Journal of Open Humanities Data

DISSILEX: A Lexico-Semantic Network and Valency Lexicon for Medieval Latin

DATA PAPER

DAVID ZBÍRAL
GIDEON KOTZÉ
SOŇA HUDÍKOVÁ
KATIA RICCARDO
KATALIN SUBA
STANISLAW BANACH
ROBERT L. J. SHAW
*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

DISSILEX is a manually compiled and curated lexico-semantic network and valency lexicon initially developed for the semantic annotation of texts portraying resistance, conflict, and prosecution connected with medieval heterodox movements. This linguistic resource currently covers 948 Latin verbal entries, 185 Latin entries modelling nouns and adjectives, and 1,673 English conceptual entries, interconnected through 6,256 distinct relations. Its main lexicographic contribution lies in mapping meanings unknown to Latin WordNet for the given lemmas, and in detailed descriptions of verbal valencies including semantic roles. It is most useful for informing (1) word-sense disambiguation in the natural language processing (NLP) of Latin, (2) categorization tasks with large language models, and (3) event extraction with rich semantic role labelling.

CORRESPONDING AUTHOR:
David Zbíral
Department for the Study of Religions, Faculty of Arts, Masaryk University, Brno, Czech Republic
david.zbiral@mail.muni.cz

KEYWORDS:
medieval Latin; lexicography; controlled vocabulary; lexico-semantic network; valency lexicon; semantic roles

TO CITE THIS ARTICLE:
Zbíral, D., Kotzé, G., Hudíková, S., Riccardo, K., Suba, K., Banach, S., & Shaw, R. L. J. (2026). DISSILEX: A Lexico-Semantic Network and Valency Lexicon for Medieval Latin. *Journal of Open Humanities Data*, 12: 143, pp. 1–7. DOI: https://doi.org/10.5334/johd.637

(1) OVERVIEW

REPOSITORY LOCATION

https://doi.org/10.5281/zenodo.23104421

We also provide a search interface for convenience: https://dissinet.cz/apps/dissilex/.

CONTEXT

DISSILEX is a digital linguistic resource mapping the distinctive and linguistically understudied domain of Latin-language medieval inquisition records. It represents a lexico-semantic network and controlled vocabulary of verbs and multi-word verbal expressions ("Actions") as well as single- and multi-word expressions representing other parts of speech ("Concepts"). It emerged from the needs of two projects within the Dissident Networks (DISSINET) research group (Zbíral & Shaw, 2022; Zbíral, Sikk, et al., 2025). The first looks into the social and religious dynamics of dissidence and inquisition, while the other focuses on medieval social exclusion through labelling and text production. Both research strands rely on data representation through Computer-Assisted Semantic Text Modelling (CASTEMO) (Zbíral et al., 2022, 2026). This data collection workflow and data model are implemented in InkVisitor, an open-source browser-based application for entity–relationship building, clause modelling, and semantic annotation (Zbíral, Hanák, et al., 2025).

Efficient computational study of both topics required a controlled vocabulary covering both original-language expressions used in our sources and generic analytical categories expressed in modern English (such as "attitude", "preacher", "punishment", "defendant", "transgressor", etc.). This vocabulary, formed of Action and Concept entries, furthermore needed to embed them in a dense network of semantic relations: some standard in ontology building and lexicography, others less conventional but providing lexicographically useful machine-operable ties. For purposes of interoperability, we linked Actions and Concepts to two major external resources wherever possible and relevant: the Linking Latin (LiLa) Lemma Bank, a key entry point into Latin Linguistic Linked Open Data (Mambrini et al., 2021), and WordNet, using the Collaborative Interlingual Index (CILI) to map between versions (Bond et al., 2016). WordNet links were made to the multilingual Princeton WordNet (PWN) (Fellbaum, 1998; Miller, 1995), rather than limiting the resource to meanings already identified by Latin WordNet (LWN) (Minozzi, 2017). As its key lexicographic contributions, DISSILEX:

1. describes the *valency frame* of any Action in terms of not only morphosyntax (case, preposition…) and entity type (which entity type can occupy the given actant slot, e.g. person, group, event…), but also *fine-grained, frame-specific semantic roles* (i.e., semantic microroles), whose specificity exceeds the general roles provided in Latin VALLEX (Passarotti et al., 2016); e.g., for Latin *do* in the sense of "to give as a gift", the actant semantic roles in DISSILEX would be "giver", "beneficiary", and "gift";
2. covers *multi-word expressions*, including their embedding in lexico-semantic relations (e.g., direct superclass, which often cannot be straightforwardly derived from the semantic relations of their constituent words: for instance, "*do penitentiam*", "to assign penance", is semantically underdefined by the meanings of its constituent parts alone, but safely lands under "sacrament" and "ritual" thanks to the superclass path of its nominal equivalent);
3. provides 89 PWN meanings for Latin Actions and Concepts *missing altogether in LWN*;
4. provides (at the time of writing) 265 lemma-PWN meaning pairs for Latin Actions and Concepts that *LWN does not attest for the given lemmas*;
5. covers 602 meanings – counting only single-word entries to allow comparison on an equal footing – which DISSILEX contributors identified as *absent from PWN*. These are either meanings specific to the thematic domain (inquisition, prosecution, criminalization of heresy, social interactions in medieval society), or meanings for which the PWN gloss was deemed inadequate.

Figure 1 shows an example entry and its relations.

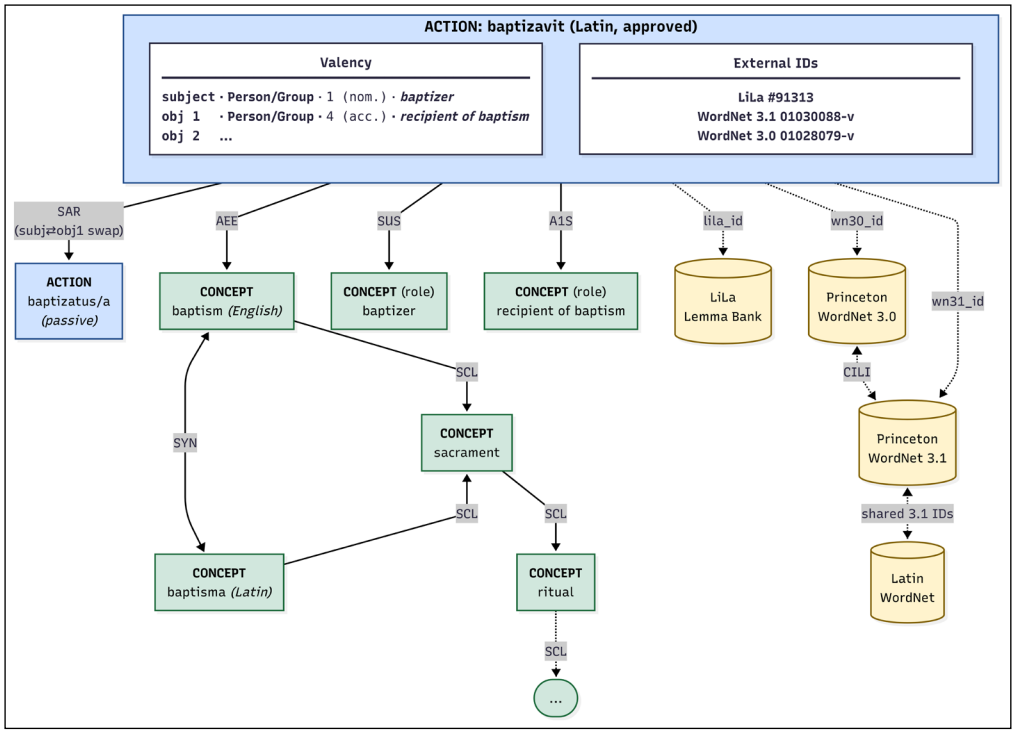


Figure 1 Action baptizavit ("to baptize") and its semantic relations (SCL = superclass; SYN = synonym; SAR = subject/actant 1 reciprocal; AEE = Action/Event Equivalent; SUS = subject semantic role, A1S = actant 1 semantic role) and references to external resources.

(2) METHOD

SAMPLING STRATEGY

Each entry in DISSILEX is a conceptual lemma-meaning unit, identified by its lemma and distinct meaning. For instance, the verb *dico* would be covered in multiple Action entries, one for each meaning of the *dico* lemma. To maintain full analytical clarity regarding semantic roles, we also treat passive and reflexive sublemmas as distinct entries. Any entry (Action for verbs and verbal expressions, Concept for all other parts of speech) has its own UUID (universally unique identifier), a machine-generated alphanumeric string that is unique to each record and allows stable referencing.

