

Sacha Beniamine, Jules Bouton, Mae Carroll, Matteo Pellegrini, Cormac Anderson, Erich Round, Olivier Bonami, Dunstan Brown, Matías Guzmán Naranjo, Borja Herce, Andrea D. Sims, Helen Sims-Williams, Farrell Ackerman, Greville Corbett, Robert Malouf, Jeff Parker, and Theodorus Fransen

Data sharing and standardization in Linguistics

Introducing the DeAR principles and Paralex

Abstract: Linguistic typology stands to gain significantly from advances in the use of extremely large datasets. However, our ability to secure these gains will depend on the availability of machine-readable data that is precise and comparable. Here we identify the challenges and opportunities ahead, relating to the quality, longevity, and (re-)usability of linguistic data in typology. Then in response, we introduce the DeAR principles (Decentralized, Automatically verified, Revisable), designed to guide and assist researchers to create diverse, high-resolution and robust datasets. We demonstrate the DeAR principles in action through the example of Paralex, a data standard (i.e., set of sci-

Sacha Beniamine, Corresponding Author: s.beniamine@surrey.ac.uk

Sacha Beniamine, Cormac Anderson, Erich Round, Helen Sims-Williams, Greville Corbett, Surrey Morphology Group, University of Surrey, Guildford, United Kingdom

Mae Carroll, Department of Linguistics and Applied Linguistics, University of Melbourne, Melbourne, Australia

Matteo Pellegrini, Theodorus Fransen, Università Cattolica del Sacro Cuore, Milan, Italy

Jules Bouton, Olivier Bonami, Université Paris Cité, CNRS, Laboratoire de linguistique formelle, Paris, France

Borja Herce, Institute of Romance Languages and Institute for the Interdisciplinary Study of Language Evolution, University of Zurich, Zurich, Switzerland

Matías Guzmán Naranjo, Department of Linguistic, Literary and Aesthetic Studies, Universitetet i Bergen, Bergen, Norway

Dunstan Brown, Department of Language and Linguistic Science, University of York, York, United Kingdom

Andrea D. Sims, Department of Linguistics, The Ohio State University, Columbus OH, USA

Farrell Ackerman, University of California San Diego, San Diego, USA

Robert Malouf, Department of Linguistics and Asian/Middle Eastern Languages, San Diego State University, San Diego, USA

Jeff Parker, Department of Linguistics, Brigham Young University, Provo UT, USA

Erich Round, University of Queensland, St Lucia, Brisbane, Australia

entific conventions) developed collaboratively for lexicons of morphologically inflected forms. Our proposals aim to foster a more resilient and equitable infrastructure for the future of linguistic research.

1 Introduction

Linguistic typology is an inherently comparative endeavour. Its success largely hinges upon documenting a rich sample of the world’s languages, collected as high quality data that is both reliable and accessible. Moreover, as new results in typology are increasingly derived with the aid of quantitative methods, the availability of datasets that are machine-readable becomes correspondingly indispensable and integral to successfully testing statistical relationships, revealing fine-grained patterns of variation, and crafting an ever-clearer picture of typological distributions. However, in our field we are confronted by significant challenges, both in creating these mission-critical datasets, and, as we wish to emphasize here, preserving them. As a consequence of the relentless obsolescence of digital technologies and the impermanence of our institutions, digital datasets, which have been created laboriously over recent decades, face the acute risk of becoming inaccessible and even lost.

Some of these issues demand attention at a collective level. For example, data creators face a range of challenges which simply cannot be solved individually, but which require collaborative resolution. Moreover, notwithstanding invaluable efforts to develop guiding principles such as ‘good practices’ (Wilkinson et al. 2016, Carroll et al. 2020) and systems of shared conventions, often termed ‘standardization’ (see for example Forkel et al. 2018), it remains difficult for many, if not most, present-day linguists to marshal the technical expertise required to carry out the necessary data management, or at times even simply to discover and use existing datasets in their research.

