Linked Open Data for West Nilotic Languages: The NILOMORPH project

Matteo Pellegrini, Matthew Baerman, Oliver Bond

University of Surrey
Surrey Morphology Group
{matteo.pellegrini,m.baerman,o.bond}@surrey.ac.uk

Abstract

In this paper, we present the NILOMORPH project, that aims at describing the complex non-concatenative morphology of West Nilotic languages and reconstructing the dynamics of its evolution from a more straightforward concatenative system. The project adopts techniques from several methodologies and draws on many kinds of data displaying different formats, tagsets and conventions. Data are also multilingual, documenting different West Nilotic varieties, and multimodal, including also audio and video recordings. This makes the process of integration of these data particularly challenging. We first describe how the data can be converted to standard formats such as CLDF and Paralex, to achieve interoperability between resources of the same kind. We then discuss how they can be modelled as Linguistic Linked Open Data in the Resource Description Framework, reusing already existing vocabularies and defining new classes and properties to meet the needs of the project, to also achieve interoperability between resources of different kinds.

Keywords: West Nilotic, morphology, Linguistic Linked Open Data

1. Introduction and Motivation

The West Nilotic languages are spoken in an area that includes South Sudan, south-western Ethiopia, the north-east of the Democratic Republic of Congo, northern Uganda and south-western Kenya. A remarkable feature of some of those languages is that they mark multiple morphological distinctions simultaneously through modifications of phonological features of the stem vowel. For instance, in the Nuer lexeme for 'rain', case and number distinctions are signalled by contrasts in tone, length, phonation type and height, as in NOM.SG /bḛ́ː/ vs. GEN.SG /bêl/ vs. NOM.PL /bɛ̌ː/ vs.GEN.PL /bɛ̂ː/ (Baerman and Monich, 2021). The NILOMORPH project aims to reconstruct in detail how such a complex and cross-linguistically rare non-concatenative system might have emerged from the more straightforward concatenative one still preserved in other West Nilotic varieties. To do that, it applies techniques taken from many different methodologies, including field linguistics, acoustic analysis, experimental linguistics, computational simulations, typology, and the historical-comparative method.

Such methodological variety poses the challenge of how to deal with a wide range of data types (Section 2): lexical resources such as lexicons of inflected forms; textual resources, such as raw texts (sometimes with translations), glossed examples, or more structured tagged corpora; as well as material derived from language experiments. Data are necessarily multilingual (to allow for the application of the comparative method), and multimodal (since audio recordings and visual documentation may provide useful information unavailable elsewhere). It is therefore crucial that these multifarious data types be interoperable with each other, allowing for the extraction of information from different sources in a unified fashion.

A natural solution to this challenge is offered by the framework of Linguistic Linked Open Data (LLOD; Cimiano et al., 2020), that aims to make data FAIR – Findable, Accessible, Interoperable, Reusable (Wilkinson et al., 2016) by leveraging Semantic Web technologies. The backbone of this enterprise is the Resource Description Framework (RDF) data model (Lassila and Swick, 1999), where all items are treated as resources with their own Unique Resource Identifier (URI), information is expressed through triples that connect a subject (a resource) to an object (a resource, or some data about it) through a property (itself a resource), and the relation between items is reflected in a hierarchical structure of sub-classes and sub-properties. RDF data can be retrieved with the SPARQL Protocol and RDF Query Language (Prud'hommeaux and Seaborne, 2008), that allows for queries that efficiently extract variegated information from different sources. The use of standard vocabularies and ontologies also makes it possible to have interoperability between datasets created by different people for different projects.

To date, RDF LLOD have been released for an increasing number of languages, as documented in the LLOD cloud.[1] Of particular note are a handful of larger-scale projects that have been launched for the integration of resources of different kinds available for a single language, starting with LiLa for Latin (Passarotti et al., 2020), later followed by LiITA for Italian (Litta et al., 2024) and MOLOR for Old Irish (Fransen et al., 2024). Efforts have been made also to achieve interoperability between languages of the same family – e.g. Romance in the ALMA project (Tittel,

[1] https://linguistic-lod.org/.

Proceedings of the 10th Workshop on Linked Data in Linguistics (LDL-2026) @ LREC 2026, pages 80–86
12 May 2026.

2023) – and between data in different modalities (Menke et al., 2013).

We illustrate here our proposal for how to model the variegated data of the NILOMORPH project as RDF LLOD. Doing so will contribute to this framework by increasing its coverage to include several under-resourced languages that were previously not represented, including Nuer, Dinka, Shilluk and Thok Reel. It will also provide the community with a discussion of the specific characteristics of the project data, including multilinguality, multimodality, and coding of cognacy between elements in different related languages.

