Manual for the
Linguistic Landscape Corpus
Version 1 (2021)

The contents of this manual are also available at the project website:
https://wou.edu/linguistic-landscape-corpus/

==========

About the LL Corpus

A resource for Linguistic Landscape scholars and students

Welcome
After years of preparation, the first version of the LL Corpus (2021) is available for use.

The LL Corpus is the full text of 383 published journal articles (1997-2017) and 165 book chapters (2008-2018) for a total of 548 items. The LL Corpus is a freely available resource intended for use by Linguistic Landscape scholars and students as well as corpus linguists. The CQPweb search interface enables users to perform anything from simple word searches to advanced corpus analysis. The LL Corpus does not infringe on copyright restrictions because the results of searches only display 200-words of text from any item. Links in the metadata for each item will direct users to the DOI or URL for the publication so that users can access the full text via institutional or individual methods.

Please see the FAQ Page for more detailed information.

==========

Sign up for access

Access to the Linguistic Landscape Corpus is only available through the Lancaster University CQPweb server. To register for a free CQPweb account follow these steps.

1. Please read the User License below.

> **Linguistic Landscape Corpus (LLC) User License**
>
> The LLC is a publicly-accessible resource. Anyone may freely access it for non-commercial research. However, many of the materials within the corpus **are not** in the public domain, with copyrights owned by various publishing companies and individuals. By registering for a CQPweb account and accessing the LLC, you are agreeing to the following conditions.
>
> * **You may** make use of the LLC only for non-commercial research and/or teaching.
> * **You may** publish the results of research that uses the corpus.
>
> … In any such publication, **you may** reproduce excerpts of the texts in the LLC only as permitted under US copyright law and fair use.
>
> … **You must** clearly identify any excerpt as originating from the LLC.

… **You must** acknowledge the use of the LLC in your research by citing the following standard reference (in your field's usual referencing style):

Troyer, Robert A. (2021). Linguistic Landscape Corpus. CQPweb at Lancaster. https://cqpweb.lancs.ac.uk/llscape202107/

*** We ask that** you inform us of any publications (see the Contact page) so that we can list them on our site.

*** You acknowledge** that the LLC is provided to you "as is", without any warranty or guarantee of utility for any specific purpose.

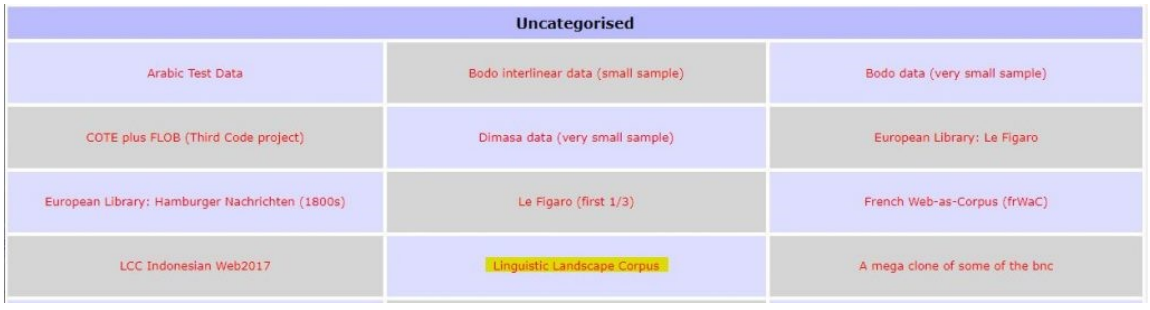
2. Create an account on cqpweb.lancs.ac.uk if you don't already have one.

- From the main page, click on **Create account**.

- Complete the form, submit it, and then click the link that will be emailed to you to validate your account.
- If you already have a CQPweb account, you do not need to create another one.

3. By registering for a CQPweb account and accessing the LLC, you affirm that you have read and understand the User License above, and that you agree to abide by its conditions.

4. Your CQPweb account will be granted access to the corpus immediately. From the CQPweb Main Page, scroll down to the "Uncategorized" corpora, and click on Linguistic Landscape Corpus to enter the search and analysis interface.



5. For help searching and analyzing the LLC, refer to the FAQ Page.

==========

FAQ

What is the LL Corpus?

- The LL Corpus is the full text of 383 published journal articles (1997-2017) and 165 book chapters (2008-2018) for a total of 548 items. The LL Corpus is a freely available resource intended for use by Linguistic Landscape scholars and students as well as corpus linguists. The CQPweb search interface enables users to perform anything from simple word searches to advanced corpus analysis. The LL Corpus does not infringe on copyright restrictions because the results of searches only display 200-words of text from any item. Links in the metadata for each item will direct users to the DOI or URL for the publication so that users can access the full text via institutional or individual methods.

- **The LL Corpus is not** a substitute for the original articles and chapters that it contains.

    - **Tables, charts, figures, numerals and non-English text will be formatted differently, removed, or reproduced inaccurately** due to the conversion process from published text to corpus-searchable plain text; however, every reasonable effort was taken to make the English text as accurate as possible.