In this article, we identify and analyse some recurrent, real-world issues that we as linguists collectively face in the compilation, publication, (re-)use and preservation of datasets that can be used for typological research (Section 2). That examination of linguistic practice leads us to a discussion of existing data management principles (Section 3), motivating the introduction of the novel DeAR principles (Section 4), which leads us to concrete, practicable solutions (Section 5). To exemplify these principles in action, we present Paralex (Section 6), a set of conventions—i.e., a ‘standard’—for the creation, use, communication and preservation of linguistic datasets that describe lexicons of morphologically inflected word forms. Embodying the DeAR principles, the Paralex standard has resulted from a sustained international collaboration

between morphologists as we sought to address these issues of coordinated scientific action tied to typologically wide-ranging linguistic data. We conclude by summarizing our main observations and suggestions and offering some speculations about future directions.

2 Problems inherent to typological datasets

We start by identifying five kinds of problems that accompany the use and production of datasets in typology: the scope of their coverage (Section 2.1), the commensurability of data comparisons (Section 2.2), the internal consistency of the formats and structures of the data representations (Section 2.3), the challenge of developing durable datasets (Section 2.4), and the technical skills demanded for the creation and use of datasets (Section 2.5).

2.1 Problems of coverage

Typological datasets come in various forms: structural linguistic feature sets, such as the World Atlas of Language Structures (WALS, Dryer & Haspelmath 2013) and Grambank (Skirgård et al. 2023); both raw and annotated machine-readable corpora, such as the Universal Dependencies corpora (Nivre et al. 2020); and lexical datasets, including the comparative word lists widely used in historical linguistics (Swadesh 1952, List et al. 2022, 2023) as well as large-scale, more intricately structured compilations, annotated with words' properties and their interrelationships (see Zeller, Šnajder & Padó 2013 for German; Vidra et al. 2019 for Czech; Namer et al. 2023 for French). The novel standard format, Paralex, which we introduce in section 6, is designed to be used for the release of datasets of this last type; it provides conventions for documenting morphological lexicons of inflected word forms.

Structural feature sets like WALS and Grambank lend themselves well to global-scale surveying and hypothesis testing of universals and statistical relationships among structural phenomena in the world's languages at large. However, the features they can address are restricted both in number (195 in Grambank) and in their degree of detail, which places an upper bound on their utility for fine-grained comparisons. Corpora and large-scale lexical datasets on the other hand offer a more fine-grained view of the data, but with a diminished and skewed typological coverage. As Elsner & Beniamine (2024: p. 3) observe, “The more information is needed, the more the analyst

must fall back on scarcer resources which tend to push towards a familiar set of well-resourced European languages” (see also Casillas et al. 2025). For example, European languages still account for most of the data covered in the Universal Dependency datasets (Nivre et al. 2020). Languages of the Americas, Africa, Oceania and South-Eastern Asia are heavily under-represented, even though these are areas with high linguistic diversity.

This tension between level of detail and cross-linguistic coverage of data is a well-known problem in typology. Resolving it will require sustained, large-scale scientific cooperation, and so it is worth asking how that cooperation can be supported. One step forward will be to ensure that our datasets are visibly disseminated, readily usable, and mutually compatible for comparative purposes. That in turn can be supported by well-designed conventions for data sharing and formatting.

2.2 Problems of commensurability

Most linguistic data do not reflect simple objective observations, but rather are the product of *analysis*. Words transcribed phonemically in the International Phonetic Alphabet (IPA), for example, depend on a cascade of choices and decisions made by the analyst about the nature of the phonological system and wordhood in the language (Chao 1934, Mosteiro Romero & Blasi 2025, Round 2023, Anderson et al. 2023, Tallman & Auderset 2023). Even for an analytical device as widely used as morphemic segmentation, there is still little consensus on a ‘correct’ or ‘best’ way to carry it out (Spencer 2012, Beniamine & Guzmán Naranjo 2021, Carroll & Beniamine 2025). These facts about analysis lead to a familiar dilemma when typological datasets are compiled from multiple sources. On the one hand, the sources are written by experts who have typically devoted years of consideration to their analysis, so it seems only reasonable to defer to their expertise. On the other hand, linguists working in different traditions, guided by different priorities and theoretical conceptions, are unlikely to have converged on the same set of choices that underpin, and then shape, their analyses. Finding solutions to issues of comparability between sources remains a central challenge for linguistic typology at large. Indeed, comparability and reproducibility are issues that face linguists more generally (Schweinberger & Haugh 2025) and the scientific community as a whole (Baker 2016).