The paper is structured as follows. Section 2 details and exemplifies the different kinds of data. Section 3 presents the model that we propose to handle the data, discussing existing standards and vocabularies relevant to our needs. Section 4 concludes and highlights the next steps.

2. The Data

Firstly, there are lexical resources. These include both traditional bilingual dictionaries available from earlier documentation efforts, and more structured resources produced more recently. An example of the former is Kiggen's (1948) Nuer-English dictionary, with information on lexical category and other morphosyntactically relevant features (e.g. transitivity-based inflection classes), giving English translation(s) and translated usage examples, and possibly additional notes and comments. Entries also list principal parts (the set of inflected forms from which other paradigm cells can be inferred; Stump and Finkel, 2013). All this information has undergone OCR and is available in machine-readable format as an Excel spreadsheet. This has been supplemented with other classifications relevant for the project goals, e.g. flagging stem-final consonants relevant for identifying suspected cognate entries. As an example of the latter type of resource, for Nuer there is also a more structured paradigmatic lexicon which provides a larger list of inflected forms in IPA transcription, and other pieces of information, such as paradigmatic patterns of vowel length and quality alternations, along with audio recordings of different forms in context (Bond et al., 2020). All this information feeds an interactive website, and it is also available as an Excel spreadsheet. For lexical resources, a crucial aim of the project is the development of a database that consist of several paradigmatic lexicons for different varieties of West Nilotic languages, and the identification of cognate lexical items across languages, to enable the application of the comparative method for the reconstruction of Proto-West Nilotic morphology.

Secondly, there are textual resources. These include both written material[2] and audio recordings of spontaneous speech (e.g. Remijsen et al., 2014-2024). In many cases, only the raw text is available, with no linguistic annotation. In some cases, we also have English translations of those texts. In other cases, richer linguistic information is provided. For written material, these may be in the form of interlinear glosses added to the text. For audio recordings, phonological transcription, morpheme segmentation, translation and comments can be provided in .eaf format using the ELAN software. For textual resources, the project aims at enriching the information provided on audio recordings with annotations in .textgrid format using the speech analysis software Praat, and at building tokenised, lemmatised, PoS-tagged and possibly morphologically analysed corpora from the texts available.

In addition, data may also consist of materials prepared for language experiments of different kinds, for instance, production and perception experiments to determine the exact phonetic realisation of sound contrasts in the languages under investigation. One example is Remijsen et al. (2022), that provides .wav audio files of forms elicited from speakers of the Bor dialect of Dinka to investigate contrasts of voice quality (modal vs. breathy) and tone (low vs. high), together with .textgrid files of annotations made with Praat.

The project will make this wealth of information publicly accessible[3] on the web in machine-readable open formats (e.g. converting Excel spreadsheets to .csv tables), and connect the different datasets with each other, and possibly with other datasets available on the web for other languages, using standard vocabularies and ontologies in the RDF data model, thus achieving the five stars in the grading system of open data recommended by Tim Berners-Lee.[4]

3. The Model

Standard formats have been proposed for many of the kinds of resources we are dealing with. The Cross Linguistic Data Formats (CLDF, Forkel et al., 2018) initiative offers standardised ways to represent in tabular format the data most often gathered while documenting and describing languages. To do that, it builds on the recommendations of the CSV on the Web W3C Working Group, namely the Model for Tabular Data and Metadata on the Web[5] and the Metadata Vocabulary for Tabular Data.[6] CLDF datasets

[2] E.g. the Nuer storybooks at https://www.nuerlexicon.com/books.php.

[3] Material protected by copyright will be standardised and converted to RDF but not released openly.

[4] https://www.w3.org/DesignIssues/LinkedData.html.

[5] https://www.w3.org/TR/tabular-data-model/.

[6] https://www.w3.org/TR/tabular-metadata/.

consist of several .csv files that refer to each other by means of unique identifiers, following the best practices of relational database management. This allows users to provide different pieces of information, avoiding redundancy. Different modules are introduced for specific kinds of dataset, namely wordlists, dictionaries, structure datasets (such as the World Atlas of Language Structures, Haspelmath et al., 2005), parallel texts in different languages, and corpora. Different components (tables) are defined to express information on different items, such as forms, lexical entries, senses, and languages. Different columns are defined to express different pieces of information on those items, e.g. the language, part-of-speech and headword of lexical entries.

