Use Case 2

Creating a linguistically annotated corpus of 19th century English novels

Background

The digital humanities provide a new conception of the world of literature. Not only is this world larger – the sheer volume of the material we can access is unprecedented – but it is open to levels of analysis that could never be achieved by human brainpower alone. Hierarchies and themes fade into the background as patterns and networks emerge. These methods simultaneously divide texts into new categories and connect them to each other to form new wholes. (Source: When computers read: Literary analysis and digital technology by Sarah Jones)

CLARIN Resources and services used

- Virtual Language Observatory (VLO)
- Language Resource Switchboard
- WebLicht

Teacher

How can my students create an annotated corpus of 19th century English novels from scratch in an easy-to-use online environment?

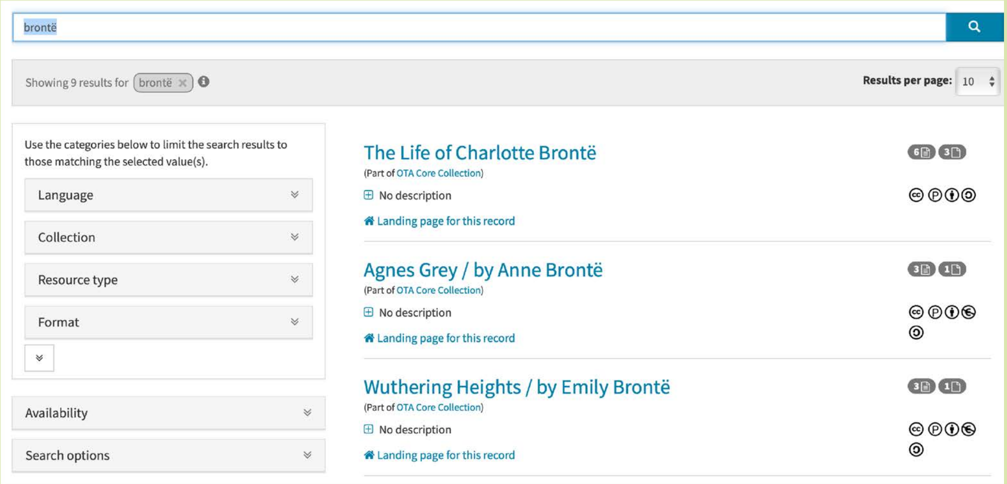
Researcher

Can you help me find resources and tools to research the stylistic differences between 19th century female and male novelists?

Step-by-step guide

The **Language Resource Switchboard (LRS)** aims at bridging the gap between resources (as identified in the VLO, Federal Content Search, and the CLARIN Virtual Collection) and tools that can process these resources in one way or another. For a given resource in question, it identifies all tools that can process the resource. It then sorts the tools in terms of the tasks they perform, and presents a task-oriented list to the user. Users can then select and invoke the tool of their choosing. (Source: Adapted from The Language Resource Switchboard by Claus Zinn)

1 Search the VLO with the simple query Brontë.



brontë

Showing 9 results for brontë ×

Results per page: 10

Use the categories below to limit the search results to those matching the selected value(s).

- Language
- Collection
- Resource type
- Format
- Availability
- Search options

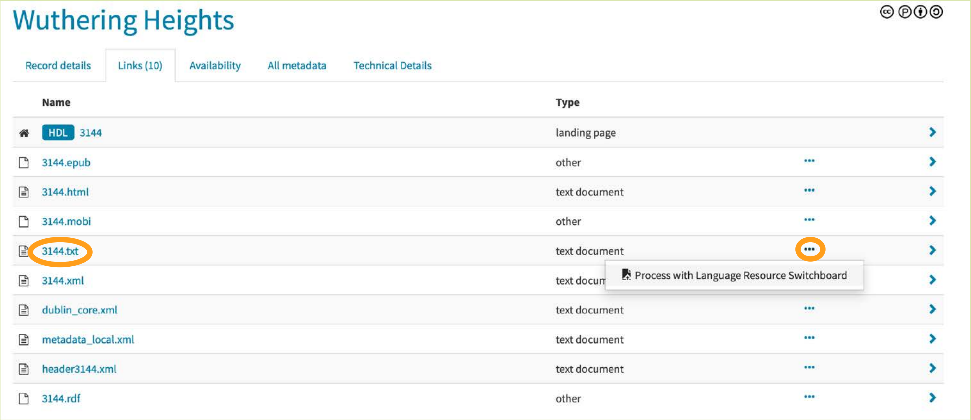
The Life of Charlotte Brontë
(Part of OTA Core Collection)
No description
Landing page for this record

Agnes Grey / by Anne Brontë
(Part of OTA Core Collection)
No description
Landing page for this record

Wuthering Heights / by Emily Brontë
(Part of OTA Core Collection)
No description
Landing page for this record

This query gives you VLO records for 19th century English novels by the Brontë sisters, such as Wuthering Heights by Emily Brontë, The Tenant of Wildfell Hall by Anne Brontë, and Jane Eyre by Charlotte Brontë.