Additionally, incommensurability can arise as information from sources is re-presented in a secondary dataset. Distinctions made in sources might be neutralized, while diversity in the degree of clarity and (un)certainty ex-

pressed in sources might be homogenized. For machine-readable datasets, it is important to implement practical solutions to manage and reduce the incommensurability arising from dataset compilation.

2.3 Problems of consistent presentation

When humans read datasets, they bring to the task a remarkable flexibility and resourcefulness. Experts in any scholarly domain are able to mentally repair inconsistencies within and across datasets by drawing on contextual information and background knowledge. Computers, by contrast, typically demand a high degree of systematicity and precision while being heedless to anything not stated literally.

As an illustration of the problems this can cause, take the original Oxford Romance Database (Maiden 2010). This dataset documents verbal paradigms for 73 Romance varieties, annotated for cognacy at the level of both lexemes and paradigm cells. Within the database, Romance verbs are coded for their corresponding main verbal Latin inflection classes. In the original database, this information was mainly indexed in the form of Roman numerals (I to IV), but occasionally Arabic numerals or full words were used, and various further conventions indicated irregular entries. Although these variations do not hinder human understanding, they inhibit full machine readability, as by default computers would take “third”, “3”, “3rd”, and “III” to be distinct values, to the detriment of analyses that rely on them. For the Oxford Romance Database, fixing these issues involved both semi-automated and manual repairs (Beniamine, Maiden & Round 2019). However, the good news is that methods exist for minimizing such problems in the first place, by using automated tools and guidelines during the construction of datasets, as we discuss in sections below.

2.4 Problems of dataset durability

In the brief period since computerized resources have existed, data has already begun to be lost.¹ We often share data through *websites*,² which serve as user interfaces designed for browsing. These may also offer some amount of data visualization or an interface for queries. In many cases, however, it is not possible to download the full underlying data. Because websites are often created by contracted external web developers towards the conclusion of research projects, there is typically limited or no funding beyond the term of the project to maintain the site and an absence of designated responsibility for ensuring the security and long term survival of the data (Windhouwer & Dimitriadis 2008). Moreover, if there is a way for the researchers to update data, these updates often become buggy as the software ages and is not adequately maintained.

As an illustration, we take again the case of the Oxford Romance Database (Maiden 2010). Ten years after its creation, the data persisted solely in the site's database, which had ceased to be maintained. The institutional knowledge and access to its backend had been forgotten, and the data itself was therefore in jeopardy of being lost. Concerted conservation work was necessary to save it, involving scraping the website's data, which was then reorganized, augmented and standardized to Cross-Linguistic Data Formats (CLDF, Forkel et al. 2018). The revised and more durable Romance Verbal Inflection Dataset 2.0 (Beniamine, Maiden & Round 2019) has been released openly.

As a matter of good science, it is imperative to secure the long-term preservation of data. The CLDF standard (Forkel et al. 2018) provides a positive example of the road forward: both novel and retro-standardized datasets are typically provided as full raw data. Visually appealing showcase websites are then generated automatically by a generic toolkit (clld, Forkel, Bank & Rzymiski 2019). Dataset changes are recorded using software which tracks the history of the project. Successive published versions are stored on an archival site (e.g. Zenodo, see for example Lexibank, List et al. 2022).

¹ One example is that of the Surrey Morphological Complexity Database, released in 2015 as a Flash web application. Flash support was dropped from all major browsers by 2021, making the data entirely unavailable. Significant work was necessary to rescue the underlying data, now available at <http://dx.doi.org/10.15126/SMG.23/1>

² By an unfortunate lexical shift, the academic community misleadingly refers to these showcase websites as *databases*. This has led to confusion between actual data and its visual rendition (the website). The confusion is increased by the fact that most websites' back-end rely on actual *databases* stored on the sites' servers, but distinct from the researcher's data.

2.5 Problems of access to technical skills

Producing machine-readable data, publishing it in sustainable ways, ensuring a high degree of coherence across datasets: these goals are essential for the future of linguistic typology. However, while much of linguistics today combines qualitative and quantitative approaches, the achievement of these goals hinges on technical skills that may often not be taught as part of linguistics degrees. In some institutions, research engineers exist who can assist linguistic researchers with technical choices and infrastructure, but this is not a given. Since it is unrealistic to expect all linguists to be trained in the relevant skills, there is significant value in making the tasks of data management more accessible and linguist-friendly.

