

# “Tea for two”: the Archive of the Italian Latinity of the Middle Ages Meets the CLARIN Infrastructure

**Federico Boschetti**  
ILC “A. Zampolli” CNR, Pisa  
& VeDPh, Venezia, Italy  
federico.boschetti  
@ilc.cnr.it

**Riccardo Del Gratta**  
ILC “A. Zampolli” CNR  
Pisa, Italy  
riccardo.delgratta  
@ilc.cnr.it

**Monica Monachini**  
ILC “A. Zampolli” CNR  
Pisa, Italy  
monica.monachini  
@ilc.cnr.it

**Marina Buzzoni**  
ALIM, Università Ca’ Foscari  
Venezia, Italy  
mbuzzoni  
@unive.it

**Paolo Monella**  
ALIM, Sapienza Università  
di Roma, Italy  
paolo.monella  
@uniroma1.it

**Roberto Rosselli Del Turco**  
ALIM, Università degli  
Studi di Torino, Italy  
roberto.rossellidelturco  
@unito.it

## Abstract

This paper aims at showing how integrating the Archive of the Italian Latinity of the Middle Ages (ALIM) into the ILC4CLARIN repository can provide mutual benefits. Making ALIM available to a large community of scholars and researchers, on the one side, represents the first step to reduce the lack of resources for Medieval Latin in CLARIN and, on the other side, constitutes an unprecedented contribution to not only linguistic investigations, but also to the studies of the culture and science at the basis of the Western European society. The paper describes the adopted approach aiming to keep intact the structure of the archive and its metadata, which are both accurately mirrored into the ILC4CLARIN repository in order to maintain existing access practices of the users. This structure can be found in exactly the same state within the CLARIN VLO. Finally, the paper illustrates the advantages of experimenting with some ALIM data, once introduced within the CLARIN Language Resource Switchboard service: first results are shown from the analysis of some texts with the UDPipe tool suite and the distant reading tool Voyant.

## 1 Introduction

The Archive of the Italian Latinity of the Middle Ages – in Italian, Archivio della Latinità Italiana del Medioevo (ALIM) – is an Italian national research project aimed to provide free online access to a large number of Latin texts produced in Italy during the Middle Ages. ALIM makes an unprecedented contribution to the study not only of Latin, but also of the culture and science at the basis of the Western European society. For several centuries, in fact, Latin represented the only language in which many of the major creations of thought, science, and literature of the Middle Ages were expressed. Even when national languages imposed themselves in written form, Latin never lost its role and prestige as a transnational language – until the end of the Middle Ages and beyond.

---

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

Federico Boschetti, Riccardo Del Gratta, Monica Monachini, Marina Buzzoni, Paolo Monella and Roberto Rosselli Del Turco 2021. “Tea for two”: the Archive of the Italian Latinity of the Middle Ages meets the CLARIN Infrastructure. *Selected papers from the CLARIN Annual Conference 2020*. Linköping Electronic Conference Proceedings 180: 180 37–46.

The general aim of this paper is to place ALIM within the framework of CLARIN-IT and CLARIN at large. Section 2 shows how ALIM may contribute to fill an important gap in the availability of literary and historical sources for Latin: query searches run on the Virtual Language Observatory for Latin-related resources demonstrate that no resource with the features and potentialities of ALIM is currently available. The internal structure and the metadata of the Archive are presented in Section 3, while the strategy for the integration of ALIM into the ILC4CLARIN repository is discussed in Section 4. Finally, ALIM's contribution in strengthening and widening the research directions in CLARIN-IT and its advantages for the CLARIN community are presented.

## 2 Latin resources in CLARIN-IT

The Italian CLARIN (CLARIN-IT) consortium<sup>1</sup> (Nicolas et al., 2018) has a strong interest in the field of Digital Classics, which is still affected by a shortage or restricted availability of language resources for historical languages such as Ancient Greek and Latin. To this end, the consortium aims to make some of the existing digital resources for Ancient Greek and Latin available through its national repository – ILC4CLARIN (Pisa).

