

THE EF CAMBRIDGE OPEN LANGUAGE DATABASE (EFCAMDAT) USER MANUAL PART I: WRITTEN PRODUCTION

Jeroen Geertzen, Theodora Alexopoulou, Rachel Baker, Henriëtte Hendriks,
Sichu Jiang, Anna Korhonen
Department of Theoretical and Applied Linguistics, University of Cambridge
and
EF Education First

Version 0.2 — July 2013

1 Preface

The EF-Cambridge Open Language Database, henceforth EFCAMDAT, is the first large project of the EF-Research Unit at the Department of Theoretical and Applied Linguistics at the University of Cambridge. The Unit was launched in 2010 to promote research in second language learning of English and innovation in language teaching through a systematic cross-fertilisation between linguistic research and teaching techniques.

Soon after the launch of the EF-Research Unit, it became apparent that we had a unique opportunity to build a large scale database of learner English. The magnitude of EF's operations as a global educational organisation meant potential access to a vast pool of learners, literally, from hundreds of countries across five continents. Together with EF's global reach came a second very critical feature of the EF learning environment: *Englishtown*, EF's online English school. *Englishtown* offers online teacher-led lessons and interactive activities, including writing assignments for various levels of proficiency. Thousands of learners access Englishtown daily to follow lessons and submit work, providing a unique opportunity for the electronic collection of significant numbers of written data. In addition, learners are asked to write on a range of topics following a common curriculum allowing considerable knowledge on their input. Crucially, data collection from Englishtown allows collection of individual longitudinal data, since learner productions could be followed for individual learners over time.

The exciting opportunity to create a database of considerable scope and size came with important questions about its structure, access and usability. Researchers working with large scale databases like EFCAMDAT cannot rely on manual inspection, annotation

or extraction of data, if the quantitative power of such a database is to be exploited. Unless linguistic annotation is automated to a considerable degree to enable automated data extraction, the scope of such databases will remain unexploited. Faced with this challenge, the obvious step was to bring computational linguistics on board to investigate how Natural Language Processing tools and technology can be applied to the task of learner English, a challenging enterprise in itself given that NLP technology has not been widely applied to L2 data. Finally, a web-based interface that supports data export and provides a search tool was built to maximise the accessibility and usability of the database.

A diverse team of linguists and computational linguists was, thus, put together to build a database of written data from *Englishtown*. The work has been sponsored by the Isaac Newton Trust, Trinity College, Cambridge and EF-Education First. EF has further provided vital logistical support in data collection and transfer, obtaining consent from EF-students and additional meta-data where needed. Last, but not least, EF has championed an open access research resource for the study of learner English.

We would like to thank a number of people that have made this project possible. From the EF side: Chris McCormick for his critical ideas in shaping this project at early stages and his co-ordinating work to take this project off ground; Eric Azumi and his team for their work on logistical and technical aspects of data collection and transfer; Yerrie Kim for all her co-ordinating work at various stages of the project. From Cambridge University: Henriëtte Hendriks for her continuing support and guidance as PI of the EF Research Unit and Head of the Dept of Theoretical and Applied Linguistics. John Hawkins for supporting this project at its early stages as director of the Research Centre for English and Applied Linguistics, the first home of the EF-Research Unit. We are also thankful to colleagues and students at the Department of Theoretical and Applied Linguistics, who have provided advice and feedback on various aspects of the project: Teresa Parodi, Norbert Vanek, Amy Hsieh, Akira Murakami and Maria Kunevich. Many thanks also to Ted Krawec for vital advice on providing a suitable user agreement and overseeing legal elements of the project. We are also grateful to Louise Balshaw for her administrative support.