Among the resources that can be represented as CLDF lexicons, there are many of the ones relevant to our project (such as dictionaries, parallel texts and corpora), but a very important one is missing, namely, paradigmatic lexicons that document inflected forms that appear in different cells of lexemes. The Paralex initiative[7] fills this gap, offering a standard format for such resources. To do that, it draws on many of the fundamental principles of CLDF, such as the use of multiple tables with a relational structure and the coding of metadata in a machine-readable format, although for the latter it relies on the frictionless[8] framework for data packages, rather than on CSVW. Consequently, different tables are defined for items on which information is provided, such as forms, lexemes, cells; and different columns are defined for the pieces of information that are provided, such as the orthographic and phonetic/phonological transcription, cell and lexeme of a form. The Paralex standard format is also well equipped for the coding of phonetic and phonological aspects that are crucial for the aims of NILOMORPH. Those can be expressed in a separate table whose lines contain segments and whose columns contain either the phonological features that define them or links to standardised repositories – such as CLTS (Anderson et al., 2018) and PHOIBLE (Moran and McCloy, 2019).

Adopting standard formats such as CLDF and Paralex introduces quite strict requirements that allow for a great degree of interoperability between resources of the same kind, e.g. between CLDF datasets of different languages but with the same module and components. Such interoperability is not only structural (pertaining to the data formats and languages), but also semantic (pertaining to the actual categories and values used). This makes it possible to develop tools that can be seamlessly applied to resources complying with the standards in order to perform fully comparable linguistic analysis, e.g. entropy-based measurements of predictability with the Qumin toolkit (Beniamine, 2018) on Paralex lexicons. Consequently, such standards have been increasingly adopted by the research communities working on these kinds of data.[9]

To achieve interoperability between resources of the same kind available for the different languages involved in NILOMORPH, and with other resources of that kind available for other languages, we start from the legacy data mentioned in Section 2. These come in various formats and use resource-specific conventions and tagsets. We convert them into the standard formats of CLDF (for traditional dictionaries and texts, glosses and additional annotations), and Paralex (for paradigmatic lexicons). In doing so, there is the potential to enrich the vocabularies of these formats to account for the complexity of West Nilotic data, e.g. the different phonological dimensions involved in vowel alternations.

However, *per se* this does not address another project requirement, i.e. interoperability between resources of different kinds, which may include ones that have been created for different purposes. RDF LLOD technology is the natural way to satisfy this requirement. Indeed, both CLDF and Paralex provide built-in ontologies that introduce RDF classes and properties for the tables and columns of the formats, and define them as sub-classes and sub-properties of resources in existing ontologies for the relevant domains. This allows seamless conversion to RDF LLOD, ensuring interoperability with a wider scope. For additional tables and properties defined for our data, it will be necessary to define additional mappings to RDF classes and properties, either by reusing existing vocabularies directly, or by extending the ontologies.

We now turn to the LLOD vocabularies of interest for the data of our project. Crucial to any work on language data are repositories of linguistic terminology, both general-purpose ones, such as GOLD (Farrar and Langendoen, 2003), and others for specific use cases, e.g. Lexinfo for lexical resources (Cimiano et al., 2011), or OLiA for annotations (Chiarcos and Sukhareva, 2015).

For lexical resources, currently the *de facto* standard is the OntoLex vocabulary (McCrae et al., 2017), which provides classes for lexical entries, their form, and meaning (senses and concepts), as well as properties that relate those items to each other. Additional modules are provided for more specific aspects, such as *decomp* for the decomposition of lexical entries, and *lime* for metadata of lexical resources. Other

[7] https://www.paralex-standard.org/.

[8] https://framework.frictionlessdata.io/.

[9] See the CLDF and Paralex communities on Zenodo (https://zenodo.org/communities/paralex, https://zenodo.org/communities/cldf-datasets, respectively) for a list of the datasets released in the two formats.

modules are also being developed, such as *morph* for morphological information (Chiarcos et al., 2022c) and *FrAC* for frequency and corpus attestations (Chiarcos et al., 2022a). Given NILOMORPH's focus, representations of morphological data using *morph* are particularly relevant, such as work by Chiarcos et al. (2022b) on German and Ionov and Rosner (2023) on Maltese.