Within the CLARIN-IT consortium, the collaboration between the Centre for Comparative Studies “I Deug-Su”, Department of Philology and Literary Criticism at the University of Siena – DFCLAM<sup>2</sup> and the ILC4CLARIN data center mostly concerns the study of methods and the development of services to offer online secure access to some digital archives of literary and historical texts. ALIM, currently hosted by the University of Siena<sup>3</sup>, is the largest digital library of the Italian Latinity, including both literary and documentary sources.

Evidence shows that the CLARIN data centers do not offer resources such as ALIM. The faceted-search functionality of the Virtual Language Observatory (VLO), performed combining *Latin text resource* and *Middle Ages*, returns 53 records, while 124 records are returned by the query *Latin* combined with the adjective *medieval*. These data are essentially images of manuscripts; no XML-TEI texts seem to be available by using these search keys. A further query with XML as ‘free text’, Latin as ‘language’, and text or corpora as ‘resource type’ returns about 1300 records, mostly consisting in Treebanks or in documents coming from the EUROPEANA platform<sup>4</sup>.

ALIM represents the first step to reduce the lack of resources for Medieval Latin in CLARIN-IT and eventually in CLARIN. On the one hand, ALIM will offer high-quality data since: (i) the resources are curated by domain experts; (ii) a strong organization is dedicated to maintenance; (iii) the resources cover a broad historical period; (iv) the resources are TEI encoded. On the other hand, the Language Resource Switchboard and the Weblicht workflow engine will use the texts provided by ALIM to produce both engaging visualizations and interesting linguistic analyses.

## 3 ALIM: history, goals, structure

ALIM is an archive of medieval Latin texts composed in the Italian area between the 8th and 15th centuries. It originated as a UAN (Unione Accademica Nazionale) project in the Nineties and was later supported by the national Ministry of Education. Its original aim was twofold: to make medieval Latin literature texts openly available and to provide a textual corpus serving as a basis to create a new dictionary of medieval Latin in its Italian variety. The latter goal explains a unique feature of ALIM: it does not only include literary sources, but also collections of documentary texts. ALIM is, therefore, divided into two sections: “Fonti letterarie” and “Fonti documentarie”. While the majority of texts are drawn from printed editions, some are new, born-digital editions<sup>5</sup>.

<sup>1</sup>The composition of the Italian Consortium is available at <http://clarin-it.it/en/content/consortium>.

<sup>2</sup>Prof. Francesco Vincenzo Stella.

<sup>3</sup><http://alim.unisi.it/il-progetto/>

<sup>4</sup><http://www.europeana.eu>

<sup>5</sup>More information on the history and scientific objectives of ALIM, with further bibliography, are in (Alessio, 2003); (Buzzoni and Rosselli Del Turco, 2016, par. 7.1.2); (Ferrari, 2017) and (D'Angelo and Monella, 2019).

### 3.1 From ALIM1 to ALIM2, from ALIM2 to CLARIN: text and metadata

Until 2016, ALIM was hosted by the servers of the University of Verona, Italy<sup>6</sup> and its texts were annotated with procedural markup, based on simple HTML markers. We shall refer to this version as “ALIM1”.

In 2016, the current version of the archive (“ALIM2”) was launched. The migration process involved the following tasks: (1) building a new open source software TEI XML-based digital library infrastructure and publishing it on the servers of the University of Siena<sup>7</sup>; (2) re-encoding text markup and metadata in TEI XML P5.

Task 1 was realised in collaboration with the external IT company Net<sup>8</sup> and completed in 2016/17, when the ALIM2 website was launched. Task 2 involved a longer process, still ongoing, curated by the “équipe di codifica” (Ferrarini, Monella and Rosselli Del Turco) to gradually improve the level of formalisation and the granularity of text markup and metadata.