3 Principles for improved research

Linguistics is not the only discipline faced with problems such as those described in Section 1, and we are certainly not the first to consider them (Berez-Kroeker et al. 2022). This enables us to benefit from progress that has already been made. In this section, we mention existing initiatives which provide a pathway to solutions for linguistic typology.

3.1 Open data

Open research is a framework for making research data, methodology and findings freely available, via an *Open source* legal license.³ It facilitates the dissemination and replication of results, and re-use of research data, while retaining sensitivity to ethical and privacy concerns and other legal restrictions. Open-source licenses can be applied to data, leading to Open Data.

Two Open Research conditions provide support for better citation practices and the enhanced traceability of research data (Bird & Simons 2003). Attribution (BY) places a legal obligation on re-use, to give credit to the original resource creators. Share-alike (SA) ensures that innovations derived from the original resource must be shared under the same terms as the original.

³ See the Open Source Software initiative, <https://opensource.org/>

3.2 FAIR principles

A major bottleneck of linguistic research is the workload involved in collating data from disparate and hard-to-find sources. The FAIR principles (Wilkinson et al. 2016) are intended to decrease bottlenecks of this kind. The four principles are that data should be:

- **Findable:** Data can only be of use if it can be found by its users.
- **Accessible:** Data are only of use if they are accessible by users.
- **Interoperable:** Data are interoperable if they use the same conventions and may be combined to increase their breadth or depth.
- **Reusable:** Data are most useful when they can be re-used to support novel research, beyond the creators' own usage and initial motivations for collecting it.

3.3 CARE principles

Shifting from a focus on data users, the CARE Principles for Indigenous Data Governance (Carroll et al. 2020) focus on the interests of language communities whose languages are described in the datasets. They call for:

- **Collective Benefit:** Data ecosystems shall be designed and function in ways that enable Indigenous Peoples to derive benefit from the data.
- **Authority to Control:** Indigenous people must have control over how data are shared and how their culture is represented and identified.
- **Responsibility:** Data creators must be accountable for how the data are used.
- **Ethics:** Indigenous Peoples' rights and wellbeing should be the primary concern at all stages of the data life cycle and across the data ecosystem.

The CARE principles can drive us to ensure that datasets are maximally beneficial not only to researchers but to the participating language communities.

3.4 Linked data

Linked Data (Berners-Lee 2006, Heath & Bizer 2011) achieves interoperability and reusability by extending the infrastructure of the World Wide Web. The core idea is to interlink not only web documents, but also databases and even individual data points.

3.5 Principles arising from linguistic database efforts

Two sets of principles which arose in the context of linguistic database elaboration and standardization are those of AUTOTYP (Bickel, Balthasar 2002, Witzlack-Makarevich et al. 2022) and CLDF (Forkel et al. 2018). The AUTOTYP principles of modularity and connectivity call for information to be distributed across files and linked together using identifiers, that is, unique names which allow us to refer to datasets and data points unambiguously. The CLDF standard is based on four design principles. Three of these focus on data interoperability, longevity and reusability: where possible, entities should be referenced using established identifiers; data should be encoded as UTF-8 text files; compatibility with existing tools, standards, and practice should be kept in mind.

Above, we saw that CARE principles target the needs of language communities. Open data, FAIR principles, Linked Data, and the three CLDF principles just mentioned target the needs of data users. A fourth CLDF principle is notable in its intent to benefit linguists as data creators: data should be editable manually and via software that the typical linguist can learn to use correctly. Facilitating data creation and maintenance by researchers is a crucial matter, and it is the motivation behind a new set of principles which we introduce next.

4 Introducing the DeAR principles

Beyond users and speakers, language data also need to be planned in ways that facilitate the work of the researchers who produce and maintain them. Thus, we introduce the DeAR principles: Decentralized, Automatically verified, and Revisable. These build on the principles introduced already.

4.1 Decentralized

When faced with the challenge of creating a large number of datasets in a standardized form, one attractive possibility is for a single, large team with ample funding to consolidate large amounts of data in a single database. This process has proven useful and efficient (see Rzymiski et al. 2020, List et al. 2022 as examples). However, initiatives of this type demand funding at a scale which is difficult to secure, and when secured, difficult to maintain beyond the term of the project. Another risk of this centralized strategy is for choices to

be made to fit a specific project or theoretical orientation, reducing reusability for other purposes and different perspectives.