For textual resources, the NLP Interchange Format (NIF, Hellmann et al., 2013) provides a way to unambiguously identify portions of texts on the web, while POWLA (Chiarcos, 2012) allows for the addition of separate layers on which different annotation levels can be operated on the same text. Furthermore, the Open Annotation vocabulary reifies annotations with their own class, introducing properties relating them to the annotated item, and to the content of the annotation. The latter vocabulary is of particular interest for this project because it supports annotation of different media types. This allows us to handle audio recordings and their annotation. For interlinear glossed text, the ligt framework (Chiarcos and Ionov, 2019) is explicitly designed to model this kind of data, and its tools are available for an automatic conversion from glossed examples in CLDF format (Ionov, 2025).

The Paralex ontology has been designed to allow for conversion to OntoLex compliant lexical resources, providing mappings to classes and properties of the OntoLex model and its modules where relevant, alongside categories from lexinfo and recommendations to reuse the Open Annotation model for the coding of morphological features, and mapping to other standard vocabularies such as the one of the Unimorph project (Kirov et al., 2018), via an OLiA annotation model (Chiarcos et al., 2020). The parallel release of data as both Paralex and OntoLex lexicons has already been tested on existing resources (Pellegrini et al., 2025). Consequently, its application to NILOMORPH data should not require any further modelling effort, except for resource-specific tables and columns introduced for the purposes of the project.

However, the CLDF ontology only defines mappings to the GOLD ontology for data.[10] Because of the status of OntoLex as a *de facto* standard for the release of lexical resources, at the stage of conversion to RDF we plan to also introduce an OntoLex-compliant modelling, defining entries, forms, senses and concepts as belonging to the corresponding OntoLex classes, and using OntoLex properties to relate them. Similarly, for textual resources we will supplement the classes and properties of the CLDF ontology with mappings to the vocabularies mentioned above for different pieces of information – NIF and POWLA for texts, ligt for glossed examples.

Furthermore, more elaborate resources do not lend themselves easily to release as either CLDF or Paralex datasets. For instance, this is the case of richer and more structured corpora that can be produced by tokenising and annotating the available texts. Such resources can be released as RDF LLOD directly, following best practices defined by Cimiano et al., (2020) and applied, for instance, to Latin corpora in the LiLa knowledge base (Mambrini and Passarotti, 2019).

Finally, once all resources are available in standardised formats and as RDF LLOD, they need to be connected with each other. Projects such as LiLa, LiITA and MOLOR follow a common strategy to achieve this, by connecting entries and lexical resources and tokens of textual resources to the respective lemma, which is defined as a sub-class of forms in the OntoLex vocabulary. Such a strategy can be applied also to our project, but it only covers the case where we have different resources for the same language. However, given the scope of the project, it is necessary to also reach interoperability between different West Nilotic languages. For this, cognacy relations between items will be identified. These can be coded using the cognates and cognate sets components of CLDF on the one hand, and through the lemonEty vocabulary for etymological information in OntoLex lexicons (Khan, 2018) as RDF LLOD on the other. Etymological information will play a pivotal role in allowing for interoperability between all the datasets relevant to the NILOMORPH project.

4. Conclusions and Future Work

We have presented here the NILOMORPH project, which aims to describe and reconstruct the complex non-concatenative morphology of the West Nilotic family, thus elaborating and creating data on under-resourced languages. We have described the multifarious data handled by the project and discussed how existing standards and ontologies are reused to model them as RDF LLOD, also showing how specific aspects of the data require us to enrich and extend these vocabularies.

We plan next to move on to other steps of the LLOD generation process as outlined by Cimiano et al. (2020), namely, i) data generation, ii) linking, iii) publication and iv) exploitation. We will i) convert the heterogeneous legacy formats into CLDF and Paralex datasets, and then convert them to RDF, ii) connect the different resources together through their citation forms and etymons and add links to other resources, iii) publish the data and metadata with open licences in

[10] For metadata, links to dcat and dcterms are also provided.

repositories that assign them a DOI, and provide access to them in a triplestore, also offering tabular and graph visualisations with lodview and lodlive,[11] and finally iv) develop tools to exploit all these interconnected resources. All these steps will ultimately be helpful to achieve the goals of the project concerning linguistic analysis of the languages involved and the reconstruction of the proto-West Nilotic system.

5. Bibliographical References

Cormac Anderson, Tiago Tresoldi, Thiago Chacon, Anne-Maria Fehn, Mary Walworth, Robert Forkel, and Johann-Mattis List. 2018. A cross-linguistic database of phonetic transcription systems. *Yearbook of the Poznan Linguistic Meeting*, 4(1):21–53.

Matthew Baerman and Irina Monich. 2021. Paradigmatic saturation in Nuer. *Language*, 97(3):e257–e275.