In the current version of the archive, each literary text is encoded as a TEI XML P5 file with a <TEI> root element, while in the documentary section, each TEI XML file includes a whole volume of a documentary collection<sup>9</sup>, has a <teiCorpus> root element and includes each individual document in a separate <TEI> element. In the latter case, both <teiCorpus> and <TEI> have their own <teiHeader> with metadata respectively regarding the whole collection and the individual document.

In the ALIM2 TEI-XML files for literary texts deriving from the initial export from ALIM1 (labeled as “encoding level ALIM2\_0”), much metadata was still included in unstructured <note> elements of the TEI. Also, most texts lacked any TEI structural markup such as <div>. In 2017/18, literary texts were gradually upgraded to “encoding level ALIM2\_1”, thanks to the work of ALIM collaborator Chiara Casali on metadata integrity and of Jan Ctibor on metadata encoding and structural markup. Ctibor’s activity was brought forth in the framework of a collaboration agreement between ALIM and the *Corpus Corporum*<sup>10</sup>, the largest full-text repository for Latin (163 M words). The current policy of ALIM requires that all new texts included in the archive must be encoded at “level ALIM2\_2”: this includes markup of work titles, quotes, speeches, persons, or place names.

The archive also includes born-digital scholarly editions directly based on handwritten medieval witnesses, whose encoding level is labeled as ALIM2\_3<sup>11</sup>.

The ALIM project provided CLARIN-IT with the TEI headers of the XML files in the archive, at the highest available encoding level, to extract metadata from them.

## 4 ALIM in CLARIN-IT

### 4.1 Structure for ALIM data into ILC4CLARIN repository

As described in Section 3, the ALIM digital library is arranged into two complementary sections: *Fonti Letterarie (Literary Sources)* and *Fonti Documentarie (Documentary Sources)*. The former is a collection of single documents (about 350), while the latter is a collection of 50 corpora that groups about 6455 texts. Since ALIM keeps these two resources separated, we decided to mirror this structure in the ILC4CLARIN repository. We created two collections, *Literary Sources* and *Documentary Sources*, under the *OPEN* community<sup>12</sup>. This structure is important for, at least, two reasons. The first underlying motivation for this partially conservative choice was that this decision would provide philologists, linguists, and historians with a user experience on the VLO consistent with the navigation in the original ALIM environment. The second reason is directly con-

<sup>6</sup><http://www.alim.df11.univr.it/>

<sup>7</sup><http://alim.unisi.it/>

<sup>8</sup><https://www.netseven.it/>

<sup>9</sup>E.g.: *Codex diplomaticus Cavensis*, volume 1: [http://alim.unisi.it/dl/fonte\\_documentaria/7381](http://alim.unisi.it/dl/fonte_documentaria/7381).

<sup>10</sup><http://www.mlat.uzh.ch/MLS/>

<sup>11</sup>See <http://alim.unisi.it/collection/nuove-edizioni-editiones-principes-e-prime-trascrizioni/> for a list of such editions. In general, on markup levels see the *Manuale di codifica dei testi ALIM in TEI XML* in <http://alim.unisi.it/documentazione/>

<sup>12</sup>Given that ILC4CLARIN uses the clarin-dspace repository, we have used the terminology community and collections. For clarity, collections are nested into communities.

nected with the VLO. In section 2, we briefly mentioned the faceted-search of the VLO. One of such facets is the collection (in the original repository) the data come from. The ALIM data are retrieved from the VLO using either “fq=collection:ALIM+Literary+Sources&fqType=collection” or “fq=collection:ALIM+Documentary+Sources&fqType=collection”<sup>13</sup>.

## 4.2 Population of the repository with ALIM data

The about 350 *Literary Sources* have complete descriptive metadata, although period, author and title are often debated in the scholarly community and, therefore, tentative in the collection. Author names have two issues: the actual authorship attribution and alternative Latin spellings of the name. Titles too are not always standardised, and the very identification of the “work”, as well as of the composition period, is problematic. However, each of these metadata fields has a value in ALIM (for the author, it can also be “Anonimo”). The 50 corpora of *Documentary Sources* group 6455 small documents. For these small documents the metadata set differs from *Literary Sources*, since they do not represent a creative work by an author. For example, private documents are actually written by a notary, but their “author” is the stakeholder (the person who buys, sells etc.), while charters are created by a public institution.