Finally, we would like to thank the research team of this project: Anna Korhonen has guided the computational strand of this research. Jeroen Geertzen has worked tirelessly and creatively to transform some millions of unstructured data into a structured, annotated and accessible research resource and has overseen the application and evaluation of NLP tools on learner data. Brechtje Post has worked with us to link the written corpus with the development of the speech corpus. Our external consultant Detmar Meurers, from the University of Tübingen has generously discussed the project with us, sharing ideas on the challenges of automated linguistic annotation for learner data and questions of building large scale databases. Sichu Jiang has extended our originally rudimentary web-based interface with critical functionality during a student internship at DTAL over Summer 2012. Finally, we would like to thank Dimitris Michelioudakis and Toby Hudson for providing manual annotations to sample parsed data and Caroline Williams for technical support at early stages of data collection.

Dora Alexopoulou and Rachel Baker

2 Database Structure

EFCAMDAT consists of essays submitted to *Englishtown*, the online school of EF Education First, by language learners all over the world (Education First, 2012). A full course in Englishtown spans 16 proficiency levels aligned with common standards such as TOEFL, IELTS and the Common European Framework of Reference for languages (CEFR) as shown in Table 1.

Table 1: Englishtown skill levels in relation (indicative) to common standards

Englishtown	1-3	4-6	7-9	10-12	13-15	16
Cambridge Esol	-	KET	PET	FCE	CAE	-
IELTS	-	<3	4-5	5-6	6-7	>7
TOEFL iBT	-	-	57-86	87-109	110-120	-
TOEIC Listening & Reading	120-220	225-545	550-780	785-940	945	-
TOEIC Speaking & Writing	40-70	80-110	120-140	150-190	200	-
CEFR	A1	A2	B1	B2	C1	C2

Learners are allocated to proficiency levels after a placement test when they start a course at EF¹ or through successful progression through coursework. Each of the 16 levels contains eight lessons, offering a variety of receptive and productive tasks. EFCAMDAT consists of scripts of writing tasks at the end of each lesson on topics like those listed in Table 2.

Table 2: Examples of essay topics at various levels. Level and unit number are separated by a colon.

ID	Essay topic	ID	Essay topic
1:1	Introducing yourself by email	7:1	Giving instructions to play a game
1:3	Writing an online profile	8:2	Reviewing a song for a website
2:1	Describing your favourite day	9:7	Writing an apology email
2:6	Telling someone what you're doing	11:1	Writing a movie review
2:8	Describing your family's eating habits	12:1	Turning down an invitation
3:1	Replying to a new penpal	13:4	Giving advice about budgeting
4:1	Writing about what you do	15:1	Covering a news story
6:4	Writing a resume	16:8	Researching a legendary creature

Given 16 proficiency levels and 8 units per level a learner who starts at the first level and completes all 16 proficiency levels would produce 128 different essays. Essays are

¹Starting students are placed at the first level of each stage, 1, 4, 7, 10, 13, or 16.

graded by language teachers; learners may only proceed to the next level upon receiving a passing grade. Teachers provide feedback to learners using a basic set of error markup tags or through free comments on students’ writing. Currently, EFCAMDAT contains teacher feedback for 36% of scripts.

The data collected for the first release of EFCAMDAT contain 551,036 scripts (with 2,897,788 sentences, and 32,980,407 word tokens) written by 84,864 learners. We currently have no information on the L1 backgrounds of learners.² Information on nationality is, thus, used as the closest approximation to L1 background. EFCAMDAT contains data from learners from 172 nationalities. Table 3 shows the spread of scripts across the nationalities with most learners.³

Table 3: Percentage and number of scripts per nationality of learners

Nationality	Percentage of scripts	Number of Scripts
Brazilians	36.9%	187,286
Chinese	18.7%	96,843
Russians	8.5%	44,187
Mexicans	7.9%	41,115
Germans	5.6%	29,192
French	4.3%	22,146
Italians	4.0%	20,934
Saudi Arabians	3.3%	16,858
Taiwanese	2.6%	13,596
Japanese	2.1%	10,672

Few learners complete all of the proficiency levels. For many, their start or end of interacting with Englishtown fell outside the scope of the data collection period for the first release of EFCAMDAT. More generally, many learners only complete portions of the program. Nevertheless, around a third of learners (around 28K) have completed 3 full levels, corresponding to a minimum of 24 scripts.⁴ Only 500 learners have completed every unit from level 1 to 6 (accounting for at least 48 scripts).