Sacha Beniamine. 2018. *Classifications flexionnelles. Étude quantitative des structures de paradigmes*. PhD Thesis, Université Sorbonne Paris Cité-Université Paris Diderot (Paris 7).

Christian Chiarcos. 2012. POWLA: Modeling linguistic corpora in OWL/DL. In Elena Simperl, Philipp Cimiano, Axel Polleres, Oscar Corcho, and Valentina Presutti, editors, *The Semantic Web: Research and Applications. 9th Extended Semantic Web Conference, ESWC 2012, Heraklion, Crete, Greece, May 27-31, 2012, Proceedings*, pages 225–239. Springer, Dordrecht.

Christian Chiarcos, Elena-Simona Apostol, Besim Kabashi, and Ciprian-Octavian Truică. 2022a. Modelling Frequency, Attestation, and Corpus-Based Information with OntoLex-FrAC. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4018–4027, Gyeongju. International Committee on Computational Linguistics.

Christian Chiarcos, Christian Fäth, and Frank Abromeit. 2020. Annotation Interoperability for the Post-ISOCat Era. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5668–5677. European Language Resources Association, Marseille.

Christian Chiarcos, Christian Fäth, and Maxim Ionov. 2022b. Unifying Morphology Resources with OntoLex-morph. A Case Study in German. In Calzolari, Nicoletta, Béchet, Frédéric, Blache, Philippe, Choukri, Khalid, Cieri, Christopher, Declerck, Thierry, Goggi, Sara, Isahara, Hitoshi, Maegaard, Bente, Mariani, Joseph, Mazo, Hélène, Odijk, Jan, and Piperidis, Stelios, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4842–4850, Marseille. European Language Resources Association.

Christian Chiarcos, Katerina Gkirtzou, Fahad Khan, Penny Labropoulou, Marco Passarotti, and Matteo Pellegrini. 2022c. Computational Morphology with OntoLex-Morph. In Thierry Declerck, John P. McCrae, Elena Montiel, Christian Chiarcos, and Maxim Ionov, editors, *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference*, pages 78–86, Marseille. European Language Resources Association.

Christian Chiarcos and Maxim Ionov. 2019. Ligt: An LLOD-native vocabulary for representing interlinear glossed text as RDF. In Maria Eskevich, Gerard de Melo, Christian Fäth, John P. McCrae, Paul Buitelaar, Christian Chiarcos, Bettina Klimek, and Milan Dojchinovski, editors, *2nd Conference on Language, Data and Knowledge (LDK 2019). LDK 2019, May 20–23, 2019, Leipzig, Germany*, page 3:1-3:15. Dagstuhl, Wadern.

Christian Chiarcos and Maria Sukhareva. 2015. OLiA – Ontologies of Linguistic Annotation. *Semantic Web*, 6(4):379–386.

Philipp Cimiano, Paul Buitelaar, John McCrae, and Michael Sintek. 2011. LexInfo: A declarative model for the lexicon-ontology interface. *Journal of Web Semantics*, 9(1):29–51.

Philipp Cimiano, Christian Chiarcos, John P. McCrae, and Jorge Gracia. 2020. *Linguistic linked data*. Springer, Dordrecht.

[11] https://github.com/LodLive/LodView.

Scott Farrar and D. Terence Langendoen. 2003. A Linguistic Ontology for the Semantic Web. *GLOT international*, 7(3):97–100.

Robert Forkel, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. 2018. Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific data*, 5(1):1–10.

Theodorus Fransen, Cormac Anderson, Sacha Beniamine, and Marco Passarotti. 2024. The MOLOR Lemma Bank: A New LLOD Resource for Old Irish. In Christian Chiarcos, Katerina Gkirtzou, Maxim Ionov, Fahad Khan, John P. McCrae, Elena Montiel Ponsoda, Patricia Martin Chozas, editors, *Proceedings of the 9th Workshop on Linked Data in Linguistics@ LREC-COLING 2024*, pages 37–43, Torino. ELRA and ICCL.

Martin Haspelmath, Matthew Dryer, David Gil, and Bernard Comrie, editors. 2005. *The World Atlas of Language Structures*. Oxford University Press, Oxford.

Sebastian Hellmann, Jens Lehmann, Sören Auer, and Martin Brümmer. 2013. Integrating NLP using linked data. In Harith Alani, Lalana Kagal, Achille Fokoue, Paul Groth, Chris Biemann, Josiane Xavier Parreira, Lola Aroyo, Natasha Noy, Chris Welty, and Krzysztof Janowicz, editors, *The Semantic Web - ISWC 2013. 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings*, pages 98–113. Springer, DordrechtNLP.