As a consequence, we decided to completely import *Literary Sources* metadata into the repository, but, at the same time, to describe only the 50 corpora of *Documentary Sources*, without importing the whole amount of data (even if technically possible).

The ratio behind this decision is related to the ALIM organization again. As noted in Section 3.1, the TEI version of each document in literary sources has its own `<teiHeader>`, corresponding to the TEI root element, that can be parsed. While for documentary sources the most informative `<teiHeader>` is extracted from `<teiCorpus>`, for literary sources metadata are extracted from the header of each files’ `<TEI>` element.

Given the large number of items to describe in the repository, we decided to use the import functionality of the repository<sup>14</sup> to batch-load the items. Since this procedure is unsupervised, as far as the content of the items is concerned, we decided to manually create a prototypical item, export it, and automatically clone it. In this way, every item is syntactically correct and can be safely imported into the repository. More in detail: (i) we took one document from literary sources and one from documentary works and kept them as prototypes; (ii) we carefully created a submission, mapping the elements of the `<teiHeader>` into the fields of the submission form of the repository; and (iii) once the internal workflow of metadata quality is passed, we exported the item.

The exported item is an archive which contains the following metadata files: **metadata\_local.xml**, **dublin\_core.xml**, and **metadata\_metashare.xml**. All of them are populated with data extracted from elements of the `<teiHeader>`. The different metadata files combine to create the descriptive items in the repository. The ALIM research team checked sample metadata from the CLARIN archive and verified that they correspond to those included in the TEI headers of the ALIM XML files and to the general project information pertaining to the archive. It is important, here, to notice that the official URL of the ALIM project (in our case, <http://it.alim.unisi.it/>) is contained in the **dublin\_core.xml** files, while **metadata\_local.xml** files contain the *demo URL*, that is to say where the resource stays in the ALIM digital library.

This mapping enforces our decision to describe the `<teiCorpus>` instead of describing every single document in the corpus. Literary Sources have a clear URL where the document resides: for example, the “Dialogus” by Gerius Aretinus is available at <http://it.alim.unisi.it/dl/resource/194>. By contrast, Documentary Sources point to URLs that report the whole corpus. For example, the “Codex diplomaticus Cavensis - 01” is available at [http://it.alim.unisi.it/dl/fonte\\_documentaria/7381](http://it.alim.unisi.it/dl/fonte_documentaria/7381). On the web page, a JavaScript function allows the user to jump to the desired documents, such as the 27<sup>th</sup> document, whose internal URL is [http://it.alim.unisi.it/dl/fonte\\_doc](http://it.alim.unisi.it/dl/fonte_doc)

<sup>13</sup>At the time of writing, only the *Literary Sources* have been imported into the ILC4CLARIN production repository. The items are available at <https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/handle/000-c0-111/130>.

<sup>14</sup><https://wiki.lyrasis.org/display/DSDOC5x/Importing+and+Exporting+Items+via+Simpl e+Archive+Format>



To make the versioning of the ILC4CLARIN repository consistent with that of ALIM, we decided to remove the *demo URL* from the old versions. In this way, users access the latest version of the document from the repository and, if they still need older data, they can contact ALIM and request them.

## 5 CLARIN Services and ALIM

### 5.1 Available analysis tools

In this section, we show the results of an analysis of an ALIM sample text with tools made available through the CLARIN infrastructure.

The Language Resource Switchboard (LRS)<sup>16</sup> (Zinn, 2018) has been used to connect the input data with suitable and available tools. Figure 3 below shows the suggested tools for the input file “Historia Langobardorum”<sup>17</sup>, which consists of about 38000 words.