Unlike standard dictionaries or genuinely lexicographic projects, DISSILEX is not guided by a balanced and systematically sampled word list, but is rather a source-driven by-product of our annotation. The annotation covers selected texts from a large (ca. 2.5 million tokens) corpus of medieval inquisition records. Depositions have received the largest coverage among document types. Regarding region and period, the first hundred years of the inquisition (the 1230s to the 1320s) in Languedoc and North-Central Italy are best covered. When compiling the entries, we used several dictionaries of medieval Latin (esp. Quetglas et al., 2021; Du Cange, 1883; Niermeyer et al., 2002).

STEPS

Initial lists of Actions and Concepts were compiled in 2019 in tabular format (Google Sheets). As the CASTEMO data model grew, we adopted a JSON document database maintained in RethinkDB (*RethinkDB*, n.d.). The DISSILEX dataset shared here is a filtered subset of that multilingual database, restricted to the most lexicographically relevant and consistently curated entities: it includes only Latin Actions and the lexico-semantic network of English- or Latin-language Concepts linked to them through any semantic relation. We further restricted the dataset to entities with "approved" or "pending" status in our database, excluding those marked "unfinished" or "discouraged".

To facilitate querying, we developed a data transformation script to produce a SQLite derivative, built upon a more accessible data model. This is the version shared in the accompanying repository. It represents DISSILEX as a series of related SQL tables (*actions*, *concepts*, *valency*, *relations*, *external_ids*, and also *_meta* for attribution of external resources). The schema of the database is visualized and described in more detail in the accompanying Zenodo repository. Using SQL querying, a user can, for example, track the coverage of external resources, list a breakdown of relation types, extract active/passive meaning pairs, study the distribution patterns of valency frames, superclass taxonomies, or DISSILEX synonym clouds (synsets), or extract typical entity type interactions including the Actions that connect them.

QUALITY CONTROL

Entries are curated at two different levels, indicated by their status field: “approved” entries have been checked to a higher standard; “pending” have been checked to a lesser standard, but are shared as work-in-progress. The definitions of “pending” entities may be less polished or missing (but often still clear from the semantic relations) and external connections (LiLa, PWN) are more frequently absent. We used a script both to verify the technical consistency of the dataset and to produce outputs for validation. This ensured a high degree of compliance of “approved” entries with the research group’s guidelines.

(3) DATASET DESCRIPTION

REPOSITORY NAME

Zenodo

OBJECT NAME

dissilex.db

FORMAT NAMES AND VERSIONS

SQLite Database File Format, version 3

CREATION DATES

2019-11-02 to 2026-08-21

DATASET CREATORS

David Zbíral: conceptualization, supervision, investigation, data curation, writing – original draft, writing – review & editing.

Gideon Kotzé: conceptualization, supervision, investigation, data curation, software, formal analysis, writing – original draft, writing – review & editing.

Soňa Hudíková: investigation.

Katia Riccardo: investigation.

Katalin Suba: investigation.

Stanislaw Banach: investigation.

Robert L. J. Shaw: conceptualization, supervision, investigation, data curation, writing – review & editing.

LANGUAGE

English, Latin.

LICENSE

CC BY 4.0

PUBLICATION DATE

2026-10-02

OPTIONAL SEARCH INTERFACE

https://dissinet.cz/apps/dissilex/

(4) REUSE POTENTIAL

DISSILEX has several limitations and avenues for future improvement. Most of these arise because it is a lexicographic by-product of historical research rather than a purpose-built linguistic resource. Among the most significant: it is source-driven rather than based on a balanced word list; it contains "pending" entities curated to a lower standard; and it is not yet aligned with formal lexicographic standards, such as TEI Lex-0 (Romary & Tasovac, 2018) or LMF (Francopoulo, 2013), or the Linguistic Linked Open Data (LLOD) models, such as OntoLex Lemon (McCrae et al., 2017) or PreMOn (Corcoglioniti et al., 2016).

Nevertheless, DISSILEX already holds lexicographic value in its current form. It provides: (1) a detailed description of valency frames with actant semantic roles, embedded in hierarchical superclass relations (hypernym paths); (2) hundreds of meanings of Latin lemmas not yet accessible in a machine-operable form through LWN or LiLa; and (3) multi-word expressions embedded in semantic relations. Its main potential for reuse lies in informing:

1. automated word-sense disambiguation for Latin;
2. transformer-based classification tasks on historical trial records;
3. automated event extraction from Latin texts with rich semantic role labelling.