Characterizing scripts quantitatively is difficult, because of the variation across topics and proficiency levels. Texts range from a list of words or a few short sentences to short narratives or articles. As learners become more proficient they tend to produce longer scripts. On average, scripts count 7 sentences (SD=3.8). Sample scripts are shown in Figure 1.

²Metadata on the L1 background of learners is being collected for the second release of the database.

³Of the 172 nationalities, 28 have over 100 learners, and 38 nationalities over 50 learners.

⁴If learners don’t receive a satisfying score on their writing, they may repeat the task, which means that a learner will have a minimum of 8 scripts per level but may have more if they repeat the task.

1. LEARNER 18445817, LEVEL 1, UNIT 1, CHINESE

Hi! Anna,How are you? Thank you to sendmail to me. My name's Anfeng.I'm 24 years old.Nice to meet you !I think we are friends already,I hope we can learn english togther! Bye! Anfeng.

2. LEARNER 19054879, LEVEL 2, UNIT 1, FRENCH

Hi, my name's Xavier. My favorite days is saturday. I get up at 9 o'clock. I have a breakfast, I have a shower... Then, I goes to the market. In the afternoon, I play music or go by bicycle. I like sunday. And you ?

3. LEARNER 19054879, LEVEL 8, UNIT 2, BRAZILIAN

Home Improvement is a pleasant protest song sung by Josh Woodward. It's a simple but realistic song that analyzes how rapid changes in a town affects the lives of many people in the name of progress. The high bitter-sweet voice of the singer, the smooth guitar along with the high pitched resonant drum sound like a moan recalling the past or an ode to the previous town lifestyle and a protest to the negative aspects this new prosperous city brought. I really enjoyed this song.

Figure 1: Three typical scripts, in which learners are asked to introduce themselves (1), describe their favourite day (2), and review a song for a website (3).

The data have been annotated with parts of speech tags and information on grammatical dependencies using the The Penn Treebank Tagset (Marcus et al., 1993) and a freely available state-of-the-art parser (Stanford parser; Klein and Manning (2003)). (Jeroen Geertzen and Korhonen, 2013) provide a detailed discussion of the linguistic annotations accompanying EFCAMDAT scripts. A substantial part of the data comes with error corrections that have been provided by teachers or correctors (See Appendix). The purpose of these corrections was to provide feedback to learners. Even though these corrections may be useful, there is no information on how reliably and systematically they have been marked by correctors.

3 The EFCAMDAT web-based interface

The database can be accessed through a web-based interface at <http://corpus.mml.cam.ac.uk/efcamdat>. Figure 2 shows the introductory page of the interface. The page lists the 16 teaching levels of EF and their alignment with Cambridge, IELTS and TOEFL exams. A list of tabs on the left gives the user a number of options.