The alternative is to coordinate a pragmatic, decentralized approach to dataset elaboration (adopted, for example, by the Universal Dependencies project, Nivre et al. 2016, 2020). In addition to distributing demands over multiple research groups, this approach ensures that responsibilities for long-term maintenance are distributed as well, as distinct creators are in charge of each dataset. Within a decentralized approach, each dataset can be made by domain experts, which enhances the scientific community’s confidence in the data quality. Moreover, the decentralized publication of datasets makes each contribution prominent on its own, which facilitates recognition for independent data creators.

Of course, this strategy presents its own demands: in lieu of a central authority prioritizing data creation and directing analytical and coding choices, there is a need for collective agreement on sets of conventions (standards) imbued with sufficient flexibility to serve the entire field while providing sufficient constraints to make data interoperable. We suggest that this need for collective effort is in fact a central reason why the resulting decentralized strategy to dataset creation is currently the most promising framework for typological data organization and use.

4.2 Automatically verified

Problems of maintaining data consistency (Section 2.3) might, at first glance, appear only to be magnified in the decentralized approach to dataset elaboration, as consistency needs to be maintained across many projects, participants and iterative updates. However, a simple technical solution exists, in the form of automated verifications. Given a specific set of expectations about the contents of a dataset, including expectations that follow from the use of a standard, it is possible to check automatically whether these expectations are upheld. This process is called *validation*. For instance, if a dataset contains vernacular words from a specific language, and a statement is provided about the permissible phonemes in the language, then automated checks can be run upon every update, generating an alert that locates any unexpected symbols. If the dataset adheres to a standard that establishes Unicode conventions to be used for IPA symbols (e.g. the Cross-Linguistic Transcription Systems (CLTS) standard, Anderson et al. 2018), then a suite of checks associated with the standard can be triggered automatically. (See Section 5.1 for methods of stating these expectations about data, using metadata.) The

strategy of validation can be applied to data points of any kind in linguistic datasets, including syntactic structures, ISO or glottocodes, and Linked Data (Section 3.4) identifiers.

Automatic validation tools, created by more technically proficient members of the community, can be made accessible to less-technical users. This serves the needs of data creators by facilitating the curation of high-quality datasets (e.g. CLDFbench, Forkel & List 2020). We describe our successful use of this approach in Section 6.

4.3 Revisable

Linguistic datasets occasionally require updating, whether due to improvements in knowledge, the creation of an expanded version, or for the correction of inadvertent errors. This is the problem of *revisability* (see also CLDF’s principle H (Forkel et al. 2018)). It can be resolved by creating datasets in standardized formats, and then publishing them not as locked websites (see Section 2.4), but with tools that automatically generate user-friendly views of the data, for instance webpages or PDF documents. These views can then be re-generated at will from the standardized data, making it simple to revise data incrementally and ensure that the published version is up-to-date.

5 Solutions to our data problems

How can the principles introduced in Sections 3 and 4 be developed in order to address the data problems identified in Section 2? For a maximally practicable pathway to applying these principles to data problems facing typology, we envision an approach, similar to CLDF (Forkel et al. 2018), that relies on the use of metadata (Section 5.1), standardization (Section 5.2), normalization (Section 5.3) and continuous development (Section 5.4).

5.1 Metadata

Metadata are any information about a dataset that are not directly part of the data but whose purpose is to facilitate the interpretation of the data. A distinction can be drawn between descriptive and administrative metadata on the one hand, and structural metadata on the other hand.

Descriptive and *administrative* metadata relate to entire datasets. They include the identity of the authors (forename, last name, institution, identifiers such as ORCID, etc.), the title and description of the dataset, how it should be cited, the list of its sources or of related datasets, its persistent identifier (DOI), the licence under which it is shared, relevant keywords, as well as version numbers. These facilitate search queries and so contribute to dataset accessibility. Archival repositories, such as Zenodo,⁴ Open Science Framework (OSF)⁵ or the Endangered Language Archive (ELAR)⁶ either collect this kind of metadata through online forms (Zenodo and OSF) or require users to upload them in a specific format (ELAR). The information is then stored within the archival platform, displayed in a human-friendly format, and queryable automatically.

