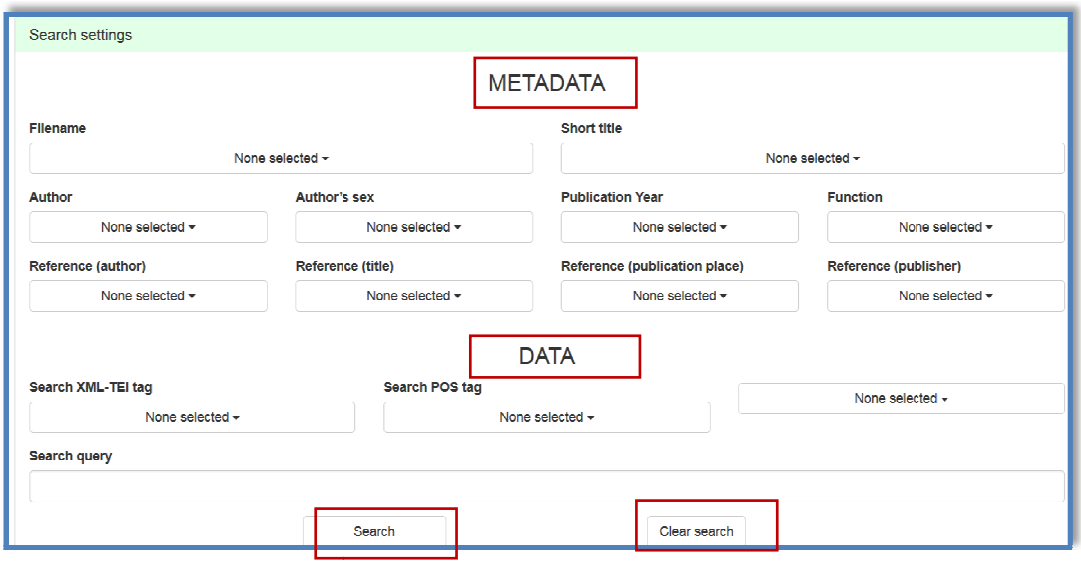


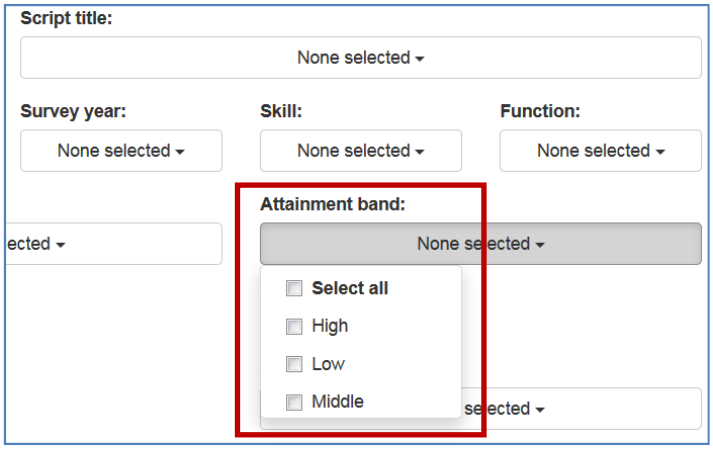
Figure 36. SEARCH tool: Basal Readers



Users can search in a particular field only or in a combination of fields. The default search value is 'None Selected'; consequently, if no restriction is made on the Search field, all/any instances in the fields displayed will automatically be retrieved. Note that the search tool is not case-sensitive.

A number of fields offer a drop-down list with predefined values; for instance "male" and "female" in Author's sex; "High", "Mid" and "Low" in "Attainment band", and so on.

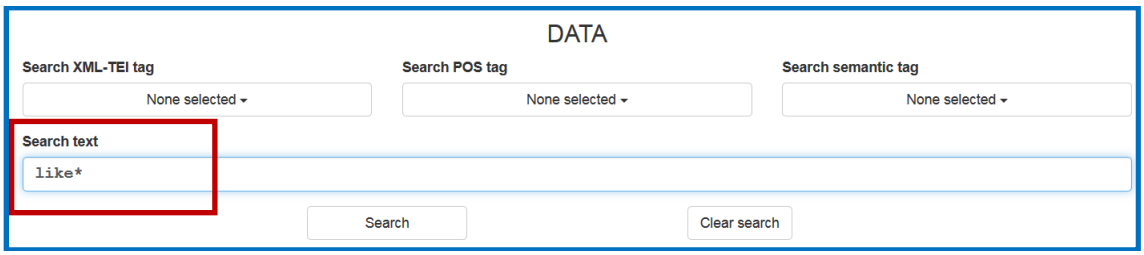
Figure 37. SEARCH tool: drop-down menu with predefined values



Two wildcards are available to users to facilitate their search, namely * and |, as follows:

Wildcard	Function	Sample	Output
*	zero or more characters	like*	*like, likes, liking*
|	search term OR search term	run|ran	*run, ran*

Figure 38. SEARCH tool: wildcards



Search data

The following options are available for searching data:

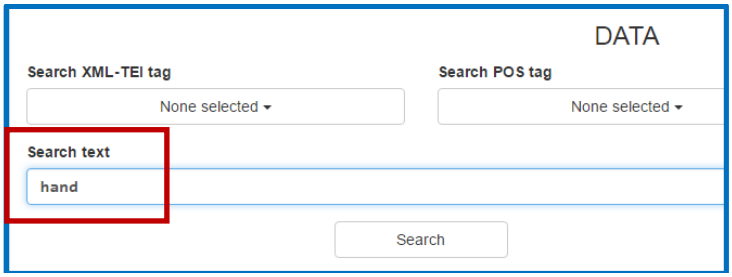
- search text only — by word or string of words, verbatim or with wildcards
- search by tags only — users can select several tags at the same time
  - search by XML-TEI tag — select from a drop-down value list
  - search by POS tag — select from a drop-down value list
  - search by semantic tag — select from a drop-down value list
- combined searches — searches that combine words/lemmas, tags and/or wildcards; users can select several tags at the same time

The combination of queries is varied. The use of wildcards is available for any search. Below are some illustrative examples.

(1) Search text only

Type any word or string of words in the box "Search text". This will retrieve hits with that word or string of words regardless of the XML/POS/semantic tag. For instance, a search for *hand* will retrieve all instances of *hand* (in that very same spelling) as both noun (singular) and verb (base form).

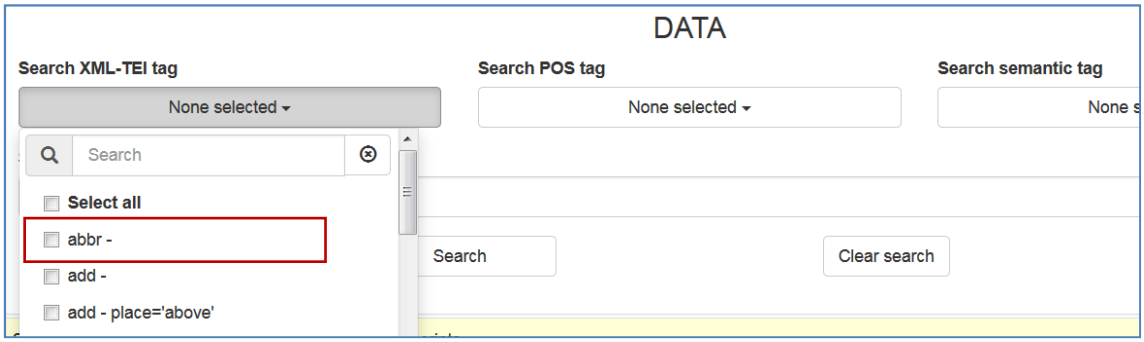
Figure 39. Search text: *hand*



(2) Search by XML tag only

Keep the box "Search text" empty and select the relevant XML-TEI tag from the drop-down menu in the box "Search TEI-XML tag". For instance, a search by the XML tag <abbr> will retrieve all instances in which a word appears in the original text in abbreviated form, e.g. *tv* for 'television', *P.E.* for 'Physical Education', etc. More than one tag can be selected in one single search. The results displayed by the concordance will *not* automatically show the XML tags; users should understand that the results will be displayed in the order in which the tags appear in the "Search XML_TEI tag" in drop-down menu (i.e. alphabetical order). For instance, if a user searches for the XML tags <abbr> and <sic>; all <abbr> hits will be displayed first, and all <sic> instances second.

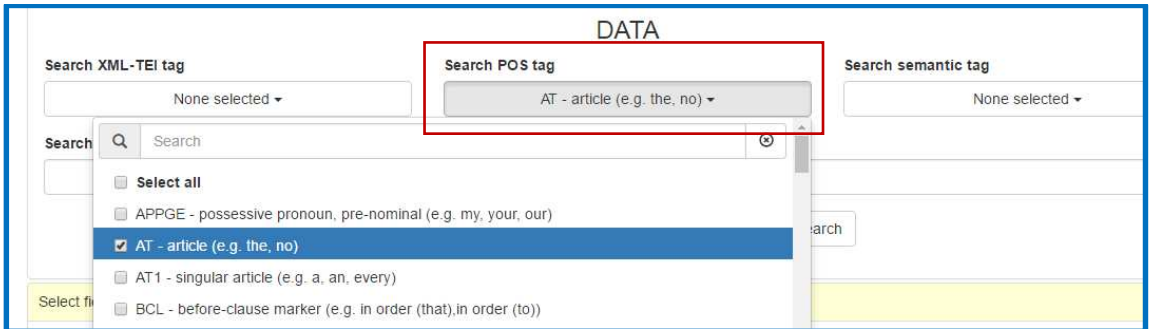
Figure 40. Search text: XML tag <abbr>



(3) Search by POS tag only

Keep the box "Search text" empty and select the relevant POS tag from the drop-down menu in the box "Search POS tag". For instance, a search by the POS tag AT will retrieve all instances of articles. More than one tag can be selected in one single search. As noted above, users should bear in mind that the results displayed by the concordance will *not* automatically show the POS tags but will organise the results based on the alphabetical order of the relevant tags.