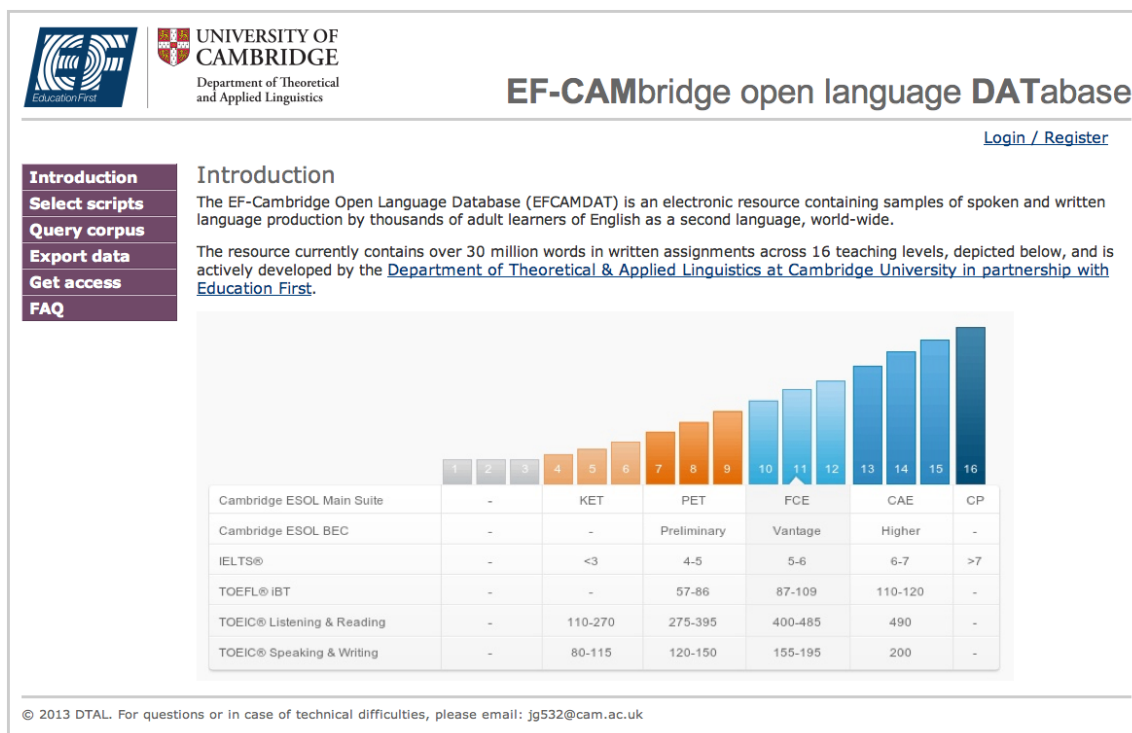


Figure 2: Introduction page of the EFCAMDAT interface

‘Get access’

The ‘Get access’ page is shown in Figure 3. Users need to download and read the End User License Agreement and accept its terms and conditions. Access to EFCAMDAT is free of charge but it is restricted to the academic research community. The user agreement sets out standard conditions protecting copyright. Users who agree to these conditions can provide some personal data and obtain access to EFCAMDAT.

UNIVERSITY OF CAMBRIDGE
Department of Theoretical and Applied Linguistics

EF-CAMbridge open language DATabase

[Login / Register](#)

Introduction
Select scripts
Query corpus
Export data
Get access
FAQ

Login

Email:

Password:

[Forgotten your password?](#)

Register

When registering for access to EFCamDat, you will be asked to agree to the end-user license and register your name, e-mail address and affiliation. Once this procedure is completed, you can access and query the corpus using your user name and password.

Please read the [End-user License Agreement](#) carefully, and select "I agree" to accept all its terms. You must accept this agreement in order to be able to use the EFCamDat Corpus.

Full name: *

Email: *

Institute:

Password: *

Confirm Password:


☐ I have read and agree with the [End-user License Agreement](#) *

© 2013 DTAL. For questions or in case of technical difficulties, please email: jg532@cam.ac.uk

Figure 3: Access to EFCAMDAT

‘Select scripts’

Figure 4 shows the page that allows users to select scripts for queries or export. Scripts can be selected from the 16 different EF levels. In Figure 4 we have selected Level 4. Each level has 8 units involving a unique writing topic. Thus, scripts are spread across different topics, e.g. *What do you do?*, *Daily routines* etc. In Figure 4 we have selected scripts from just one the topic *what does she look like?*. We have also selected scripts from that level and that topic written by Russian learners. The numbers in brackets indicate the number of scripts for the specific choice. For instance, there are 7,101 scripts from Russians at Level 4, of which 752 are on the selected topic, *What does she look like?*. The top box provides a summary of the selection.



UNIVERSITY OF
CAMBRIDGE
Department of Theoretical
and Applied Linguistics

EF-CAMbridge open language DATAbase

Jeroen Geertzen [Log out](#)

[Introduction](#)
[Select scripts](#)
[Query corpus](#)
[Export data](#)
[Get access](#)
[FAQ](#)
[Admin](#)

Select scripts

If you are interested in querying or exporting scripts from specific levels, activities, or learner backgrounds, please specify this below. Numbers in parentheses indicate the number of scripts available.

