

Отчет об экспертизе usability лингвистического ресурса

Corpus of Modern Scottish Writing

<http://www.scottishcorpus.ac.uk/cmsw>

Выполнен студентками 2 группы (БКЛ-171)

Моховой Анной (ann.mokhova@gmail.com)

Петуховой Анной (aapetukhova_2@edu.hse.ru)

Сигдел Элиной (sigdelina@gmail.com)

Введение

Для описания лингвистического ресурса мы выбрали Corpus of Modern Scottish Writing (CMSW) - корпус письменных текстов 1700-1945 годов шотландского варианта английского языка, созданный в университете Глазго. Дополняет Helsinki Corpus of Older Scots (1450-1700) и the Scottish Corpus of Texts & Speech (1945-...). Корпус содержит более 350 документов, около 5,5 млн слов. Источники: библиотека Митчелла в Глазго, Национальная библиотека Шотландии и архив университета Глазго.

Дизайн

Дизайн корпуса шотландских текстов достаточно прост.

Сайт сделан в фиолетовом цвете, который окружает белую вставку с текстом. В ней содержится информация о корпусе (независимо от того, на какой странице сайта находится пользователь).

Текст выделяется на фоне фиолетового белым цветом, подчеркивая заголовки и сочетаясь с фоном информации в центре страницы. В целом, материал расположен очень логично: структурирован на страницах сайта с помощью альбомной ориентации, это позволяет не потерять важное, не прокручивая страницу вниз. К тому же, в верхней части сайта присутствуют функциональные элементы корпуса, а в нижней – ссылки на почтовый адрес, современный корпус Scottish Corpus of Texts & Speech, эмблему и информацию об университете Глазго.

Результаты выдачи представлены красиво: искомые словоформы ярко выделяются на фоне самих текстов. Кроме того, каждый текст подсвечивается отдельными не слишком контрастными цветами, что концентрирует внимание на определенном авторе, дате создания, частоте, годе создания и так далее.

Все ссылки подсвечены, позволяют без проблем найти их и перейти на необходимый ресурс. Таким образом, в дизайне не хочется ничего менять, он очень приятный по цветовой гамме, предоставляет подсветку необходимой информации, обеспечивает комфортную работу пользователю. Минусов в дизайне замечено не было.

Onboarding глазами новичка

Данный ресурс несложно найти: если не знать названия, ссылка на основной корпус или Scottish Corpus of Texts & Speech представлена в Wikipedia, если же знать название, то сайт будет первым среди запрошенных.

Будучи новичком, можно разобраться в работе и структуре корпуса, не затрачивая при этом много усилий. Необходимо лишь знать английский язык на базовом уровне, чтобы найти необходимую информацию. Все интуитивно понятно, легко, без проблем выдается короткий и важный для работы с корпусом материал. Например, в Search the Corpus кроме поискового окна, в которое новичок может вписать нужные ему слова, документы и так далее, есть материал о структуре ресурса. Благодаря альбомной ориентации не надо искать или переходить по ссылкам, чтобы узнать, сколько документов и вхождений зафиксировано, - все это и похожая информация представлена в правой части страницы.

Поиск словоформ настолько очевиден, насколько интуитивно понятно само устройство сайта. Тот факт, что все необходимые условия для поиска представлены в одном блоке, облегчает работу и не вызывает трудностей. Здесь без проблем можно либо ввести слово, либо задать условия для поиска не по конкретным словам.

Помощь

Отдельного раздела “Помощь” на сайте CMSW нет. Образцы с поисковыми запросами для новичков также не представлены. Несмотря на это, дизайн и структура сайта устроены так, что даже у пользователя, который до этого не работал с лингвистическими корпусами, не возникнет затруднений с использованием этого ресурса. Немного вводит в заблуждение тот факт, что на сайте сестринского корпуса The Scottish Corpus of Texts & Speech (SCOTS) такой раздел есть, но не все инструкции по работе подходят и для корпуса CMSW. Мы бы предложили разработчикам сайта добавить раздел “Помощь” с подробным объяснением возможностей корпуса.

Функционал

Доступность

Главными плюсами корпуса являются бесплатный открытый доступ к нему как онлайн, так и оффлайн, и отсутствие необходимости регистрации.

Типы текстов

В корпусе содержатся как напечатанные, так и рукописные тексты: письма, книги, документы, статьи. Представленные жанры: личные письма; административная, экспозиционная, художественная, учебная, религиозная литература; стихотворения/драмы; журналистика; Orthoepists.