Structural metadata describe the component parts of the dataset, the relationships among them, and their possible contents, formats and significance. For example, assuming a dataset made of a few tables of data (where a ‘table’ is a spreadsheet-like subset of data organized in rows and columns), one could ask questions such as:

- What tables are present in the dataset?
- How are they written to disk?
- What does each represent?
- What are the relations between tables?
- What columns are present in each of the tables?
- What do these columns mean?
- What content is expected in each column (e.g. numbers, text in a specific script, true/false values encoded as 1s and 0s, missing data encoded as blanks)?
- Which columns serve as identifiers?

These are questions which humans can be good at answering just from inspection, based on expert knowledge and experience, though they might hesitate on some of the details. Computers, however, need this information spelt out explicitly. Structural metadata perform that function.

Standards already exist for writing metadata formally. A widely-used example is the Frictionless standard (Fowler, Barratt & Walsh 2018), which arranges metadata into attribute-value matrices, expressed in the JSON data

⁴ <https://zenodo.org/>

⁵ <https://osf.io/>

⁶ <https://www.elararchive.org/>

format. Figure 1 shows an excerpt of the descriptive metadata from the Lat-InfLexi dataset (Pellegrini & Passarotti 2018), in frictionless compliant JSON. Figure 2 shows an excerpt of the structural metadata for the same dataset, describing a column of a sounds table which codes the distinctive feature \pm consonantal.

```

1  {
2    "name": "latinflexi",
3    "title": "LatInfLexi",
4    "licenses": [
5      {
6        "name": "CC BY-SA 4.0",
7        "title": "Creative Commons Attribution-
8                ShareAlike 4.0 International",
9        "path": "https://creativecommons.org/licenses/by-sa/4.0/"
10     }
11   ],
12   "contributors": [
13     {
14       "title": "Matteo Pellegrini",
15       "role": "author",
16       "organization": "Università Cattolica del Sacro Cuore"
17     }
18   ]
19 }

```

Fig. 1: Excerpt of descriptive json metadata extracted from LatinFlexi.

Readers of this article who have endured the filling out of reams of descriptive and administrative metadata might recoil from the prospect of furnishing yet more, structural metadata for their datasets. We agree. Typing such metadata manually is not a good use of our time and is prone to error. Instead, metadata is best generated automatically when possible. Adhering to standards aids the process, as it enables structural metadata to be inferred and filled in automatically.

5.2 Agreed-upon community conventions, a.k.a. Standards

Standards are agreements on conventions. The most well-known example of a standard might be the metric system. Measuring entities using the metric system frees us from the need to continually convert between specialized units. Moreover, a standard can be set up in a way that is maximally practical: e.g. the metric system is decimal, making calculations easy.

```

1  {
2    "name": "consonantal",
3    "title": "Whether the sound displays
4             the feature consonantal",
5    "description": "Binary feature (+/-) indicating whether the
6                   segment displays the feature consonantal",
7    "type": "boolean",
8    "trueValues": [ "+" ],
9    "falseValues": [ "-" ]
10 }

```

Fig. 2: Excerpt of structural json metadata extracted from LatinFlexi.

Standards may arise either from planned agreements or *de facto*. The main source of planned standards is the International Organization for Standardization (ISO), which edits standards for almost all aspects of modern infrastructure (e.g. the 2 and 3-letter languages codes, ISO 639). In contrast, *de facto* standards often emerge through the widespread use of a tool.

Standards are useful for expressing specific data points, such as sequences of phonemes, codes for countries, or gloss abbreviations. Table 1 describes some common data points in linguistic datasets and possible standards for them. At a larger scale, standards can specify how to organize entire datasets, and provide conventions to do so. Some prominent examples of standards meant for entire datasets are the CLDF standard (for tabular data such as word lists and dictionaries, see Forkel et al. 2018) and the Text Encoding Initiative (for structured XML data such as glossed examples or dictionaries, see TEI Consortium 2025). Table 2 summarizes a few types of linguistic datasets, and existing standardized formats for them.