The current selection contains **752** scripts (**±58908** words) from:

- 1711 learners
- 1 nationalities
- 1 unit(s) from level(s): 4

☐ scripts only from learners who have completed all the selected units

Teaching levels and units

--Please select levels--
--Select all (43049)--
Level 1 (10547)
Level 2 (4783)
Level 3 (3878)
Level 4 (7101)
Level 5 (3505)
Level 6 (2236)
Level 7 (4531)
Level 8 (4682)

--Please select units--
--Select all--
Level 4
--Select whole level--
What do you do? (1736)
Daily routines (1235)
Having a party (912)
Sports and hobbies (772)
What does she look like? (752)
Talking about past events (637)

--Selected units--
Level 4
--Select whole level--
What does she look like? (752)

AddRemove Clear

Learner nationalities

--Please select continents--
--Select all (8912)--
Africa (45)
Asia (2629)
Europe (2184)
North America (908)
South America (3137)
Oceania (9)

--Please select countries--
--Select all--
Europe
--Select whole continent--
Russia (752)
Germany (466)
Italy (412)
France (382)
United Kingdom (22)
Spain (147)

--Selected countries--
Europe
--Select whole continent--
Russia (752)

AddRemove Clear

© 2013 DTAL. For questions or in case of technical difficulties, please email: jg532@cam.ac.uk

Figure 4: Selection of EFCAMDAT scripts according to teaching level, lesson and learner nationality

‘Query corpus’

The corpus can be queried by providing word patterns within the set of scripts that has been selected. Figure 5 shows the page that allows users to select an example or already entered query, or construct a novel query and look for sentences that match the pattern.

UNIVERSITY OF CAMBRIDGE
Department of Theoretical and Applied Linguistics

EF-CAMbridge open language DATAbase

Jeroen Geertzen [Log out](#)

- Introduction
- Select scripts
- Query corpus
- Export data
- Get access
- FAQ
- Admin

Query the corpus

Retrieve instances of patterns of interest, consisting of words, part-of-speech, and dependency syntax.

The current selection contains **752** scripts (± 58908 words) from:

- 1711 learners
- 1 nationalities
- 1 unit(s) from level(s): 4
- ☐ scripts only from learners who have completed all the selected units

Search for a pattern

Select a query:

Or specify:

Add PoS:

Add GR:

Display:

returned 234 matches, showing 1-234 :

- 01 IPRP started VBD to TO study VBI inIN 1995 CD atIN elementary JJ school IN J.
- 02 IPRP went VBD to TO school VB when IN was VBD six CD years NNS old JJ J.
- 03 IPRP started VBD to TO draw VB and IN IPRP went VBD to TO art NN school IN IN 1998 CD J.
- 04 After IN IPRP graduated VBD J. IPRP wanted VBD to TO travel VB J.
- 05 IPRP went VBD to TO school VB inIN 1986 CD J.
- 06 After IN wedding VBG we IPRP went VBD to TO travel VB J.
- 07 After IN IPRP graduated VBD IPRP wanted VBD to TO travel VB J.
- 08 I went to graduate study in 2003 as a teacher.

Sentence ID: U14296:4
Sentence length: 10 tokens
Teaching level: 4
Learner nationality: Russia
[\[show tree\]](#)

© 2013 DTAL. For questions or in case of technical difficulties, please email: jg532@cam.ac.uk

Figure 5: Query page

Word patterns can be specified as follows:

1. Each word specification is enclosed with brackets
2. A word can be specified by means of a word token (e.g. `[word="cars"]`), a lemma (e.g. `[lemma="car"]`), a part-of-speech tag (e.g. `[pos="NNS"]`), its relation to the head it attaches to (e.g. `[dg-rel="nsubj"]`), and properties of the head (`dg-hword` / `dg-hlemma` / `dg-hpos`)
3. A word specification may contain multiple properties. For instance, a plural noun that attaches as a direct object to a past-tense verb can be specified as:
`[(pos="NNS") & (dg-rel="dobj") & (dg-hpos="VBD")]`.