Maxim Ionov. 2025. Ligt: Towards an Ecosystem for Managing Interlinear Glossed Texts with Linguistic Linked Data. In Mehwish Alam, Andon Tchechmedjiev, Jorge Gracia, Dagmar Gromann, Maria Pia di Buono, Johanna Monti, and Maxim Ionov, editors, *Proceedings of the 5th Conference on Language, Data and Knowledge*, pages 100–105. UniorPress, Napoli.

Maxim Ionov and Michael Rosner. 2023. Beyond Concatenative Morphology: Applying OntoLex-Morph to Maltese. In Sara Carvalho, Anas Fahd Khan, Ana Ostroški Anić, Blerina Spahiu, Jorge Gracia, John P. McCrae, Dagmar Gromann, Barbara Heinisch, and Ana Salgado, editors, *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 385–391, Vienna. NOVA CLUNL.

Fahad Khan. 2018. Towards the Representation of Etymological Data on the Semantic Web. *Information*, 9(12):304.

Jan Kiggen. 1948. *Nuer-English Dictionary*. Missiehuis, Steyl bij Tegelen.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sabrina J. Mielke, Arya McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. UniMorph 2.0: Universal Morphology. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, Japan.

Ora Lassila and Ralph R. Swick. 1999. Resource Description Framework (RDF) Model and Syntax Specification.

Eleonora Litta, Marco Passarotti, Paolo Brasolin, Giovanni Moretti, Valerio Basile, Andrea Di Fabio, and Cristina Bosco. 2024. The Lemma Bank of the LiITA Knowledge Base of Interoperable Resources for Italian. In Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 517–522, Pisa, Italy. CEUR Workshop Proceedings.

Francesco Mambrini and Marco Passarotti. 2019. Linked Open Treebanks. Interlinking Syntactically Annotated Corpora in the LiLa Knowledge Base of Linguistic Resources for Latin.

John P. McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The Ontolex-Lemon model: development and applications. In Iztok Kosem, Carole Tiberius, Miloš Jakubíček, Jelena Kallas, Simon Krek, and Vít Baisa *Proceedings of eLex 2017 conference*, pages 19–21.

Peter Menke, John Philip McCrae, and Philipp Cimiano. 2013. Releasing Multimodal Data as Linguistic Linked Open Data: An Experience Report. In Christian Chiarcos, Philipp Cimiano, Thierry Declerck, and John P. McCrae *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and linking lexicons, terminologies and other*

language data, pages 44–52, Pisa, Italy. Association for Computational Linguistics.

Marco Passarotti, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, and Rachele Sprugnoli. 2020. Interlinking through lemmas. The lexical collection of the LiLa Knowledge Base of linguistic resources for Latin. *Studi e Saggi Linguistici*, 58(1):177–212.

Matteo Pellegrini, Marco Passarotti, Francesco Mambrini, and Giovanni Moretti. 2025. PrinParLat: a Lexicon of Principal Parts of Latin Verbs Linked to the LiLa Knowledge Base. *Language Resources and Evaluation*:1–41.

Eric Prud'hommeaux and Andy Seaborne. 2008. SPARQL Query Language for RDF.

Gregory T. Stump and Raphael A. Finkel. 2013. *Morphological Typology: From Word to Paradigm*. Cambridge University Press, Cambridge.

Sabine Tittel. 2023. Ceci n'est pas un dictionnaire. Adding and Extending Lexicographical Data of Medieval Romance Languages to and through a Multilingual Lexico-Ontological Project. In Marek Medveď, Michal Měchura, Carole Tiberius, Iztok Kosem, Jelena Kallas, Miloš Jakubíček, and Simon Krek *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference*. Lexical Computing CZ sro, Brno.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, and Philip E. Bourne. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9.

6. Language Resource References

Oliver Bond, Tatiana Reid, Irina Monich, and Matthew Baerman. 2020. Nuer Lexicon. www.nuerlexicon.com. Accessed 6 February 2026.

Steven Moran and Daniel McCloy. 2019. PHOIBLE 2.0.

Bert Remijsen, Mirella L. Blum, and Jon Pen de Ngong. 2022. A dataset on Voice Quality, Tone, and Vowel Quality in the Bor South dialect of Dinka.

Bert Remijsen, Otto Gwado Ayoker, and Maria Bocay Onak. 2014-2024. Collection of Shilluk narratives and songs.