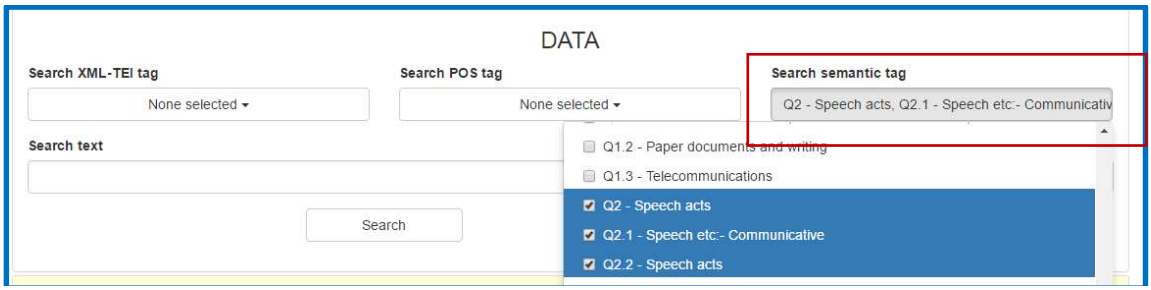
Figure 41. Search text: POS tag AT



(4) Search by semantic tag only

Keep the box "Search text" empty and select the relevant semantic USAS tag from the drop-down menu in the box "Search semantic tag". For instance, a search by the semantic tag Q2 will retrieve all instances of 'Speech acts'. More than one tag can be selected in one single search. As noted above, users should bear in mind that the results displayed by the concordance will *not* automatically show the semantic tags but will organise the results based on the alphabetical order of the relevant tags.

Figure 42. Search text: semantic tag Q2 'Speech acts'

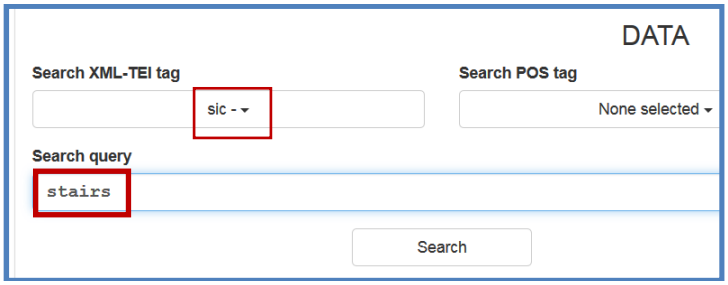


COMBINED SEARCHES

(5) Search text + XML tag

(a) Type the word under consideration the box "Search text" and select the relevant XML-TEI tag from the drop-down menu in the box "Search TEI-XML tag". For example, the search for *stairs* and the tag *sic* will retrieve all cases for which the word *stairs* has been spelled in a non-standard fashion, e.g. *stiars*, *staris*, *staires*, *stair's*.

Figure 43. Search text: *stairs* + <sic>



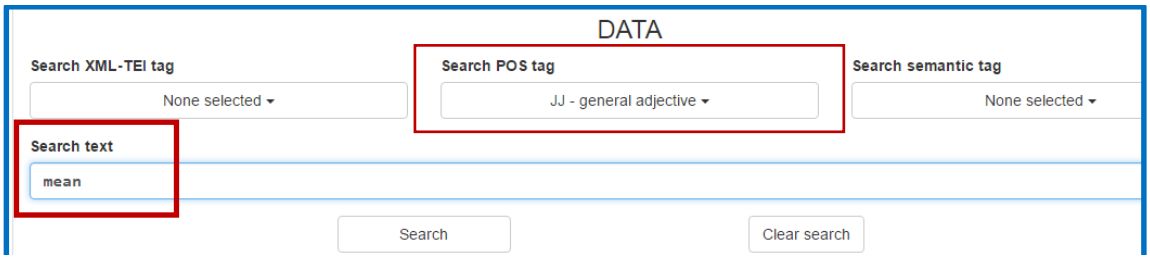
(6) Search text + POS tag

There are two options here.

(a) Type any word in the box "Search text" and select the relevant POS-tag from the drop-down menu in the box "Search POS tag". For instance, a search for *mean* and the POS-tag _JJ will retrieve all instances in which the word *mean* has been tagged as a general adjective (compared to its function as a verb or noun).

(b) Similar results can be obtained by introducing the search string [mean_JJ] in the "Search text" box.

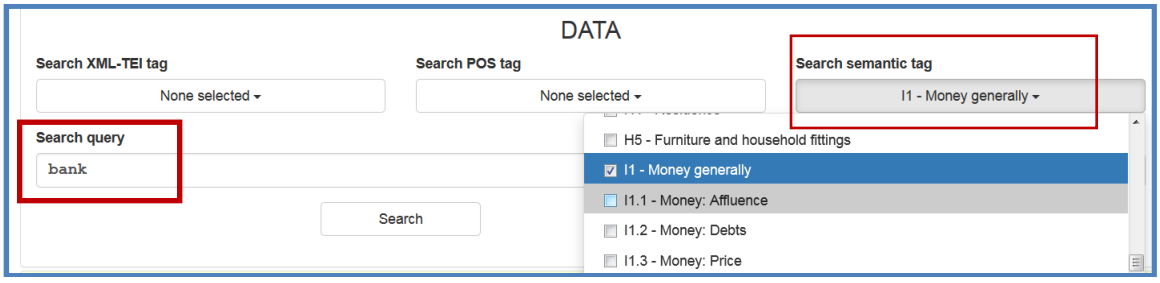
Figure 44. Search text: *mean* tagged as adjective



(7) Search text + semantic tag

As above, this type of search can be carried out by introducing a relevant word in the box "Search text" and selecting the relevant semantic tag from the drop-down menu in the box "Search semantic tag". For instance, a search for the word *bank* and the semantic tag _I1 will retrieve all instances in which the word *bank* has been tagged in relation to "money generally", compared to *bank* coded with the tag "geographical term" (e.g. *river bank*). Similar results can be obtained by introducing the search string [bank_I1] in the "Search text" box.