The screenshot displays the LRS interface. At the top, under the 'Resources' section, a file named 'Paulus Diaconus - Historia Langobardorum(1).txt' (264.08 KIB) is shown. To its right are dropdown menus for 'Mediatype' (set to 'text/plain') and 'Language' (set to 'Latin'). Below this is the 'Matching Tools' section, which includes a 'Group by task' checkbox and a 'Search for tool' input field. Three tool categories are visible: 'Dependency Parsing' with 'UDPipe' (LINDAT logo), 'Distant Reading' with 'Voyant Tools' (Voyant logo), and 'Text Analytics' with 'WebLicht Advanced Mode' (WEBLIGHT logo).

Figure 3: “Historia Langobardorum” and connected tools.

The LRS lists a distant reading tool, Voyant<sup>18</sup> and a dependency parsing tool, UDPipe (Straka and Straková, 2020; Straka and Strakova, 2017).

Figure 4 displays the *Cirrus* and *TermsBerry*, while Figure 5 provides some textual statistics on the examined text<sup>19</sup>.

<sup>16</sup><https://switchboard.clarin.eu/>

<sup>17</sup><http://hdl.handle.net/20.500.11752/OPEN-152>.

<sup>18</sup><https://voyant-tools.org/>

<sup>19</sup>We invite the interested readers to replicate the experiment by providing the URL <http://alim-admin.unisi.it/download.txt?id=201> to the LRS and eventually use *Voyant tools*.



# text = Inter haec moritur Godehoc, cui successit Clafoo, filius suus									
1	Inter	inter	ADJ	AAXXX----1A----	Degree=PoslForeign=YeslPolarity=Pos	0	root	_	TokenRange=0:5
2	haec	haec	NOUN	AAXXX----1A----	Degree=PoslForeign=YeslPolarity=Pos	1	flat:foreign	_	TokenRange=6:10
3	moritur	moritur	NOUN	NNMS1-----A----	Animacy=AnimlCase=NomiGender=MascplNumber=SinglPolarity=Pos	1	flat:foreign	_	TokenRange=11:18
4	Godehoc	Godehoc	PROPN	NNMS1-----A----	Animacy=AnimlCase=NomiGender=MascplNameType=GivlNumber=SinglPolarity=Pos	1	flat:foreign	_	SpaceAfter=NoI TokenRange=19:26
5	,	,	PUNCT	Z:-----	_	7	punct	_	TokenRange=26:27
6	cui	cui	PRON	PQ--4-----	Animacy=InanlCase=AcclPronType=Int,Rel	7	obl:arg	_	TokenRange=28:31
7	successit	successit	VERB	VpYS---XR-AA---	Aspect=PerflGender=MascplNumber=SinglPolarity=PoslTense=PastlVerbForm=PartlVoice=Act	3	acl:reicl	_	TokenRange=32:41
8	Clafoo	Clafoo	PROPN	NNMS1-----A----	Animacy=AnimlCase=NomiGender=MascplNameType=GivlNumber=SinglPolarity=Pos	7	nsubj	_	SpaceAfter=NoI TokenRange=42:48
9	,	,	PUNCT	Z:-----	_	10	punct	_	TokenRange=48:49
10	filius	filius	NOUN	NNMS1-----A----	Animacy=AnimlCase=NomiGender=MascplNumber=SinglPolarity=Pos	8	appos	_	TokenRange=50:56
11	suus	suus	NOUN	NNIS1-----A----	Animacy=InanlCase=NomiGender=MascplNumber=SinglPolarity=Pos	10	nmod	_	SpaceAfter=NoI TokenRange=57:61

Figure 6: A sample UDPipe table.