Standards also exist for metadata (see section 5.1). Common terms for describing metadata were standardized as the Dublin Core Metadata Terms (DCMT, or ISO 15836). In the case of XML documents, data and metadata are written and standardized conjointly. For tabular data, metadata is often provided as separate JSON files, with specifications given either as part of data standards (as in CLDF), or separately (as with Fowler, Barratt & Walsh 2018). Although some are described in ordinary prose (the Leipzig Glossing Rules), standards can themselves be described in standardized ways (see Bradner 1997), which facilitates unambiguous interpretation.

The main benefit of standards is to provide unambiguous semantics, which ensures interoperability. For example, in a gloss, does the abbreviation IMP stand for “imperfect”, “imperfective”, “impersonal”, or “imperative”? If we know that data creators followed the Leipzig Glossing Rules, the answer is trouble-free: it is imperative. The use of standards is crucial for long-term

Data points	Some standards & conventions	Reference
Parts of Speech	LexInfo	Cimiano et al. 2011
Gloss abbreviations	Leipzig glossing rules	Comrie, Haspelmath & Bickel 2008
	Corbett's conventions	Corbett 2000, 2006
	Creissels' conventions	Creissels 2006
	Bernard Fradin's synthesis of the above	http://www.llf.cnrs.fr/fr/node/60
	Universal Dependency Tagset	Nivre et al. 2016, 2020
	UniMorph	Sylak-Glassman et al. 2015
Interlinear glosses	Leipzig glossing rules	Comrie, Haspelmath & Bickel 2008
Languages	Glottocodes	Hammarström & Forkel 2022
	ISO 639	ISO Central Secretary 2023
	IETF BCP 47	Phillips & Davis 2006, 2009
Countries	ISO 3166	ISO Central Secretary 2020
Locations on earth	ISO 19111	ISO Central Secretary 2019b
Phonemes	IPA	International Phonetic Association 1999
	CLTS	Anderson et al. 2018, List et al. 2024
Meanings	Concepticon	List, Cysouw & Forkel 2016
Dates	ISO 8601	ISO Central Secretary 2019a
Writing systems	ISO 15924	ISO Central Secretary 2022

Tab. 1: Examples of common data points in linguistic datasets, and existing standards or conventions for these.

Data nature	Formats	Standards	References
Corpora	TSV	CONLL-U	Nivre et al. 2016
	CSV	CLDF	Forkel et al. 2018: TextCorpus
	XML	TEI	TEI Consortium 2025: Chap. 16
Dictionaries	CSV	CLDF	Forkel et al. 2018: Dictionary
	XML	TEI	TEI Consortium 2025: Chap. 10
Glossed examples	CSV	CLDF	Forkel et al. 2018: ExampleTable
	XML	TEI	TEI Consortium 2025: Chap. 3.4
	TXT	Shoebox MDF	
Word lists	CSV	CLDF	Forkel et al. 2018: Wordlist
	XML	TEI	TEI Consortium 2025: Chap. 10
Speech transcriptions	XML	TEI	TEI Consortium 2025: Chap. 8

Tab. 2: Examples of common linguistic dataset, and existing standards for these. CLDF entries refer to the CLDF RDF namespace (<https://cldf.cldf.org/v1.0/terms.rdf#>)

intelligibility, as it is more likely that knowledge of a single common standard will be retained in the future than that of numerous independent and partly overlapping conventions. In addition, standards avoid duplication of work by relieving data creators and users of some of the responsibility of inventing, documenting, maintaining, interpreting, and converting conventions.

There are some costs associated with the use of standards. First, data creators and users need to learn specific norms. Second, formal compatibility across datasets may still mask fundamental differences: it is important to remember that interoperability is not the same thing as comparability. Finally, standards which do not successfully cover all the needs of data creators might lead them to use the same term to mean different things. However, these risks can all be mitigated by making sure standards are produced with and for their users, and updated as needed. Overall, the benefits created by standardization far outweigh their costs.

5.3 Normalization

Normalization refers to two processes of data tidying: terminological normalization and database normalization.