Figure 45. Search text: *bank* tagged as I1



Combinations of the different search types described above are also possible. For instance, Figure 46 reflects the results of a combined search by lemma, POS-tag and wildcard aimed at retrieving from the corpus all Noun Phrases which begin with the determiner *some*, followed by an adjectival premodifier and a common single noun as phrasal head: some *_JJ *_NN1.

Figure 46. Combination of searches: some *_JJ *_NN1

Figure 46-a. Search query

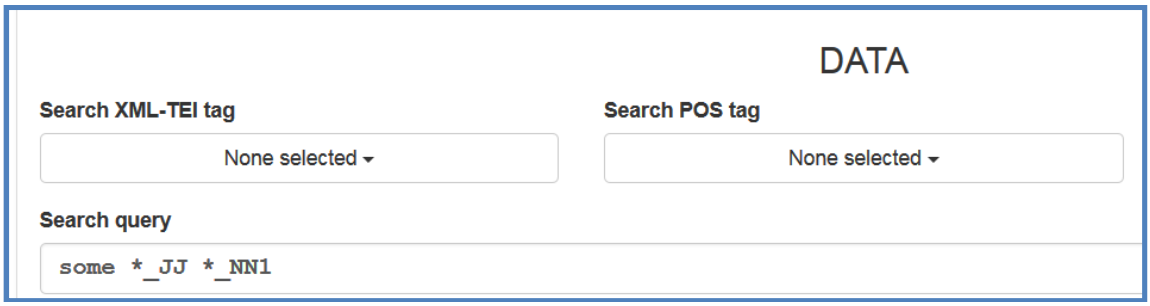
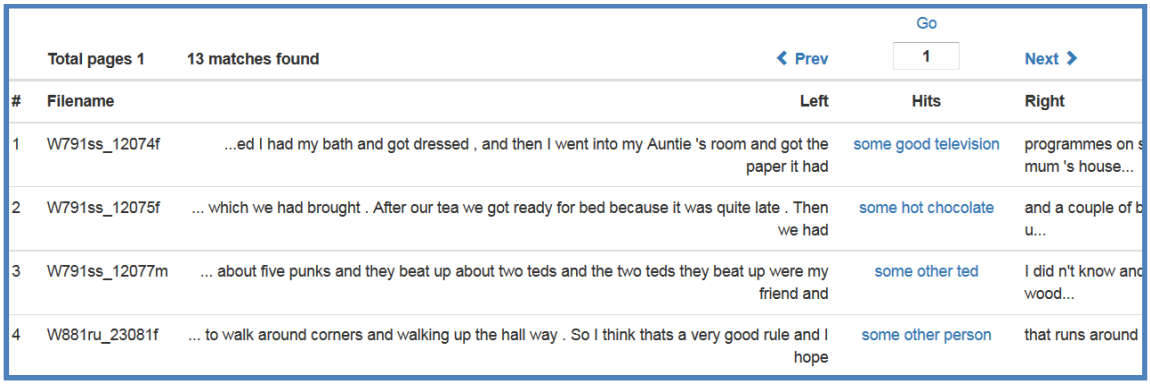


Figure 46-b. Search results



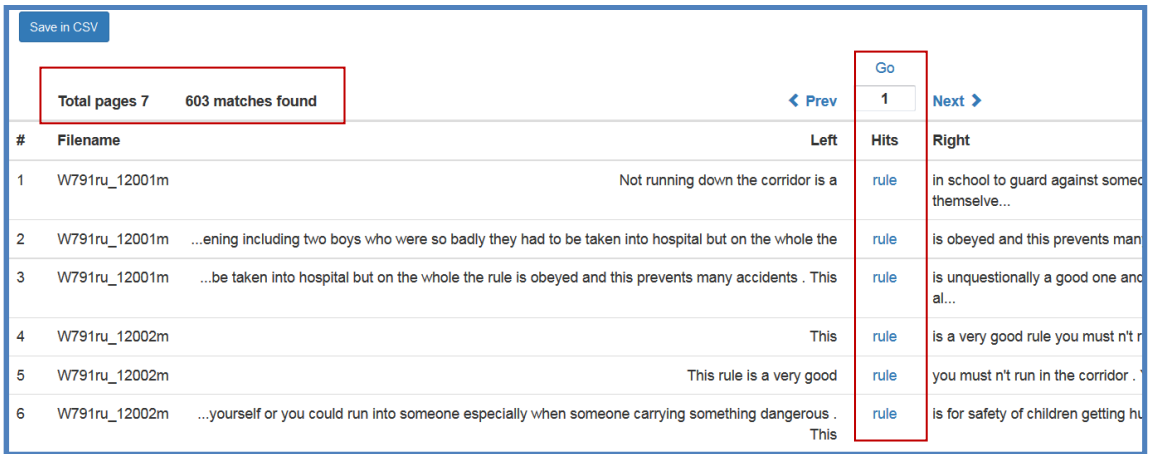
Total pages 1 | 13 matches found | Prev | Go 1 | Next

#	Filename	Left	Hits	Right
1	W791ss_12074f	...ed I had my bath and got dressed , and then I went into my Auntie 's room and got the paper it had	some good television	programmes on s mum 's house...
2	W791ss_12075f	... which we had brought . After our tea we got ready for bed because it was quite late . Then we had	some hot chocolate	and a couple of b u...
3	W791ss_12077m	... about five punks and they beat up about two teds and the two teds they beat up were my friend and	some other ted	I did n't know and wood...
4	W881ru_23081f	... to walk around corners and walking up the hall way . So I think thats a very good rule and I hope	some other person	that runs around

Search hits

On clicking “Search”, the Search Results page will display as in Figure 47.