    - Many of the works contained in the LL Corpus are under **copyright** and/or not freely available. For this reason, users can only see 100 words to the left and 100 words to the right of a search term/phrase of each text—the full texts are not available. However, the links provided in the metadata should take you stable webpages where you can see either the full text or information on how to purchase or find institutional access to items.

How do I cite the LL Corpus? (APA)

Troyer, Robert A. (2021). Linguistic Landscape Corpus. CQPweb at Lancaster. https://cqpweb.lancs.ac.uk/llscape202107/

What details about the LL Corpus should I be aware of?

- In compiling the corpus, reasonable efforts were taken to ensure that the text of each article and chapter are accurate reflections of the published texts. Some details to be aware of:

- **spellings** were not standardized–British and American variants remain as they were in the original articles, and any non-standard or infrequently used spellings (or misspellings) of words were maintained; thus, if you want to retrieve examples of both "neighborhood" and "neighbourhood" you will need to specify both forms in your search using parentheses and the alternative symbol | so that the search is typed as (neighborhood|neighbourhood). The corpus is lemmatized and lemma searches can be incorporated into alternates so that ({neighborhood}|{neighbourhood}) will retrieve both the singular and plural of both spelling variants.

- we attempted to remove all **hyphens** in the original text **that were used when a word was divided at the end of a line of printed text** so that in the corpus the word appears as a whole (and can be found in searches); however, some of these divided words might have gone undetected, and some intentional hyphens (in compound words that have optional hyphens, and in long URLs) may have been deleted in the process. These minor inconsistencies should not be statistically significant or detract from the usability of the corpus; however, if you notice mistakes, please let us know so that we can fix them in a future version.

- as stated above, whenever possible the text in **tables, charts, and figures** was maintained though not in the original format; however, when the words in illustrations were part of image files (not printed text) in the originals, the words could not be included in the corpus files.

- **images/figures** in original files and any linguistic items in the images are not included in the corpus files, but every attempt was made to include the labels and captions for all of these items so that they can be found in searches. As stated above, text in languages other than English, especially those that are not written in Latin/Roman script, will vary greatly in accuracy and completeness in the corpus. Furthermore, all part of speech tagging and lemmatization is based on English structure. Creating a functioning multilingual corpus is beyond the scope of this project, so please rely only on the English text for search and analysis of the corpus while keeping in mind that some foreign words may have been automatically tagged with English part of speech tags.

What is the purpose of the LL corpus?

- The primary aim of the LL Corpus is to enable LL scholars to find more detailed and accurate information from previous studies than is available from the LL Bibliography or from large academic databases that include publications that are not related to the field of Linguistic Landscape Studies.

- Another aim is to encourage the democratization of LL research by making the work of lesser-cited authors just as accessible as that of more frequently cited scholars. When users perform a search for words or phrases they are interested in, they will obtain results from any and all publications in the corpus, and it is our hope that this leads scholars to publications they would not otherwise have discovered.

- Because the LL Corpus contains metadata categories for year of publication, the corpus can be used to explore historical developments in the field of LL Studies. Similarly, the metadata category of publication type (article or book chapter) allows for comparisons between these two major publishing venues. From a corpus linguistics perspective, it is rarely feasible to create a specialized, discipline-specific corpus of publications that is highly representative of an academic field (See the following section for representativeness). Because the LL Corpus is hosted on the CQPweb server, users can perform genre studies—for example, keyword analysis of the LL Corpus in comparison to the British National Corpus or to any of the other publicly available corpora that are also on CQPweb.

How representative is the corpus?

- The LL Bibliography on Zotero contains complete reference information for 1115 items (books, book chapters, journal articles, dissertations and theses, reports, and the annual LL Workshops). The LL Bibliography lists 427 journal articles from 1997-2017 and 247 book chapters from 2006-2018--the LL Corpus contains the full text of 383 of these articles (90%) and 165 chapters (67%) respectively from those years; thus, the LL Corpus contains approximately 80% of LL publications in journals and books during the respective periods. It is worth noting that the 10% of journal articles not included in the corpus are typically ones that were very difficult to access while the 33% of book chapters that were not included were present in a very wide variety of volumes the whole of which were not focused on LL Studies. On the other hand, the complete texts of the following edited collections of LL work are included in the LL Corpus.
    - *Linguistic Landscape: Expanding the Scenery*. 2009
    - *Linguistic Landscape in the City*. 2010
    - *Semiotic Landscapes: Language, Image, Space*. 2010
    - *Linguistic Landscapes, Multilingualism and Social Change*. 2012
    - *Minority Languages in the Linguistic Landscape*. 2012
    - *Conflict, Exclusion and Dissent in the Linguistic Landscape*. 2015
    - *Negotiating and Contesting Identities in Linguistic Landscapes*. 2016
    - *Expanding the Linguistic Landscape*. 2018
- A complete list of the metadata for each item in the LL Corpus is available on the project website as an Excel file as well as in a pdf organized alphabetically by author's last name.

How do I access the LL Corpus?

- The LL Corpus is only available through CQPweb. See the Sign Up For Access page.

How do I search and analyze the LL Corpus?

