Searchable Language Documentation Corpora: DoReCo meets TEITOK

Maarten Janssen
Faculty of Mathematics and Physics
Charles University, Czechia
janssen@ufal.mff.cuni.cz

Frank Seifart
SeDyL CNRS, France
Humboldt-Universität zu Berlin, Germany
frank.seifart@cnrs.fr

Abstract

In this paper, we describe a newly created searchable interface for DoReCo, a database that contains spoken corpora from a worldwide sample of 53, mostly lesser described languages, with audio, transcription, translation, and - for most languages - interlinear morpheme glosses. Until now, DoReCo data were available for download via the DoReCo website and via the Nakala repository in a number of different formats, but not directly accessible online. We created a graphical interface to view, listen to, and search these data online, providing direct and intuitive access for linguists and laypeople, including members of speech communities. The new interface uses the TEITOK corpus infrastructure to provide a number of different visualizations of individual documents in DoReCo and provides a search interface to perform detailed searches on individual languages.

1 Introduction

Over the past 30 years, spoken corpus data have been produced through linguistic fieldwork on hundreds of languages around the world, often in attempts to document languages that are threatened of becoming extinct (Seifart et al., 2018). Typically, these were archived as part of language documentation collections in repositories such as TLA[1] and ELAR[2]. However, within these collections, the corpus data are often not easily identifiable and subject to access restrictions. Recently, the DoReCo database brought together selected high-quality corpus data from such collections, harmonized their annotations, and made them available for download (Seifart et al., 2024).

But even in DoReCo, these data are served only as raw source data, that is, as files from their respective tools. So in order to access them, users have to download the data, install the corresponding tool, and use the data locally using that tool, for instance, ELAN[3]. This means that it is not trivial for casual users to access such language documentation corpora, even though they could be valuable resources for simply getting an impression of the language or for university teaching, as well as linguistic research.

In this paper, we demonstrate how to make spoken corpora stemming from fieldwork-based language documentation directly accessible online, including for online corpus searching, by converting the source data to a corpus search tool with an online interface, building on existing tools and formats. In the example here, we convert DoReCo to TEITOK, a web-based corpus management platform that provides specific tools for spoken data and interlinear glossed data (Janssen, 2016). We first briefly describe DoReCo and TEITOK and then describe how the DoReCo data were converted into a TEITOK corpus. And finally we will demonstrate how the TEITOK online interface of the DoReCo data can be used to quickly and efficiently access the fieldwork data. The use of TEITOK also enables the corpus for use with NLP pipelines, either using the data to train NLP models or to use NLP models to further enrich the data.

2 DoReCo

DoReCo (Language Documentation Reference Corpus) is a collection of spoken corpora on a diverse set of 53 languages from around the world, with a focus on small and endangered languages. It was conceived to make data that were painstakingly collected in fieldwork in often remote areas available for cross-linguistic and cross-cultural research. As such, it addresses the problem of overreliance on what has been termed WEIRD (Western Educated Industrial Rich Democratic) populations and their languages in cognitive science (Henrich et al.,

[1] https://archive.mpi.nl/tla/
[2] https://www.elararchive.org/
[3] https://www.mpi.nl/corpus/html/elan/

Proceedings of the Fourth Workshop on NLP Applications to Field Linguistics (FieldMatters), pages 58–64

2010; Blasi et al., 2022).

Most of the corpora in DoReCo stem from efforts to document endangered languages in the framework of documentary linguistics (Himmelmann, 1998). From such documentary collections, DoReCo selected data that were suitable for cross-linguistic corpus-based research (Schnell and Schiborr, 2022). The selection criteria included the quality and consistency of annotation, the quality of the accompanying audio-recordings and that these materials could be made available using CC-BY licenses. The majority of DoReCo data are spontaneously produced traditional or personal narratives, in addition to some conversations and stimulus retellings, but not isolated examples. Each corpus had been transcribed, translated and, for 39 languages, also morphologically annotated by experts on the language prior to their inclusion in DoReCo. These experts are also the authors of the individual corpora that are edited and made available through DoReCo.