Figure 47. Search results display for the item *rule*



Save in CSV

Total pages 7 | 603 matches found | Prev | Go 1 | Next

#	Filename	Left	Hits	Right
1	W791ru_12001m	Not running down the corridor is a	rule	in school to guard against someo themselve...
2	W791ru_12001m	...ening including two boys who were so badly they had to be taken into hospital but on the whole the	rule	is obeyed and this prevents man
3	W791ru_12001m	...be taken into hospital but on the whole the rule is obeyed and this prevents many accidents . This	rule	is unquestionally a good one and al...
4	W791ru_12002m	This	rule	is a very good rule you must n't r
5	W791ru_12002m	This rule is a very good	rule	you must n't run in the corridor .
6	W791ru_12002m	...yourself or you could run into someone especially when someone carrying something dangerous . This	rule	is for safety of children getting hu

The total number of hits is indicated at the top. The default settings will display 100 records per page. Users can move from page to page with the buttons on top of the table.

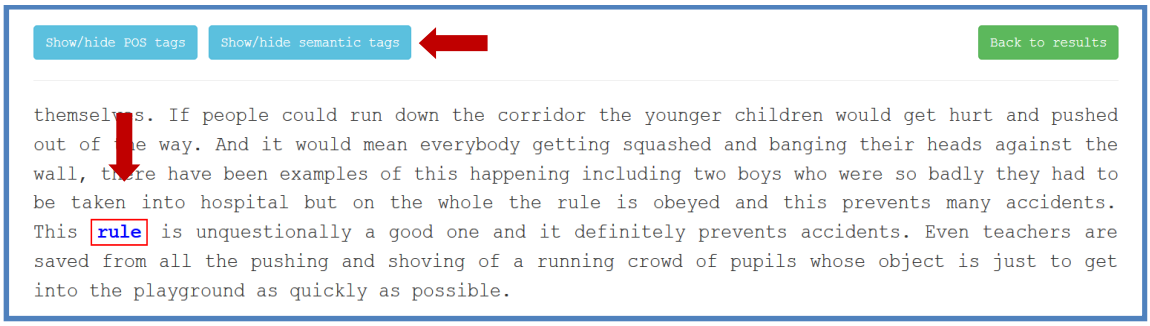
The search hits are displayed in KWIC concordance format; that is, *Key Word In Context*, aligned in the centre of the record line.

Clicking on the search hit of any record will show the expanded context for that particular instance (see Figure 48). The searched word is highlighted for easy identification. For copyright reasons, the search engine must limit the amount of context that can be viewed and downloaded to approximately 600 characters Left and Right. The limits should be adequate for most linguistic purposes.

Notice that the expanded view shows the text in normalised spelling, since this is the necessary input for tagging with W-Matrix (see section "Formats").

Users can show or hide POS tags or semantic tags by clicking on the top-left tabs.

Figure 48. Search results display, expanded view for individual records



Download

Users may download the search results by clicking on the tab "Save CSV" on the top-left corner of the screen displaying the result hits (Figure 49). Users are given the option to select the fields they wish to download (Figure 50). The downloadable file is saved in CSV and then easily converted to XLS (Excel) format.[1]

Figure 49. Download search results I

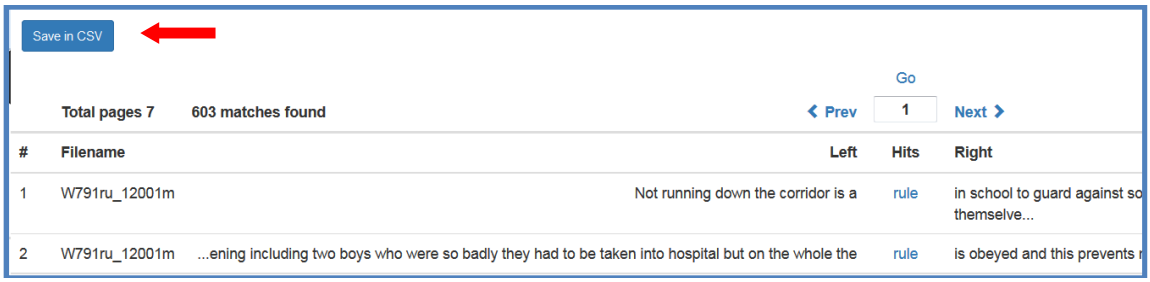
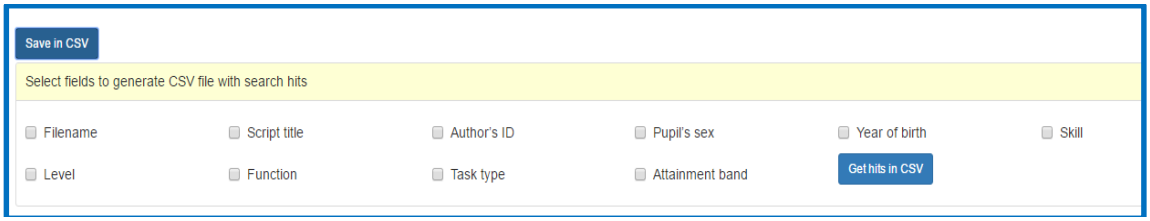