4. Properties may also be underspecified, using `.*`. For instance, any noun that attaches as a direct object to any verb can be specified as:

```
[(pos="N.*") & (dg-rel="dobj") & (dg-hpos="V.*")]
```

5. Word gaps in patterns can be indicated with empty brackets `[]`, such that the pattern `[word="to"] []{1,3} [word="for"]` will match the word token "to", followed by at least one and at most three words, followed by the word token "for".

Construction of the patterns is aided by drop-down list boxes for the available part-of-speech tags and grammatical relations. Results can be visualised in plain sentences or in part-of-speech tag annotated sentences, and meta-information is provided, such as teaching level and learner nationality. A dependency tree can be visualised upon request.

‘Export data’

Scripts that have been selected, or sentences that have been queried, can be exported to XML files. Figure 6 shows the page that allows users to select what unit of interest to export (scripts or queried sentences), what information should be included (raw script text, syntactic annotations, or error corrections), and whether to compress the resulting XML file.

The screenshot displays the 'Export corpus data' interface of the EF-CAMbridge open language DATAbase. The page header includes the University of Cambridge logo and the title 'EF-CAMbridge open language DATAbase'. A user profile for 'Jeroen Geertzen' is visible in the top right corner. The left sidebar contains navigation links: 'Introduction', 'Select scripts', 'Query corpus', 'Export data' (highlighted), 'Get access', 'FAQ', and 'Admin'. The main content area is titled 'Export corpus data' and includes a sub-header 'Export XML-structured corpus data according to selected criteria'. A summary box states: 'The current selection contains 752 scripts (±58908 words) from: 1711 learners, 1 nationalities, 1 unit(s) from level(s): 4, and a checkbox for 'scripts only from learners who have completed all the selected units'. Below this, the 'Specify what to export' section has three subsections: 'Unit of interest' with radio buttons for 'Selected scripts' (selected) and 'Last queried sentences'; 'Information included' with checked checkboxes for 'Raw script text', 'Syntactic annotations', and 'Error corrections'; and 'Export format' with radio buttons for 'XML compressed (zipped)' and 'XML uncompressed' (selected). An 'Export data' button is located below these options. At the bottom, a download link is provided: 'Download: EF20130315_selection55.xml (6.1 MB)'. The footer contains copyright information: '© 2013 DTAL. For questions or in case of technical difficulties, please email: jg532@cam.ac.uk'.

Figure 6: Export

It is generally recommended to choose to download zip compressed XML, unless the selection is rather small. Zipped XML files, depending on the selection, may range from tens of KB to about 500MB at most.

The XML data contains a header with information about the corpus version, the selection of levels and nationalities, followed by either scripts or sentences and provided with information according to the requested information to be included (see Figure 7). The XML structure of a full script with all available information is exemplified in Figure 8.

The screenshot displays the EF-CAMbridge open language DATabase web interface. On the left is a navigation menu with links: Introduction, Select scripts, Query corpus, Export data, Get access, FAQ, and Admin. The main content area is titled 'Export corpus data' and shows a preview of XML data for a selection. The XML structure includes a <meta> block with fields like title, version, url, key, user, date, nationalities, and units. It also contains a <writings> block with a <writing> element that includes learner, nationality, topic, date, grade, and text. The text field contains a paragraph in Russian about a person named Natalia. At the bottom of the interface, there is a copyright notice: '© 2013 DTAL. For questions or in case of technical difficulties, please email: jg532@cam.ac.uk'.

Figure 7: XML

Frequently Asked Questions

The 'FAQ' page provides further information on how to use the corpus interface and contains documents with information on EFCAMDAT. We ask users to cite the following paper when using EFCAMDAT:

J. Geertzen, T. Alexopoulou, A. Korhonen, (2013) Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT) in Proceedings of the 31st Second Language Research Forum (SLRF), Carnegie Mellon, Cascadia Press.

References

Education First (2012). Englishtown. <http://www.englishtown.com/>.