Слой разметки

Corpus of Modern Scottish Writing имеет только один слой разметки - метатекстовый. Это делает CMSW менее удобным и функциональным по сравнению с другими языковыми корпусами, которые имеют не только метатекстовую, но и морфологическую, синтаксическую и другие. Таким образом, поиск в корпусе возможен только по точной форме слова и параметрам текста.

Интерактивность

Отличительной особенностью корпуса является наличие медиатеки, где представлены фотографии каждой страницы текста, идущие по порядку, что дает возможность посмотреть на рукописи 18 века.

Можно найти подробную информацию о тексте и авторе: идентификационный номер текста, заголовок, промежуток времени создания, жанр, год публикации, количество слов в документе, идентификационный номер автора, имя, фамилия, пол, год рождения, место рождения, места рождения отца и матери, профессия, профессии отца и матери, образование, место жительства.

Текст можно посмотреть в формате txt и загрузить.

Возможности поиска

В корпусе представлены две опции поиска: быстрый и расширенный. Быстрый поиск находится в правом верхнем углу страницы и доступен при просмотре любой вкладки. С помощью быстрого поиска можно найти только точную форму слова.

Расширенный поиск находится во вкладке “search” и выдает запросы по следующим параметрам:

- Слово
- Автор (имя, пол)
- Текст (название, жанр, период создания, год публикации: с, по).

Такой поиск сильно ограничен вследствие наличия только одного уровня разметки.

Примеры поисковых запросов

1) Поиск по автору.

Зададим имя автора - Burns - и пол - male (мужской):

Поиск выдал:

Информация о документе:

Poems, Chiefly in the Scottish Dialect

Author(s): Burns, Robert

Text

THE Simple Bard, unbroke to
He pours the wild effusions
And if inspir'd, 'tis Nature's
Her's all the melting thrills
ANONYMOUS
KELMARNOCK:
PRINTED BY JOHN WILSON,
M.DCCCLXXXVI
Entered in Stationers-hall.
PREFACE.
The following trifles are not
of the Poet, who, with all the
of learned art, and perhaps
and idleness of upper life,
a rural theme, with an eye to
Virgil. To the Author of this
celebrated names their com

Poems, Chiefly in the Scottish Dialect

Document Information

Document ID	108
Title	Poems, Chiefly in the Scottish Dialect
Year group	1750-1800
Genre	Verse/drama
Year of publication	1786
Wordcount	27458

Author information: Burns, Robert

Author ID	215
Forenames	Robert
Surname	Burns
Gender	Male
Year of birth	1759
Place of birth	Alloway, Ayrshire, Scotland
Occupation	Farmer, author, exciseman
Father's occupation	Farmer
Education	Little formal schooling
Locations where resident	Alloway, Scotland

2) Быстрый поиск.
Введем слово - existence:

Corpus of
Modern
Scottish
Writing

existence

Go

Home Search Browse Microsites About

Поиск выдал:

385 occurrences of **existence** (0.01% of corpus tokens)
52 documents matching criteria (14.61% of corpus)
3,706,558 total words in matching documents (68.20% of corpus)

Showing documents 1 - 50 of 52
Page: < prev next > **1** 2 | Results per page: 20 **50** 100 1000 All | Max word occurrences per document: 5

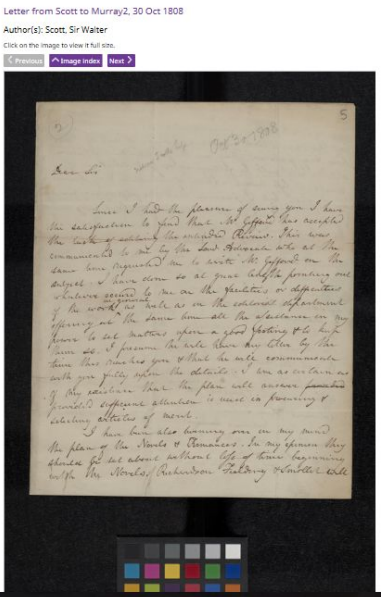
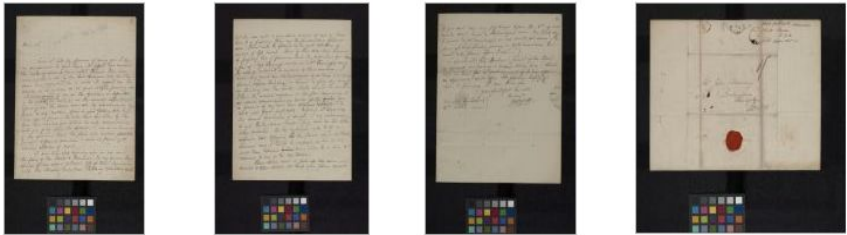
Title	Author	Year	Word count	Freq.	Norm.
An Essay on the Scoto-English Dialect	Collin, Zacharias	1862	35,322	3	0.08
... tongue. So the interlineary glosses came into existence 2). Fortunately for us, four, or at least...					
... bian, have, in modern English, both lost their existence in their own right, and coalesced to one ...					
.... Vir. p. 49, 12. These may suffice to prove the existence of this form during the mentioned time. — ...					
A New Revelation	Doyle, Arthur Conan	1918	20,284	10	0.49
... my finite brain could get no further than its existence . Right and wrong I saw also as great obvious...					
... our duty towards humanity during our transitory existence . This was my frame of mind when Spiritual...					
..., just as it would be no argument against the existence of a musician if you tampered with his violin...					
... arguing about some impersonal thing such as the existence of Atlantis or the Baconian controversy. ...					
... few years. If no other spiritual books were in existence than five which have appeared in the last...					
Anderson, Isabel					