Figure 50. Download search results: select fields



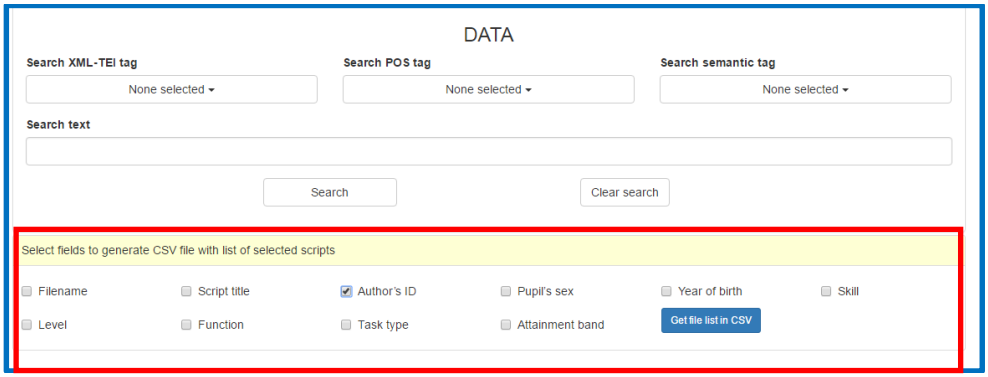
[1] Technical note: if the CSV file is opened with Excel, the comma-separated-values should automatically display in separate columns, as shown in Figure 51. Should that not be the case, you will need to adjust the settings in Excel; this depends on the version of the operating system.

Figure 51. Download search results: csv file

	1	2	3	4	5	6	7	8
1	filename	pupil_sex	function	attainment_b	left	center	right	
2	W791ru_12001m	male	Argumentative,	High	Not running down the corridor is a	rule	in school to guard against someone runni	
3	W791ru_12001m	male	Argumentative,	High	Not running down the corridor is a rul	rule	is obeyed and this prevents many accider	
4	W791ru_12001m	male	Argumentative,	High	Not running down the corridor is a rul	rule	is unquestionally a good one and it defin	
5	W791ru_12002m	male	Argumentative,	Low	This	rule	is a very good rule you must n't run in the	
6	W791ru_12002m	male	Argumentative,	Low	This rule is a very good	rule	you must n't run in the corridor . You mus	
7	W791ru_12002m	male	Argumentative,	Low	This rule is a very good rule you must	rule	is for safety of children getting hurt and i	

Users can also download metadata information directly from the Search window, by selecting fields at the bottom of the page (Figure 52).

Figure 52. Download search results: metadata



Users' Corner

The USERS' CORNER is an optional tool for users to document their work with APU, be it in the form of academic publication, teaching materials or any other type of research and/or educational resources. The Users' Corner intends to be a shared space for APU users to disseminate and publicise their own work and/or practice, as well as a way of creating an international community of practice for scholars and educators interested in corpus-based approaches to English language teaching and learning in the UK. As illustrated in Figure 54, prospective users are asked to provide a brief description of their work, URL if applicable, and they can upload any materials they wish in zip format. Please note that, at this stage, only documents in Word, .TXT or PDF format are allowed. Copyrights of any materials uploaded rest with the author(s).

For any questions concerning the Users´ Corner, please e-mail us at apucorpus@liv.ac.uk.

Figure 53. Users' Corner I

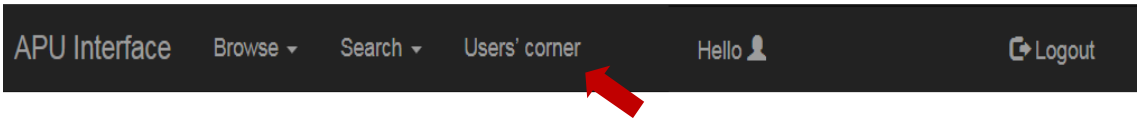
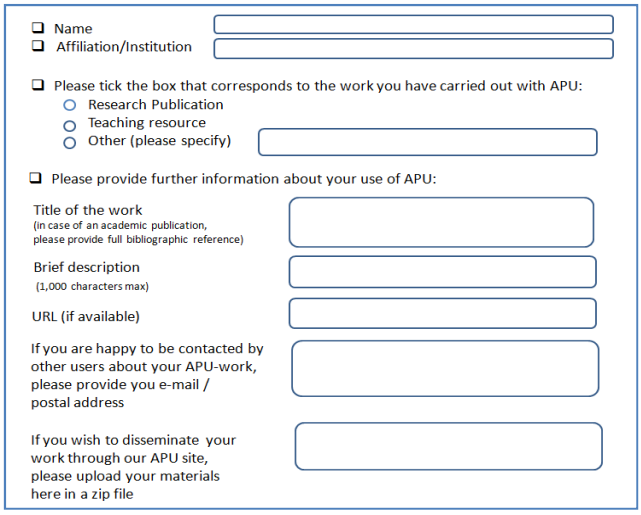