AI DECLARATION

During the preparation of this article, we used Anthropic's Claude Sonnet 4.6 to refine the clarity, flow, concision, and English language of the original, manually drafted text; the model's suggestions were applied only after critical consideration. In the production of the dataset, we used Anthropic's Claude Sonnet 4.6 to (a) advise on meaning definitions within the dataset, again applying, modifying or rejecting suggestions on the basis of human decision-making; and (b) assist in flagging validation issues (esp. diagnosing the semantic match between synonyms, and between our definitions and the linked WN definitions). Finally, we used Anthropic's Claude Sonnet 4.6 and Claude Opus 4.8 as coding assistance when producing scripts for converting DISSILEX data between formats and for internal validation. AI was not used to directly generate or modify DISSILEX data in any way.

ACKNOWLEDGEMENTS

We would like to thank Tomáš Hampejs (T. H.), Petr Hanák (P. H.), Adam Mertel (A. M.), Ján Mertel (J. M.), and Davor Salihović (D. S.) for contributions to: the initial stages of data model development (T. H., A. M.); data collection (D. S.); data transformations from RethinkDB format (T. H.); the development of the InkVisitor environment (A. M., P. H., J. M.); and a script for counting user contributions (A. M.). We also thank all anonymous reviewers, and all contributors to WordNet, the Collaborative Interlingual Index (CILI), Latin WordNet, and LiLa.

FUNDING STATEMENT

This research received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme, grant agreement No. 101000442, project "Networks of Dissent: Computational Modelling of Dissident and Inquisitorial Cultures in Medieval Europe" (David Zbíral, Soňa Hudíková, Katia Riccardo, Katalin Suba, Stanislaw Banach, and Robert L. J. Shaw), and from the European Regional Development Fund, grant agreement No. CZ.02.01.01/00/22_008/0004595, project "Beyond Security: Role of Conflict in Resilience-Building" (Gideon Kotzé and publication fees). We also used computational resources provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

David Zbíral: conceptualization, supervision, investigation, data curation, writing – original draft, writing – review & editing.

Gideon Kotzé: conceptualization, supervision, investigation, data curation, software, formal analysis, writing – original draft, writing – review & editing.

Soňa Hudíková: investigation.

Katia Riccardo: investigation.

Katalin Suba: investigation.

Stanislaw Banach: investigation.

Robert L. J. Shaw: conceptualization, supervision, investigation, data curation, writing – review & editing.

AUTHOR AFFILIATIONS

David Zbíral orcid.org/0000-0002-7903-2069
Department for the Study of Religions, Faculty of Arts, Masaryk University, Brno, Czech Republic
Gideon Kotzé orcid.org/0000-0001-7318-2245
Department for the Study of Religions, Faculty of Arts, Masaryk University, Brno, Czech Republic
Soňa Hudíková orcid.org/0009-0000-3851-4195
Department for the Study of Religions, Faculty of Arts, Masaryk University, Brno, Czech Republic
Katia Riccardo orcid.org/0000-0002-1997-3588
Department for the Study of Religions, Faculty of Arts, Masaryk University, Brno, Czech Republic
Katalin Suba orcid.org/0009-0001-4580-6425
Department for the Study of Religions, Faculty of Arts, Masaryk University, Brno, Czech Republic
Stanislaw Banach orcid.org/0009-0002-6725-5991
Department for the Study of Religions, Faculty of Arts, Masaryk University, Brno, Czech Republic
Robert L. J. Shaw orcid.org/0000-0003-0164-8100
Department for the Study of Religions, Faculty of Arts, Masaryk University, Brno, Czech Republic

REFERENCES

Bond, F., Vossen, P., McCrae, J. P., & **Fellbaum, C.** (2016). CILI: The Collaborative Interlingual Index. In C. Fellbaum, P. Vossen, V. B. Mititelu, & C. Forascu (Eds.), *Proceedings of the 8th Global WordNet Conference (GWC)* (pp. 50–57). Global Wordnet Association. https://doi.org/10.18653/v1/2016.gwc-1.9

Corcoglioniti, F., Rospocher, M., Aprosio, A. P., & **Tonelli, S.** (2016). PreMOn: A Lemon Extension for Exposing Predicate Models as Linked Data. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 877–884). European Language Resources Association (ELRA). https://aclanthology.org/L16-1141/ (last accessed: 12 December 2025).

Du Cange, C. D. F. (1883). *Glossarium mediae et infimae latinitatis*. L. Favre. http://ducange.enc.sorbonne.fr/ (last accessed: 15 May 2026).