Выберем произвольный документ:
Фотографии текста:

Letter from Scott to Murray2, 30 Oct 1808

Author(s): Scott, Sir Walter

[Return to text](#)



3) Поиск по жанрам и периоду.
Выберем жанры - Religious prose, Imaginative prose - и период - 1800-1850:

Search the Corpus

Reset search boxes

Word

Author

Surname

Gender

- All

Document

Title

Genre (s)

☒ Religious prose

☐ Verse/drama

☒ Imaginative prose

☐ Journalism

☐ Orthoeipists

Year group (s)

☐ 1700-1750

☐ 1750-1800

☒ 1800-1850

☐ 1850-1900

☐ 1900-1950

Year published from

Year published to

Search

356 documents in the corpus

147,129 total types

5,434,830 total tokens

The content of CMSW is mainly written texts. Documents are selected on the basis of date and genre. Some texts have been chosen as 'written records of speech', e.g. minutes of meetings and transcripts of court proceedings. Other genres include personal writing, expository prose, verse/drama, journalism, and writing by orthoeipists or commentators on language.

You can download the entire corpus as plain text.

Поиск выдал:

Найдем слово design в тексте с помощью поля highlight:

Private Memoirs and Confessions of a Justified Sinner

Author(s): Hogg, James

Text

THE PRIVATE MEMOIRS

WRITTEN BY HIMSELF:

WITH A DETAIL OF CURIOUS TRADITIONAL FACTS, AND
LONDON:

PRINTED FOR LONGMAN, HURST, REES, ORME, BROWN,
MDCCCXIV.

EDINBURGH:

PRINTED BY JAMES CLARKE AND CO.

1824.

TO

THIS WORK IS RESPECTFULLY INSCRIBED,

THE

IT appears from tradition, as well as some parish registers
still extant, that the lands of Dalcastell (or
Dalchastell, as it is often spelled) were possessed by a
family of the name of Colwan, about one hundred and
fifty years ago, and for at least a century previous to
that period. That family was supposed to have been a
branch of the ancient family of Colquhoun, and it is
certain that from it spring the Cowans that spread
through the Border. I find, that in the year 1687

"Then, thanks be to Him whose name you have
profaned, I am none. If one of the party be a
gentleman, I do hope in God I am not!"

singular favour, they wisely restrained one another
from inflicting the punishment that each of them

It was now apparent to them all that he was courting
obloquy and manual chastisement from their
hands, if by any means he could provoke them to
the deed; and, apprehensive that he had some sinister
and deep-laid **design** in hunting after such a
yearned to bestow, personally, and which he so well
deserved.

But the unpopularity of the Younger George

Colwan could no longer be concealed from his associates.

It was manifested wherever the populace

were assembled; and his young and intimate friend,

Adam Gordon, was obliged to warn him of the circumstance,

that he might not be surprised at the

gentlemen of their acquaintance withdrawing themselves

from his society, as they could not be seen

Options:

- Information about document and author
- View 404 colour images
- View 404 greyscale images
- View plain text
- Download plain text
- Cite this document

Highlight word:

design

There are 5 matches.

< Previous Next >

Заключение

В целом, корпус понравился приятным дизайном, удобной поисковой системой и доступностью, наличием медиаданных. Понятно и легко ориентироваться на сайте. Однако CMSW представляет собой скорее архив с функцией нахождения слов, чем лингвистический корпус в привычном понимании. Важность корпуса все же велика, так как тексты были написаны в значимый период истории Шотландии: переход от литературы Эпохи Возрождения до литературы Нового времени, стандартизация письменного английского.