Figure 54. Users' Corner II



Multidimensional analysis

MD approach

In line with the educational trends of the time, the APU language team believed in approaching the study of language from a **functional perspective** (Stubbs 1986: 29, Hudson 2003). More specifically, the assumptions underlying the language assessment framework were mainly that (a) language is a *purposeful* activity, (b) that performance and attitudes are *inter-related*, and (c) that assessment should be focused not only on *what* is said but also on *how* it is said (see further Gorman 1986: 2-4, Foxman et al. 1991: 28-31). The overall aim was therefore to develop tasks "related as closely as possible to ways in which language is actually used" in their daily life (Thornton 1987: 2, also White 1986: 1). Thus the Writing Surveys were based on a range of tasks that reflected the "purposes and uses for which pupils produced writing" (Gorman 1986: 15). Of the twelve main functions of language considered by the APU team, the APU corpus contains two: **argumentation-cum-persuasion and narrative-cum-descriptive**.

In current corpus-based studies, one of the most well-known approaches to the study of genre and text-type variation is the **multidimensional approach to linguistic analysis (MDA)**. It was first introduced by Biber (1988), applied to synchronic register variation in English and in adults' writings. Later studies have proven its applicability to historical register variation, to university students' writings, and further to other languages (e.g. Biber 1995, Biber 2006, Biber & Finegan 1997). The MD method is based on complex computational and quantitative tools, whereby co-occurring patterns are identified empirically and quantitatively with a **Factor Analysis** (a multivariate statistical technique). Central to the MDA is the notion of **linguistic co-occurrence** and **frequency scores**, in the belief that linguistic features which co-occur frequently tend to share the same communicative function. The so-called '**dimensions of linguistic variation**' are thus sets of co-occurring variables along a continuum representing a particular situation or function, namely Dimension 1 "Involved vs. Informational Discourse", Dimension 2 "Narrative vs. Non-Narrative Concerns", Dimension 3 "Explicit vs. Situation-Dependent Reference", Dimension 4 "Overt Expression of Persuasion", Dimension 5 "Abstract vs. Non-Abstract Information", Dimension 6 "Online Informational Elaboration". As inferred from the labels, Dimension 2 and Dimension 4 are particularly relevant for the analysis of the data compiled in the APU corpus – **D2 for the Story task**, and **D4 for the Rule task**.

MD and APU

Given the importance of the MD approach for the study of genre/register variation, we have carried out a number of MD analyses with the data compiled in the APU corpus. The software used is the **Multidimensional Analysis Tagger** (MAT), which automatically tags the input files (based on an expansion version of the Stanford Tagger), runs statistical tools, and provides multidimensional functional analyses in a variety of output files, including plot boxes for each dimension. In essence, MAT aims to replicate Douglas Biber's methodology as applied in *Variation across Speech and Writing* (1988).

We have **applied MAT to the APU corpus** in various sets of files, to meet users' interests as much as possible. The materials available via the APU online interface are provided in zip files, classified as indicated in Table 7.

Table 7. APU materials run by MAT

Combination	Sets of zip files
Task – Year	• Rule 1979, Rule 1988 • Story 1979, Story 1988
Sex – Year	• Female 1979, Female 1988 • Male 1979, Male 1988
Task – Sex	• Rule female, Rule male • Story female, Story male
Task – Sex – Year	• Rule female 1979, Rule female 1988 • Rule male 1979, Rule male 1988 • Story female 1979, Story female 1988 • Story male 1979, Story male 1988
Attainment band – Year	• High 1979, High 1988 • Middle 1979, Middle 1988 • Low 1979, Low 1988
Task – Attainment band	• Rule high, Rule middle, Rule low • Story high, Story middle, Story low
Task – Attainment band – Year	• Rule high 1979, Rule high 1988 • Rule middle 1979, Rule middle 1988 • Rule low 1979, Rule low 1988 • Story high 1979, Story high 1988 • Story middle 1979, Story middle 1988 • Story low 1979, Story low 1988

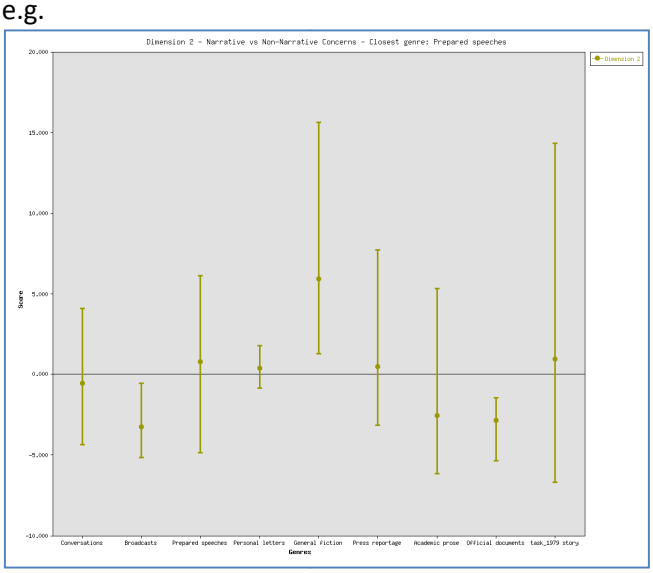
Each zip file contains **ten individual files**: three files in Excel format (Dimensions, Statistics, Z-scores)[2] and seven images in PNG format (one for each dimension plus one for text-type analysis). Details as follows:

- Option for "no correction" for z-scores.
- VASW tags only.