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. MIT Press (Ed.). https://doi.org/10.7551/mitpress/7287.001.0001

Francopoulo, G. (2013). *LMF Lexical Markup Framework*. ISTE Ltd. https://doi.org/10.1002/9781118712696

Mambrini, F., Passarotti, M., Litta, E., & **Moretti, G.** (2021). Interlinking valency frames and WordNet synsets in the LiLa knowledge base of linguistic resources for Latin. In M. Alam, P. Groth, V. De Boer, T. Pellegrini, H. J. Pandit, E. Montiel, V. Rodríguez Doncel, B. McGillivray, & A. Meroño-Peñuela (Eds.), *Studies on the Semantic Web*. IOS Press. https://doi.org/10.3233/SSW210032

McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P., & **Cimiano, P.** (2017). The Ontolex-Lemon model: Development and applications. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek, & V. Baisa (Eds.), *Electronic lexicography in the 21st century: Proceedings of eLex 2017 conference* (pp. 587–597). Lexical Computing CZ. https://john.mccr.ae/papers/mccrae2017ontolex.pdf (last accessed: 15 October 2026).

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*(11), 39–41. https://doi.org/10.1145/219717.219748

Minozzi, S. (2017). Latin WordNet, una rete di conoscenza semantica per il latino e alcune ipotesi di utilizzo nel campo dell'information retrieval. In P. Mastandrea (Ed.), *Strumenti digitali e collaborativi per le Scienze dell'Antichità*. Università Ca' Foscari Venezia, Italia. https://doi.org/10.14277/6969-182-9/ANT-14-10

Niermeyer, J. F., van de Kieft, C., & **Burgers, J. W. J.** (2002). *Mediae Latinitatis lexicon minus* (Vols. 1–2). Brill.

Passarotti, M., Saavedra, B. G., & **Onambele, C.** (2016). Latin Vallex. A treebank-based semantic valency lexicon for Latin. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 2599–2606). European Language Resources Association (ELRA). https://aclanthology.org/L16-1414/ (last accessed: 15 October 2026).

Quetglas, P. J. et al. (2021). *Glossarium Mediae Latinitatis Cataloniae digital*. Editorial CSIC. https://gmlc.imf.csic.es/glossarium/gmlc (last accessed: 15 October 2026).

RethinkDB. (n.d.). Retrieved 18 May 2026, from https://rethinkdb.com/ (last accessed: 18 May 2026).

Romary, L., & **Tasovac, T.** (2018, September). TEI Lex-0: A Target Format for TEI-Encoded Dictionaries and Lexical Resources. *TEI Conference and Members' Meeting*. https://inria.hal.science/hal-02265312

Zbíral, D., Hanák, P., Mertel, A., Mertel, J., Ondrejka, P., Hampejs, T., & **Shaw, R. L. J.** (2025). *InkVisitor 1.5* [Computer software]. https://github.com/DISSINET/InkVisitor/ (last accessed: 15 October 2026).

Zbíral, D., & **Shaw, R. L. J.** (2022). Hearing voices: Reapproaching medieval inquisition records. *Religions*, *13*(12), 1175. https://doi.org/10.3390/rel13121175

Zbíral, D., Shaw, R. L. J., Hampejs, T., & **Mertel, A.** (2022). Model the source first!: Towards Computer-Assisted Semantic Text Modelling and source criticism 2.0. *Zenodo*. https://doi.org/10.5281/zenodo.6963579 (last accessed: 15 October 2026).

Zbíral, D., Shaw, R. L. J., Hanák, P., Hampejs, T., & **Mertel, A.** (2026). *From texts to structured data: Building knowledge graphs through Computer-Assisted Semantic Text Modelling (CASTEMO)*. Masaryk University. https://docs.religionistika.phil.muni.cz/books/from-texts-to-structured-data-building-knowledge-graphs-through-computer-assisted-semantic-text-modelling-castemo (last accessed: 15 October 2026).

Zbíral, D., Sikk, K., & **Shaw, R. L. J.** (2025). Bridging digital history methods and source criticism: A research agenda for the study of inquisition records. *Acta Historica Tallinnensia*, *31*(1), 46–72. https://doi.org/10.3176/hist.2025.1.02

TO CITE THIS ARTICLE:
Zbíral, D., Kotzé, G., Hudíková, S., Riccardo, K., Suba, K., Banach, S., & Shaw, R. L. J. (2026). DISSILEX: A Lexico-Semantic Network and Valency Lexicon for Medieval Latin. *Journal of Open Humanities Data*, 12: 143, pp. 1–7. DOI: https://doi.org/10.5334/johd.637

Submitted: 22 August 2026
Accepted: 05 October 2026
Published: 03 November 2026

COPYRIGHT:
© 2026 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See https://creativecommons.org/licenses/by/4.0/.

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.