Within DoReCo, these data have been processed to add time alignment of transcription and audio through a combination of automatic forced alignment and manual corrections (Paschen et al., 2020). As a result, the start and end times for each phone, morph, and word unit are now annotated - a design motivated by research questions on phonetic lengthening (Blum et al., 2024). Other data processing steps in DoReCo included the harmonization, across the 53 corpora, of the tier structure and tier names, the documentation of the phonetic value of symbols used in the transcription, and the creation of csv files for each language, one with one word per line and another with one phone per line. DoReCo was first published in 2022, and the latest major update, containing 53 languages, was published in 2024. All DoReCo data are distributed under CC BY(-NC)(-ND-/SA) licenses.

3 TEITOK

TEITOK is an online platform for creating, managing, visualizing, and searching annotated corpora. All corpus documents in TEITOK are stored in a tokenized TEI/XML format[4]. It has a modular setup with various search and visualization methods. The default search is performed using Corpus WorkBench (CWB) (Evert and Hardie, 2011), which allows rich queries that can combine various token attributes, sequences of tokens, and can take metadata into account. The default document visualization shows linguistic information and is designed to display lemmatization, POS tagging, and dependency data. But there are also visualization modules for facsimile-aligned manuscript-based corpora, for time-aligned audio-based corpora (Janssen, 2021), and for interlinear glossed text corpora.

TEITOK was initially developed for the diachronic corpus PostScriptum (Vaamonde et al., 2014) and the learner corpus COPLE2 (Mendes et al., 2016), and has since been used for a wide variety of corpora including the multilingual Universal Dependencies corpus[5], the parliamentary corpus ParlaMint (Janssen and Kopp, 2024), dialectal corpora such as Madison[6], and corpora on less-resourced languages like CoDiaJE on Judeo-Spanish (Quintana, 2020). A list of publicly accessible TEITOK projects can be found on the TEITOK website[7].

TEITOK actively supports corpus editing and does not typically rely on corpora that have been fully developed outside of the platform. It allows users to run NLP pipelines by default using UDPIPE[8] on their data from the interface, in order to easily enrich a corpus with NLP data such as tagging, lemmatization, and dependency parsing. For fieldwork data, there typically are no NLP pipelines available, but TEITOK also allows training a tagger on the manually annotated data in the corpus, to automatically pre-tag subsequent documents with the recently trained tagger. And it provides an intuitive interface to add and correct annotations, so that errors in the automatic annotations can be corrected. This mechanism has been used, for instance, in the CoDiaJE corpus mentioned above to create a POS tagger from scratch for a language for which no NLP tools were available.

TEITOK is actively maintained and extended with new functionalities, and has an active user base. It is open source and can be easily installed anywhere from the repository[9], or run in a virtual environment from DockerHub[10]. TEITOK has been generally well received both by corpus creators and corpus users. The fact that it makes use of well established formats and tools such as TEI/XML

[4] https://tei-c.org/

[5] https://lindat.mff.cuni.cz/services/teitok/ud214/index.php

[6] http://teitok.clul.ul.pt/madison/

[7] http://www.teitok.org/index.php?action=projects

[8] https://lindat.mff.cuni.cz/services/udpipe/

[9] https://gitlab.com/maartenes/TEITOK/

[10] https://hub.docker.com/r/maartenpt/teitok

and the Corpus WorkBench means that many people will be familiar with various aspects of the interface even if they do not know the tool itself.

4 DoReCo in TEITOK

In order to create a searchable version of DoReCo in TEITOK, all original DoReCo files were converted to the TEITOK file format. The interface follows the same design layout as the DoReCo website, even though it is hosted on a different server, highlighting that it provides a visualization of the existing DoReCo version, not a re-edition.

Spoken corpora, including the DoReCo corpora, often closely transcribe what is said by the speakers, keeping track of pauses, corrections, false starts, etc. Transcription of such phenomena is typically performed in fieldwork-specific tools such as the Field Linguist Toolbox[11] or in speech-driven tools like ELAN. Since such tools use plain text for the transcription, all labels (or codes) for speech phenomena like corrections or false starts are transcribed by using special characters and labels inside the transcription.