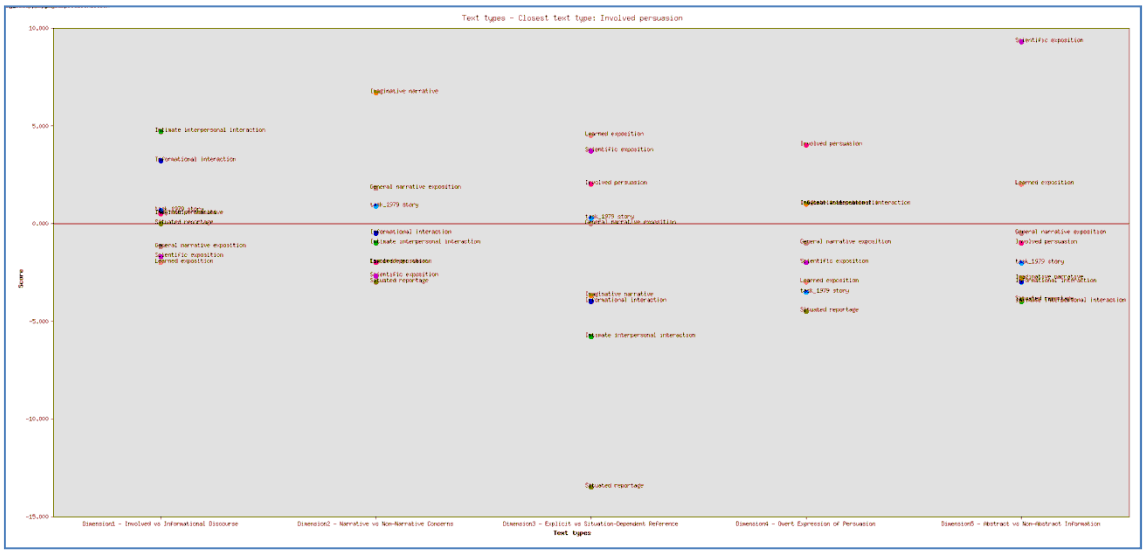
[2] The output files in MAT are provided as tab delimited .txt files, which we have converted to Excel format.

- Tool “Tag and Analyse”.
- Type-token ration at 400 (default value).
- File “Statistics.xlsx”: an Excel file that shows the frequency per 100 tokens for the linguistic variables found in the input files. It displays the tags used in Biber (1988).
- File “Zscores.xlsx”: an Excel file that includes the z-scores of the linguistic variables for the input files. The average for the selected files is also showed. The z-scores are calculated on the basis of the means and standard deviations presented in Biber (1988: 77). For each text and for the selected files as a whole, the program will flag all the z-scores with a magnitude higher than 2 as ‘Interesting variables’. It displays the tags used in Biber (1988).
- File “Dimensions.xlsx”: an Excel file that contains the scores for the Dimensions as well as the averages for the selected files. The Dimension scores are calculated using the z-scores of the variables that presented a mean higher than 1 in the chart presented in Biber (1988: 77). The program classifies each text according to its closer text type as proposed by Biber (1989) using Euclidean distance. The average for the selected files is also provided. We have chosen to not use the z-score correction.
- Images “Dimension#.png”: a graph that displays the location of the input texts’ Dimension score compared to a number of genres as shown in Biber (1988: 172). The graph displays the mean and the range for each genre. The mean and the range for the corpus are displayed too. The program will print the closest genre to the user’s selected texts next to the title of the graph. We have chosen the option to produce graphs for the six Dimensions.

  e.g.



- Image “Text_types.png”: a graph representing the location of the analysed selection of texts in relation to Biber’s (1989) eight text types. The program will print the closest text type to the user’s selection of texts next to the title of the graph. Text types are assigned using Euclidean distance.
  e.g.



Acknowledgement

MAT has been developed by **Andrea Nini**. For information about the tagger, manual, publications, etc., see his website at http://sites.google.com/site/multidimensionaltagger. We are very grateful to Andrea Nini for kindly giving us permission to post in our interface the results of running the APU files by MAT (version 1.3, February 2017).

References on the MD approach

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, Douglas. 1989. A typology of English texts. *Linguistics* 27, 3-43.

Biber, Douglas. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.

Biber, Douglas. 2006. *University language: A corpus-based study of spoken and written registers*. Amsterdam & Philadephia: John Benjamins.

Biber, Douglas & Edward Finegan. 1997. Diachronic relations among speech-based and written registers in English. In Terttu Nevalainen & Leena Kahlas-Tarkka (eds.), *To explain the present: Studies in the changing English language in honour of Matti Rissanen*, 253-75. Helsinki: Société Néophilologique.

Nini, Andrea. 2015. *Multidimensional Analysis Tagger* (Version 1.3). Available at: http://sites.google.com/site/multidimensionaltagger (Accessed February 2017)