- The best resource for learning how to search and analyze the corpora that are available on CQPweb is the CQPweb Help page which you can access from the link here, or from bottom section of the navigation bar to the left of the CQPweb interface.

- The CQPweb Help system is composed of a series of user-friendly YouTube tutorials that explain how to perform everything from the most basic searches to more advanced corpus linguistic methods. The YouTube tutorials can be reached from the Help system page or directly from the "Video tutorials" link or the links here.

Where is the LL Corpus hosted?

- The Linguistic Landscape Corpus is generously hosted on the CQP Web server at Lancaster University. https://cqpweb.lancs.ac.uk/

- You can cite CQPweb as follows:

> Hardie, A (2012) CQPweb – combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17 (3): 380–409. [DOI to Full text on publisher's website] [Alternative source for PDF]

How was the corpus created?

- At the conceptual level, around 2016 we began creating a version of the LL Corpus which contained journal articles and book chapters from 1997 through 2017 for a presentation at the 10th Linguistic Landscape Workshop (LLX in Bern, Switzerland).

> Troyer, R. (May 2018). 20 Years of Linguistic Landscape Studies: A Corpus Analysis of Publications. Presentation at the 10th annual Linguistic Landscapes Workshop. Bern, Switzerland.

- That corpus served its purpose well, but we felt the best path forward would be to create a version that could be accessible to LL scholars. Many of the 357 articles and chapters in the first corpus had been modified so that variant spellings (i.e., American vs. British) were standardized and some elements of formatting that were not relevant for individual research use were inconsistent. The current LL Corpus of 548 items, however, has been created as an accurate (within the parameters discussed in this Manual) representation of the items with as much of the text of articles reproduced as possible with consistent formatting and extensive metadata that includes abstracts and web links to the publications.

- As for the nuts and bolts of corpus organization and creation, we maintain a master database of LL publications. All of these publications are fully referenced on the LL Bibliography on Zotero. We make every effort to obtain electronic full text of each of these items. When obtained (nearly always in pdf form), we use software to convert from pdf format to plain text.

What were the challenges in creating the LL Corpus?

- Journal articles published prior to around 2013 were often in pdf formats that required extensive manual editing to obtain a accurate text versions, and even those published through 2017 were often problematic to convert accurately. The good news is that with newer conversion software and newer pdfs, adding journal articles from 2018 onward should be much faster and easier.

- The frequency with which LL publications contain tables, charts, figures, images and text in non-English languages and non-Roman script makes rendering clean, corpus-searchable files complicated with decisions needing to made about what to include and exclude and how much time can be allotted to manual editing.

- Obtaining electronic versions of book chapters is very difficult. In the cases of the specifically LL-themed collections listed above, this was not challenging, and we thank the editors of several of the

above volumes for sharing the full texts of chapters with us. However, edited collections that are devoted to other or broader linguistic or sociological topics but that contain a chapter focused on the LL are very challenging to acquire and convert to text format—thus, only around 67% of the chapters listed in the LL Bibliography were able to be included in the LL Corpus. Growth of the LL Studies. While we are reasonably certain that the LL Bibliography and LL Corpus up to 2018 is as comprehensive as possible, the maturation of the field has spawned so many publications in varied sources that maintaining a 'comprehensive' collection of LL publications may not be tenable after 2020 unless significantly more funding is available. Whether the LL Bibliography and LL Corpus can be continually updated or if they eventually become an archive of the first two decades of LL scholarship, we are confident that they provide the most inclusive repository of LL work that exists and we are committed to maintaining these resources in freely available online formats.

Is the LL Corpus able to be downloaded by users?

- Due to copyright restrictions on most of what is in the LL Corpus, the actual corpus files **are not publicly available**, and access to the corpus is limited to the 200-word window around the node word of search results in the CQPweb interface. However, if researchers are interested in collaborating on a project that would necessitate use of the entire corpus outside of CQPweb, please send inquires to Rob Troyer (see Contacts page).

Will there be a future version?

- If there is enough user interest, future additions to the LL Corpus can include as many as possible of the 53 PhD Dissertations and Masters Theses that are referenced in the LL Bibliography on Zotero. Likewise, we possess pdfs of the vast majority of journal articles and book chapters published from 2017 through 2020, and following conversion and editing, these could be added to a future version of the LL Corpus.
- If you use the LL Corpus, please let us know (see the Contacts page) so that we can make a case for additional funding.

How was the creation of the LL Corpus funded?

- Creation of the corpus, directed by Rob Troyer, was enabled by funding for undergraduate research assistants provided by Western Oregon University's Community Internship Program as well as Faculty Development Funding for a Major Projects Research Grant as well as Funding for travel to Linguistic Landscape Workshops and institutional subscriptions to journals and purchases of individual publications.

==========

Contact us

For Inquiries, Comments, and Corrections, regarding the Linguistic Landscape Corpus please contact Rob Troyer (troyerr@wou.edu), Professor of Linguistics, Western Oregon University

If you experience **technical** problems with the **signup website** or the **CQPweb system**, please refer to this CQPweb User Page.