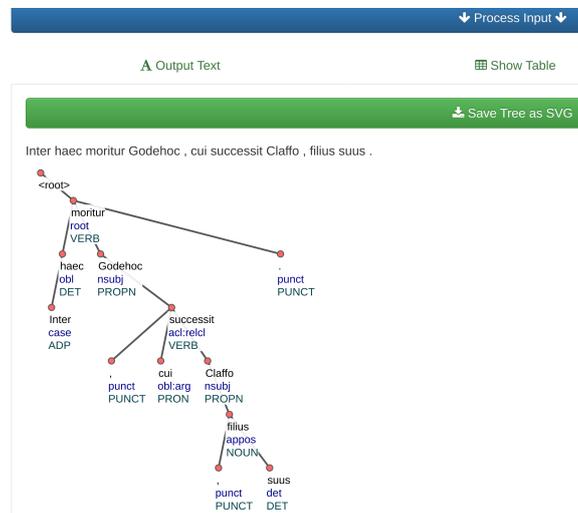


Figure 7: A sample UDPipe tree.

A few words on the use of WebLicht for the ALIM texts. As a workflow engine, WebLicht can run as many tools as needed. Figure 8 reports the analysis chain which can be run on the same sentence. The chain consists of three modules: a tokenizer, a tagger, and a parser. The tokenized text can be further used in the Federated Content Search (Stehouwer et al., 2012), cf. Section 6.

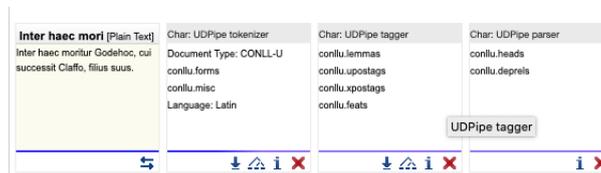


Figure 8: WebLicht analysis chain.

## 5.2 Are similar resources available?

The “Historia Langobardorum” was already available from the VLO as a digital facsimile of the medieval manuscript provided by e-codices, the Virtual Manuscript Library of Switzerland (<https://>

[//www.e-codices.ch/en/csg/0635/1/0/](http://www.e-codices.ch/en/csg/0635/1/0/)). Even if the direct source of the text of “Historia Langobardorum” in ALIM is the digitization of the nineteenth-century edition by L. Bethman - G. Waitz<sup>24</sup>, the manuscript available from the VLO is one of its witnesses. Indeed, the increased use of the IIF ([www.iif.io](http://www.iif.io)) framework to publish digital facsimiles of medieval manuscripts means that for many of the ALIM texts there are already one or more digitized witnesses freely available on the Web. Not only that, this opens the way for a future connection of an ALIM text to those facsimiles: for the most advanced ALIM critical editions, those belonging to level 3 of the encoding, i.e. a full use of the elements of the Critical Apparatus TEI module (cf. Section 3.1), this could also mean generating automatically the corresponding witness and linking it to the digitized images of the manuscript. s

## 6 Concluding Remarks

The DFCLAM committed itself to offering data and free online access to some digital archives of literary and historical texts: one of them is ALIM (the Archive of the Italian Latinity of the Middle Ages), the largest digital library of the Italian Latinity including both literary and documentary texts, encoded in TEI XML from philologically checked printed editions or published directly from manuscripts produced in Italy during the Middle Ages, in new born-digital scholarly editions. Strategies for importing the metadata of ALIM in the ILC4CLARIN repository through a shared TEI header are under study, as well as procedures for delivering dedicated tools for textual and linguistic analysis through the CLARIN channels. This would allow meta-queries and cross-queries on semantic items which could connect Latin and modern European languages derived from Latin and allow to develop semantic trees and networks of lexical derivations at the very heart of the European shared vocabulary.

ALIM complements the Latin resources in CLARIN by providing access to a large corpus of medieval literary and documentary Latin texts with granular curated metadata. On the other hand, participating in CLARIN provides ALIM with a valuable opportunity in terms of sustainability, long term preservation, persistent identification, format migration, and visibility for the ALIM research outputs. The VLO makes the resources produced and described in the ILC4CLARIN repository, including ALIM metadata, available to a wider audience in the SSH community, while the CMDI model ensures high quality metadata curation. Also, CLARIN offers ALIM the possibility to use technology and text analysis tools available at CLARIN data centers to deal with multilingual data. For example, Weblicht allows to combine web services so as to handle and exploit textual data, while the Language Resource Switchboard can connect the ALIM texts to visualization tools such as Voyant. A further reciprocal advantage is that CLARIN contributed in enhancing the ALIM strategies on Open Access and open source policies by supporting ALIM in planning the actions necessary to provide FAIR (Findable, Accessible, Interoperable, Reusable) data (de Jong et al., 2018).