The encoding of these phenomena by means of special characters is not ideal for a number of reasons. The first is that the labels tend to vary from corpus to corpus, so it is always necessary to provide a legend along with the corpus to explain the labels. The second is that these manually added labels are often not computer readable if they are not applied 100% consistently. The third is that the labels impede easy searching of the corpus: if we use the symbol / for a pause, then searching for "the man" will not yield results that have a pause in the middle (the / man).

The TEITOK conversion converts all DoReCo labels for disfluencies etc. into TEI/XML markup. XML is a formal language that has to be used systematically and TEI provides a set of standardized, well-described markers. This makes the resulting TEI documents compatible with other spoken data. The meaning of the markers used can be looked up in the TEI documentation for those who are not familiar with them. And they do not interfere with searches because searches are done on sequences of tokens ignoring markers. In the next section, we show how the conversion was done, and then how the converted corpus can be used for visualization and searching.

[11] https://software.sil.org/toolbox/

4.1 Conversion

The conversion from DoReCo to TEITOK was done completely automatically by a custom script that combines the metadata from the DoReCo metadata table with the transcription data from the ELAN (EAF) files. The script reads each line in the metadata table and then for each line creates a TEI/XML file in TEITOK style and saves it under the identifying name of that line. The metadata are placed in their appropriate TEI fields in the header (teiHeader), while the transcription is placed in the body (text). As a corpus search environment, TEITOK does not work with tiers, but rather with running text. For spoken corpora with multiple speakers, this typically implies an "interview style" representation of the text, in which speech turns are presented in chronological order, determined by their start time, also in case of overlapping turns.

The technical implementation of the conversion of the EAF files is as follows. Each annotation unit on the DoReCo REF tier(s), which represents a chunk of speech defined by the corpus creators as sentences, intonation units, or larger units like paragraphs, and which is associated with a translation unit, is turned into an utterance (u). The utterances are ordered chronologically by their start time to generate the interview-style representation of the text. The utterances get adorned with attributes taken from all the tiers that correspond to the utterance: the start and end time from the interval, the speaker identifier (who) from the name of the tier (ref@XX), the identifier (id) from the REF tier, the text from the TX tier, and the translation from the FT tier.

Inside the utterance, it creates tokens (tok) for each annotation within the range of the utterance from the WD tier, with the inner text corresponding to the content of the WD tier and attributes from all dependent tiers. Within each token, it creates morphemes (m) from the MB tier with its respective attributes. When the start and end times are available for tokens and morphemes, they are also added to the respective nodes.

The CWB searches do not work with units smaller than the token, which means that morphs and phones (approximated by units transcribed with X-SAMPA symbols) are not directly searchable. Therefore, the content of the MB and PH tiers are (also) kept as single string on the token, concatenating the content of the various elements. Morphemes are separated by a dot, while the X-SAMPA charac-

ters are separated by a space to increase readability. The identification of morpheme breaks follows the language-expert annotations in DoReCo.

The disfluency labels in the DoReCo transcription are converted into their respective TEI codes, as shown in Table 1. This way, the custom codes used in DoReCo are converted to standardized markers and separated from the text.

An example of a one-word sentence *Эвīлэн* from the file `2007_Ekonda_Udygir_Viktor_FSk3` in the corpus on the Siberian language Evenki (Kazakevich and Klyachko, 2024) is given in Table 2 (with some details left out for clarity).

The header of the TEI file tracks whether or not the audio file for the transcription is available. Generally, DoReCo corpora consist of a core set of annotations which have been time-aligned and for which the audio is available, and a larger set for which that is not the case. For eight DoReCo corpora, however, the audio files are only available at repositories outside DoReCo after registration, so these were not made available in the TEITOK corpus either. When the audio is not available, audio related functions are disabled for that file.

The entire conversion is fully automatic and would work not only for possible future versions or extensions of DoReCo, but also for any ELAN data following the DoReCo set-up.