Jeroen Geertzen, T. A. and Korhonen, A. (2013). Automatic linguistic annotation of large scale L2 databases: The ef-cambridge open language database (efcamdat). In *Selected*

```

<writing id="U661262" level="1" unit="6">
  <learner id="22782633" nationality="eg"/>
  <topic id="4478">Signing up for a dating website</topic>
  <date>2012-07-01 16:38:03.920</date>
  <grade>94</grade>
  <text>
    Hi!<br>My name's Mustafa, <change><selection>i'm</selection><tag><symbol>SP</symbol><correct>I'm</co
    rrect></tag></change> 27.<br>I'm short and slim. I have short<change><selection></selection><tag><symbol>PU</sym
    bol><correct></correct></tag></change> black hair and brown eyes.<br><br><change><selection>bye</selection><ta
    g><symbol>C</symbol><correct>Bye</correct></tag></change>!<br><br>Mustafa
  </text>
  <annotation id="syn1">
    <sentences count="3">
      <sentence id="2" tokencount="9">
        <token id="1" lemma="my" pos="PRP$" head="2" rel="poss">My</token>
        <token id="2" lemma="name" pos="NN" head="4" rel="nsubj">name</token>
        <token id="3" lemma="s" pos="POS" head="4" rel="cop">'s</token>
        <token id="4" lemma="Mustafa" pos="NNP" head="8" rel="ccomp">Mustafa</token>
        <token id="5" lemma="," pos="," head="8" rel=",">,</token>
        <token id="6" lemma="i" pos="RB" head="8" rel="nsubj">i</token>
        <token id="7" lemma="m" pos="VBP" head="8" rel="cop">'m</token>
        <token id="8" lemma="27" pos="CD" head="0" rel="root">27</token>
        <token id="9" lemma="." pos="." head="8" rel=",">.</token>
      </sentence>
      <sentence id="3" tokencount="6">
        <token id="1" lemma="I" pos="PRP" head="3" rel="nsubj">I</token>
        <token id="2" lemma="m" pos="VBP" head="3" rel="cop">'m</token>
        <token id="3" lemma="short" pos="JJ" head="0" rel="root">short</token>
        <token id="4" lemma="and" pos="CC" head="5" rel=",">and</token>
        <token id="5" lemma="slim" pos="JJ" head="3" rel="conj_and">slim</token>
        <token id="6" lemma="." pos="." head="5" rel=",">.</token>
      </sentence>
      <sentence id="4" tokencount="9">
        <token id="1" lemma="I" pos="PRP" head="2" rel="nsubj">I</token>
        <token id="2" lemma="have" pos="VBP" head="0" rel="root">have</token>
        <token id="3" lemma="short" pos="JJ" head="5" rel="amod">short</token>
        <token id="4" lemma="black" pos="JJ" head="5" rel="amod">black</token>
        <token id="5" lemma="hair" pos="NN" head="2" rel="dobj">hair</token>
        <token id="6" lemma="and" pos="CC" head="8" rel=",">and</token>
        <token id="7" lemma="brown" pos="JJ" head="8" rel="amod">brown</token>
        <token id="8" lemma="eye" pos="NNS" head="5" rel="conj_and">eyes</token>
        <token id="9" lemma="." pos="." head="8" rel=",">.</token>
      </sentence>
    </sentences>
  </annotation>
</writing>

```

Figure 8: XML

Proceedings of the 2012 Second Language Research Forum, Somerville, MA, USA. Cascadilla Proceedings Project.

Klein, D. and Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, pages 423–430, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Appendix: Error codes

Code	Meaning
$x \gg y$	change from x to y
AG	agreement
AR	article
CO	combine sentences
C	capitalization
D	delete
EX	expression of idiom
HL	highlight
I(x)	insert x
MW	missing word
NS	new sentence
NWS	no such word
PH	phraseology
PL	plural
PO	possessive
PR	preposition
PS	part of speech
PU	punctuation
SI	singular
SP	spelling
VT	verb tense
WC	word choice
WO	word order