2 In each VLO record under the "Links" tab, we can find the complete novels in the form of .txt files. Each file can be processed through the Language Resource Switchboard by clicking on "…" next to the .txt file (in our case, 3144.txt).



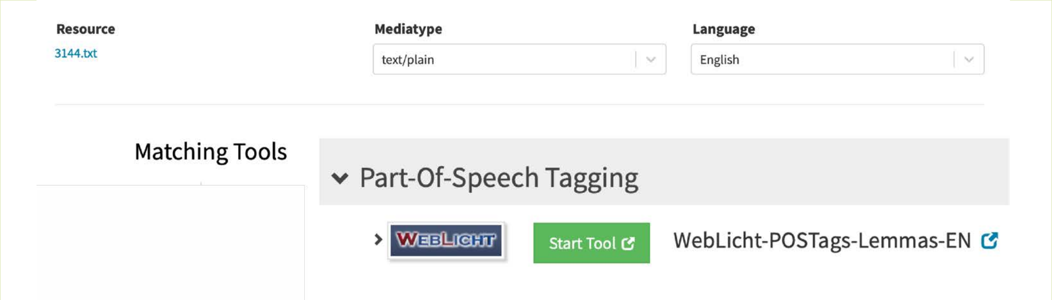
Wuthering Heights

Record details | Links (10) | Availability | All metadata | Technical Details

Name	Type
HDL 3144	landing page
3144.epub	other
3144.html	text document
3144.mobi	other
3144.txt	text document
3144.xml	text document
dublin_core.xml	text document
metadata_local.xml	text document
header3144.xml	text document
3144.rdf	other

Process with Language Resource Switchboard

3 The first step in linguistic annotation typically involves **part-of-speech tagging**, with which each word in a corpus is assigned a part of speech, like *noun, verb*, and *adjective*. In the LRS, we see that part-of-speech tagging is performed by WebLicht.



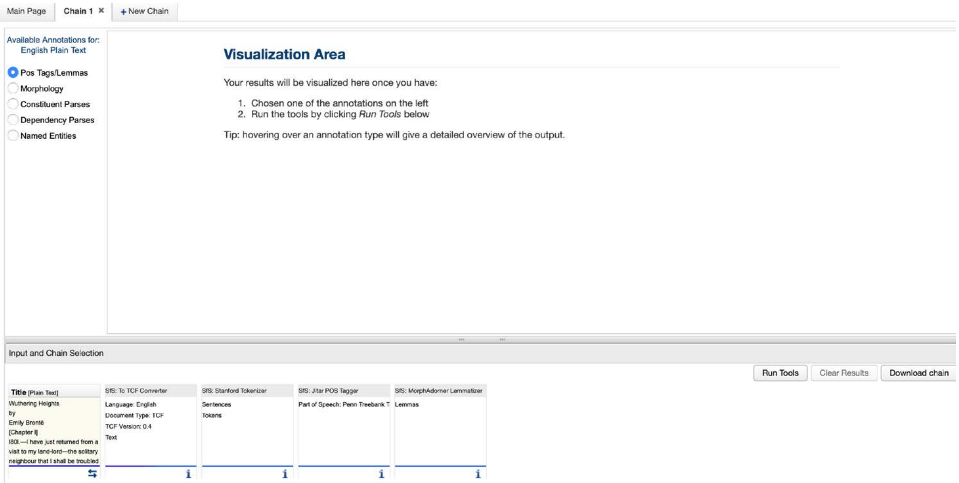
Resource	Mediatype	Language
3144.txt	text/plain	English

Matching Tools

Part-Of-Speech Tagging

WebLicht | Start Tool | WebLicht-POSTags-Lemmas-EN

4 In the WebLich application, we select *PoS tags/lemmas* under “Available Annotations for English Plain text”



Available Annotations for:
English Plain Text

● Pos Tags/Lemmas
○ Morphology
○ Constituent Parses
○ Dependency Parses
○ Named Entities

Visualization Area

Your results will be visualized here once you have:

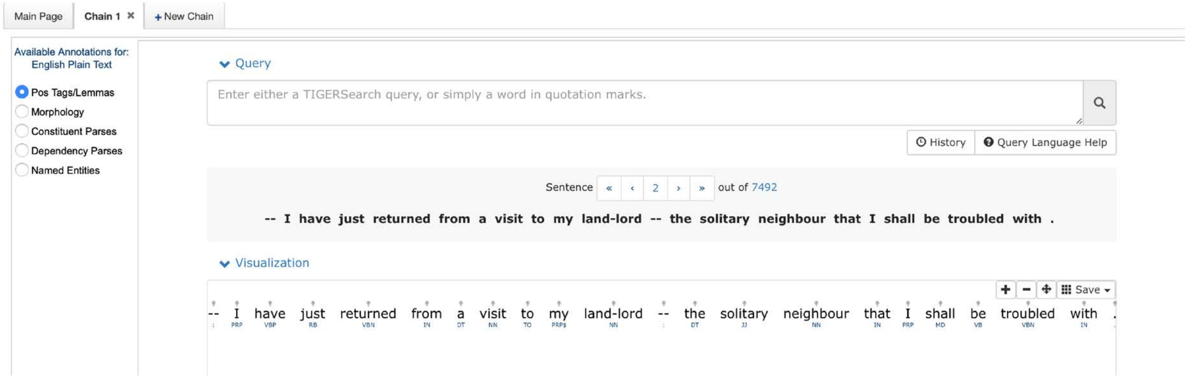
1. Chosen one of the annotations on the left
2. Run the tools by clicking *Run Tools* below

Tip: hovering over an annotation type will give a detailed overview of the output.

Input and Chain Selection

Title [Plain Text]	SfS: To TCF Converter	SfS: Stanford Tokenizer	SfS: Jitar POS Tagger	SfS: MorphAdorner Lemmatizer
Wuthering Heights by Emily Brontë [Chapter I] I80I.—I have just returned from a visit to my land-lord—the solitary neighbour that I shall be troubled	Language: English Document Type: TCF TCF Version: 0.4 Text	Sentences Tokens	Part of Speech: Penn Treebank T	Lemmas

5 After clicking on *Run Tools*, the entire *Wuthering Heights* novel becomes tagged for parts of speech.



Available Annotations for:
English Plain Text

● Pos Tags/Lemmas
○ Morphology
○ Constituent Parses
○ Dependency Parses
○ Named Entities

Enter either a TIGERSearch query, or simply a word in quotation marks.

Sentence « ‹ 2 › » out of 7492

-- I have just returned from a visit to my land-lord -- the solitary neighbour that I shall be troubled with .

--	I	have	just	returned	from	a	visit	to	my	land-lord	--	the	solitary	neighbour	that	I	shall	be	troubled	with	.
:	PRP	VBP	RB	VBN	IN	DT	NN	TO	PRP$	NN	:	DT	JJ	NN	IN	PRP	MD	VB	VBN	IN	.

The annotated novel can now be queried like a regular corpus either for simple words or by using the TIGERSearch corpus query language. To find all the adjectives in the newly tagged corpus, type [**pos** = /**J**.*/] in the Query field. Make sure to enclose the query in square brackets. Try visualizing the results. Which are the most and least common adjectives in the novel? Hint: In the Statistics visualisation under “Add/remove columns”, try adding the values PoS and lemma.

WebLicht also allows you to create additional annotation chains (“New Chain”). Try to repeat the task above by also tagging The Tenant of Wildfell Hall for **parts of speech** as well as for **named entities**. By creating several annotation chains in this way, you are able to create a full-fledged linguistically annotated corpus, consisting of several novels which were originally in simple plain text.

Research bite

By analysing a corpus of novels by Charles Dickens, Mahlberg et al. (2019) have studied how fictional dialogue is used by the author to create a sense of realism and authenticity. The authors have shown that Dickens consistently writes dialogue characterised by linguistic features that fictional and real people share (e.g., question fragments, set expressions conveying politeness and vagueness), which contributes to a sense of naturalness to speech in fiction. By contrast, the range of frequent word combinations in fictional dialogue is more limited than that in spoken fiction, so it is possible that literature adds an iconic or heightened meaningful effect onto these forms. Particularly, shorter lexical combinations (e.g. *I mean, you know* and *I don't know*) are less frequent in fiction. (Source: Speech-bundles in the 19th-century English novel by Michaela Malmberg, Viola Wiegand, Peter Stockwell, and Anthony Hennessey)