4.2 Visualizations

There are three main ways to visualize the individual files in the DoReCo-TEITOK: a linguistic view, a speech-oriented view, and an interlinear glossed view. The linguistic view displays the full text of the transcription with the audio file displayed on top. Moving the mouse over one word will display a pop-up that shows all the information available for that word: the word itself, the morphological breakdown with their glosses (where available), the POS tag, and the X-SAMPA transcription.

The speech-oriented visualization displays the waveform of the audio file on top and below that the transcription of the utterances. Clicking on an utterance will play that utterance. Playing the audio will highlight which word in the transcription is currently being pronounced, and the word will also appear as a caption in the waveform image. While the waveform scrolls horizontally, the transcription scrolls vertically. Naturally, the speech-oriented visualization is available only for those transcriptions that have audio files that accessible. And the word-level visualization is only available for the files that were time-aligned at the word level. An example of a waveform visualization from the Evenki DoReCo corpus (Kazakevich and Klyachko, 2024) is given in Figure 1.

The default visualization of DoReCo-TEITOK is set to the interlinear glossed text (IGT) visualization. This is because for transcriptions that have a morphological breakdown, neither the linguistic nor the speech-oriented visualization will display the morphemes. The IGT view displays each utterance in sequence, with first the utterance, then the words of the utterance with below each word the token-level annotations such as POS, gloss, and X-SAMPA. Below that, it displays the morphemes of each word with the morpheme level annotations such as form and gloss, and, finally, the utterance-level annotations such as full text translation and the option to listen to the utterance. An example of an IGT visualization from the Evenki DoReCo corpus is given in Figure 2.

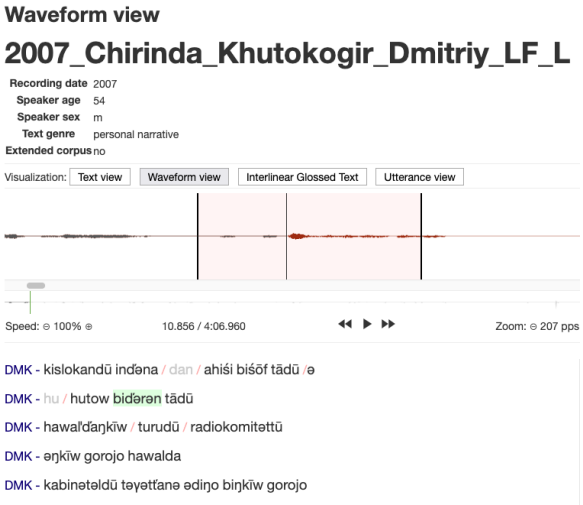

Figure 1: Waveform view example (Evenki corpus (Kazakevich and Klyachko, 2024))

4.3 Searches

From the collection of converted TEI/XML files, an indexed corpus is created in Corpus WorkBench (Evert and Hardie, 2011), making the various kinds of metadata searchable, along with all the attributes on the utterances and the tokens. The corpus can be searched using the Corpus Query Language (CQL). CQL is a well established and powerful query language used by many tools including for instance CQPWe(Hardie, 2012) and SketchEngine(Kilgarriff et al., 2014), and should be familiar to many potential users - but for people not familiar with it, TEITOK provides a user friendly

Filled pause	<<fp>uhm>	<vocal><desc>uhm</desc></vocal>
	<<fp>>	<pause type="filled"/>
Prolongation	<<pr>looonger>	<tok obs="prolongued">longer</tok>
Backchannel	<<bc>mm>	<vocal><desc>mm</desc></vocal>
False start	<<fs>fal->	<del type="falsestart">fal-</del>
Ideophone	<<id>tick>	<tok obs="ideophone">tick</tok>
Onomatopeic	<<on>moo>	<vocal type="onomatopeic"><desc>moo</desc></vocal>
Foreign material	<<fm>Weberei>	<foreign><tok>Weberei</tok></foreign>
Unidentifiable	<<ui>vubi>	<unclear><tok>vubi</tok></unclear>
	<<ui>>	<gap reason="unidentifiable"/>
Singing	<<sg>>	<gap reason="singing"/>
Silent pause	<p:>	<pause type="silent"/>
Word-internal pause	<<wip>>	<pause type="word-internal"/>