Finally, ILC4CLARIN is an endpoint of the Federated Content Search (FCS)<sup>25</sup>, a tool to query data distributed across local collections available at the various CLARIN centres. At the time of writing, the only CLARIN collection including Latin texts (`lat_wikipedia_2012_100K`, about 1.5M tokens) is available at the FCS aggregator from the Automatische Sprachverarbeitung - Universität of Leipzig. The addition of further sources to the aggregator will be of fundamental importance to increase the number of high-quality Latin texts available through the Federated Content Search.

---

<sup>24</sup>MGH SS rer. Lang., Hannover 1878, pp. 45-18

<sup>25</sup><https://contentsearch.clarin.eu/>.

## References

- Gian Carlo Alessio. 2003. Il progetto alim (archivio della latinità italiana del medioevo). In Francesco Santi, editor, *In Biblioteche elettroniche. Letture in Internet: una risorsa per la ricerca e per la didattica*, volume 1, pages 73–81. SISMEL - Edizioni del Galluzzo.
- Marina Buzzoni and Roberto Rosselli Del Turco. 2016. Evolution or revolution? digital philology and medieval texts: History of the discipline and a survey of some italian projects. In *Mittelalterphilologien heute. Eine Standortbestimmung. Band 1: Die germanischen Philologien*, pages 265–294. Königshausen und Neumann.
- Edoardo D’Angelo and Paolo Monella. 2019. ALIM (Archivio della Latinità Medievale d’Italia). Storia, attualità, prospettive di una banca-dati di testi mediolatini. In Roberto Gamberini, Paolo Canettieri, Giovanna Santini, and Rosella Tinaburri, editors, *La Filologia Medievale. Comparatistica, critica del testo e attualità. Atti del Convegno (Viterbo, 26-28 settembre 2018)*, volume 3 of *Filologia Classica e Medievale*. L’Erma Di Bretschneider.
- Franciska de Jong, Bente Maegaard, Koenraad De Smedt, Darja Fišer, and Dieter Van Uytvanck. 2018. CLARIN: Towards FAIR and responsible data science using language resources. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).
- Edoardo Ferrarini. 2017. ALIM ieri e oggi. *Umanistica Digitale*, 1:7–17.
- Erhard Hinrichs, Marie Hinrichs, and Thomas Zastrow. 2010. WebLicht: Web-based LRT services for German. In *Proceedings of the ACL 2010 System Demonstrations*, pages 25–29. Association for Computational Linguistics.
- Lionel Nicolas, Alexander König, Monica Monachini, Riccardo Del Gratta, Silvia Calamai, Andrea Abel, Alessandro Enea, Francesca Biliotti, Valeria Quochi, and Francesco Vincenzo Stella. 2018. CLARIN-IT: State of Affairs, Challenges and Opportunities. In *Selected papers from the CLARIN Annual Conference 2017, Budapest, 18-20 September 2017*, Linköping electronic conference proceedings (Print), pages 1–14.
- Herman Stehouwer, Matej Durco, Eric Auer, and Daan Broeder. 2012. Federated search: Towards a common search infrastructure. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).
- Milan Straka and Jana Strakova. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipes. pages 88–99, 01.
- Milan Straka and Jana Straková. 2020. Udpipes at evalatin 2020: Contextualized embeddings and treebank embeddings. *arXiv preprint arXiv:2006.03687*.
- Claus Zinn. 2018. The language resource switchboard. *Comput. Linguist.*, 44(4):631–639, December.