Table 1: Disfluency code conversion

```
<u who="VNU" tier="ref" start="317.92" end="318.42" eid="0089_doreco_even1259_2007_Ekonda_Udygir_Viktor_FSk3" text="¯." gloss="He began to play." id="u-86">
    <tok who="VNU" tier="wd" start="317.92" end="318.42" form="wīln" pos="v" phon="@wi:l@n" morph="wɪ.-l.-.-n" id="w-358">
        wīln
        <m who="VNU" tier="mb" start="317.92" end="318.15" form="wī" gloss="" id="m-358-1"/>
        <m who="VNU" tier="mb" start="318.15" end="318.25" form="-l" gloss="INCH" id="m-358-2"/>
        <m who="VNU" tier="mb" start="318.25" end="318.28" form="-" gloss="NFUT" id="m-358-3"/>
        <m who="VNU" tier="mb" start="318.28" end="318.42" form="-n" gloss="3SG" id="m-358-4"/>
    </tok>
</u>
```

Table 2: Example utterance in TEITOK/XML (Evenki corpus (Kazakevich and Klyachko, 2024))

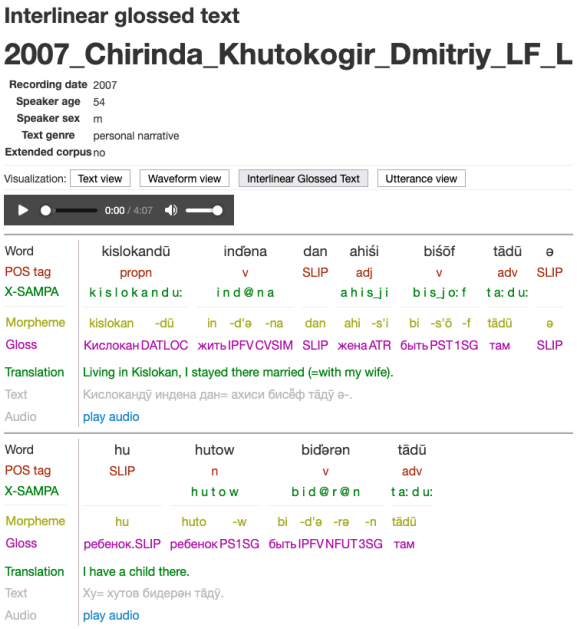

Figure 2: Interlinear glossed text view example (Evenki corpus (Kazakevich and Klyachko, 2024))

GUI to build search queries.

CQL can be used to search for words (or sequences of words) and to restrict that search to specific documents or utterances. These searches can combine any of the attributes present in the corpus: the form and X-SAMPA representation of the word, the part-of-speech (POS) tag (when available) or glosses. They can also be restricted to utterance by speakers of a certain sex or age, and to documents of a specific genre or to the core vs. extended (without time-alignment and audio) corpus sections.

This makes it possible to quickly find examples in the corpus, which facilitates its use, for instance, in teaching in linguistics programs. The results can also be used for statistical data by grouping the results by one of the categories. This makes it possible, for instance, to see the distribution of words over the different POS tags, to see whether certain types of words are more frequently used by women, or in narrative texts. The search results are rendered as utterances, and when a sound file is available, it will have a play button next to the result, making it possible to directly listen to the utterances.

Since all text-based codes in the original DoReCo data have been converted to TEI/XML codes according to Table 1, all text is searchable and disfluencies, gaps, and other markings do not hamper the search, while the information they provide is still available.

5 Conclusion

In this paper, we have shown how we created a searchable, directly accessible version of the DoReCo corpus making use of the built-in capacities of the TEITOK platform. This TEITOK ver-

sion of DoReCo is much easier to use for casual users and allows expressive searches and frequency counts to quickly find examples or quickly extract some general information on the language, for example in teaching settings.

TEITOK visualization and search functions focus on textual information present in DoReCo. It disregards the speech-related aspects of DoReCo, especially the time-alignment of annotation with audio at the phone-, morph- and word-level. For analyses taking these into account, the original DoReCo files in combination with speech-specific tools like ELAN and Praat offer functions that a corpus tool like TEITOK does not.

Currently, the DoReCo corpus in TEITOK only represents the information already provided by DoReCo. But having fieldwork corpora from ELAN made available in TEITOK not only makes it possible for casual users to search the corpus online, but also makes it possible for the corpus creators to enrich their corpus data with further annotations, such as a lemmatization, POS tags, Named Entities, or full dependency treebanks in the framework of Universal Dependencies[12]. The platform has been designed to help provide the necessary manual annotations, and furthermore provides an interface to then use the manually annotated data to train NLP tools like taggers, parsers, and named entity recognition tools.

The current project applied TEITOK to DoReCo data, but the visualization and search functions shown here would work for many other language documentation corpora. By providing fieldwork corpus data on more languages in the same interface and following the same principles, the cross-linguistic coverage for comparative corpus research could be enhanced even further. Therefore, it would be beneficial to the community if more language documentation corpora were made available in the same fashion.

References

Damián E. Blasi, Joseph Henrich, Evangelia Adamou, David Kemmerer, and Asifa Majid. 2022. Over-reliance on English hinders cognitive science. *Trends in Cognitive Sciences*, 26(12):1153–1170.

Frederic Blum, Ludger Paschen, Robert Forkel, Susanne Fuchs, and Frank Seifart. 2024. Consonant lengthening marks the beginning of words across a diverse sample of languages. *Nature Human Behaviour*, 8(11):2127–2138.

Stefan Evert and Andrew Hardie. 2011. Twenty-first century corpus workbench: Updating a query architecture for the new millennium. In *Corpus Linguistics 2011*.

Andrew Hardie. 2012. CQPweb – combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17(3):380 – 409.

Joseph Henrich, Steven J. Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3):61–83.

Nikolaus P. Himmelmann. 1998. Documentary and descriptive linguistics. *Linguistics*, 36(1):161–195.

Maarten Janssen. 2016. TEITOK: Text-faithful annotated corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4037–4043, Portorož, Slovenia. European Language Resources Association (ELRA).

Maarten Janssen. 2021. A corpus with Wavesurfer and TEI: Speech and video in TEITOK. In *Text, Speech, and Dialogue: 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6–9, 2021, Proceedings*, page 261–268, Berlin, Heidelberg. Springer-Verlag.

Maarten Janssen and Matyáš Kopp. 2024. ParlaMint in TEITOK. In *Proceedings of the IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora (ParlaCLARIN) @ LREC-COLING 2024*, pages 121–126, Torino, Italia. ELRA and ICCL.

Olga Kazakevich and Elena Klyachko. 2024. Evenki DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovvář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine: ten years on. *Lexicography*, pages 7–36.

Amália Mendes, Sandra Antunes, Maarten Janssen, and Anabela Gonçalves. 2016. The COPLE2 corpus: a learner corpus for portuguese. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Ludger Paschen, François Delafontaine, Christoph Draxler, Susanne Fuchs, Matthew Stave, and Frank Seifart. 2020. Building a time-aligned cross-linguistic reference corpus from language documentation data (DoReCo). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*,

[12] https://universaldependencies.org/

pages 2657–2666, Marseille, France. European Language Resources Association.

Aldina Quintana. 2020. CoDiAJe–the annotated diachronic corpus of judeo-spanish. *Scriptum digital. Revista de corpus diacrònics i edició digital en Llengües iberoromàniques*, (9):209–236.

Stefan Schnell and Nils Norman Schiborr. 2022. Crosslinguistic Corpus Studies in Linguistic Typology. *Annual Review of Linguistics*, 8:171–191.

Frank Seifart, Nicholas Evans, Harald Hammarström, and Stephen C. Levinson. 2018. Language documentation 25 years on. *Language*, 94(4):e324–e345.

Frank Seifart, Ludger Paschen, and Matthew Stave, editors. 2024. *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Gael Vaamonde, Ana Luísa Costa, Rita Marquilhas, Clara Pinto, and Fernanda Pratas. 2014. Post Scriptum: archivo digital de escritura cotidiana. *Janus. Humanidades digitales: desafíos, logros y perspectivas de futuro.*, pages 473–482.