5.3.1 Terminological normalization

Terminological normalization consists in ensuring data points are always written in the same way and belong to homogeneous sets, with no spurious duplicates referring to the same entity, for example variation in case (e.g. “VERB” vs “verb”), writing system (“III” vs “third”), abbreviation (“VERB” vs “V”), etc. Although normalized values may seem an obvious desideratum, they can be genuinely difficult to achieve by sheer human discipline, without technical constraints on input or validation. Hence, normalization is greatly aided by the presence of exhaustive structural metadata, which can be used as the basis of checks on the validity of each data point.⁷

⁷ For instance, by enumerating a closed class of valid codes, by linking to a catalog or vocabulary which in turn lists valid codes, or enforcing valid patterns, such as a regular expression pattern for string-valued data.

5.3.2 Database normalization

For tabular databases, especially *relational databases* where multiple tables are linked together, data normalization (Codd 1970, Wilson et al. 2017) aims to eliminate data redundancy and facilitate its maintenance (see also the notion of *tidy data*, Wickham 2014). We take as an example Table 3, which presents three languages, with a few main pieces of information on each.

name	family	family size	639-3	country
English	Indo-European	586	eng	UK, USA, Canada, Australia, Ireland, New Zealand
Ewe	Niger–Congo	1,540	ewe	Ghana, Togo, Benin
Aguna	Niger–Congo	1,540	aug	Togo, Benin
Malay	Austronesian	1,274	msa	Malaysia

Tab. 3: Table with information on three languages

In Table 3, rows are identified by the ‘name’ column. However, language names might present variations, contain spaces or special characters. To improve machine readability, distinct, unique row identifiers can be added. They are conventionally given as a first column. A second issue is that of data duplication. In Table 3, there is a dependency between the “family size” column and the “family” column, as the first describes not languages, but their families. Each time a language from the same family would occur in the table, the same number would need to be repeated. Duplication introduces problems for two reasons: it leads to excessive storage space, and it multiplies the risks for a change in one location but not the other to lead to inconsistent data. In our example, one risks updating a family size in some places, but not all rows where the family occurs. A third issue in Table 3 is the presence of multiple values in some cells. For example, a language can be spoken in more than one country, and a country may have more than a single language.

Table 4 illustrates a partial solution. It introduces row identifiers, separate tables for each entity described (languages, countries, families), and a table for the many-to-many relation between countries and languages. Rows of the country/language relation table can be read as “the language X is spoken in the country Y”.

In a normalized database, each table describes homogeneous things (entities or relations), represented in rows and labelled with identifiers (primary

language tables			
lang_id	name	family_id	639-3
lg-eng	English	fam-1-ie	eng
lg-ewe	Ewe	fam-2-nc	ewe
lg-aug	Aguna	fam-2-nc	aug
lg-msa	Malay	fam-3-aus	msa

country table	
country_id	name
country-1	United Kingdom
country-2	USA
country-3	Australia
country-4	Ireland
country-5	New Zealand
country-6	Ghana
country-7	Togo
country-8	Benin
country-9	Malaysia

family table		
family_id	family	size
fam-1-ie	Indo-European	586
fam-2-nc	Niger-Congo	1,540
fam-3-aus	Austronesian	1,274

country/languages relations table		
relation_id	country_id	language_id
rel-1	country-1	lg-eng
rel-2	country-2	lg-eng
rel-3	country-3	lg-eng
rel-4	country-4	lg-eng
rel-5	country-5	lg-eng
rel-6	country-6	lg-ewe
rel-7	country-7	lg-ewe
rel-8	country-8	lg-ewe
rel-7	country-7	lg-aug
rel-8	country-8	lg-aug
rel-9	country-9	lg-msa

Tab. 4: Normalization of table 3

keys). No rows or columns are duplicated. Each column states a single piece of information about the row (not about values in other columns), and each cell contains a single value (it is atomic).

Normalized databases ensure no duplication, and minimize the risks of introducing inconsistencies. Yet for linguists, Table 3 may be more intuitive to browse. In the context of research data, it is crucial to find a balance between browsability and normalization. In Section 6, we show with the example of Paralex that we can benefit from some of the practices of data normalization, while letting researchers read and write the data conveniently.

5.4 Continuous Deployment

The last solution we highlight is a common practice in software development, called *continuous deployment*.

Usually, software code exists in two simultaneous environments: one that the developers are working on (the *development* environment) and one accessible to users (the *production* environment). Continuous deployment streamlines the deployment from the development to the production environment. It verifies each change to ensure it will not degrade the software. If updates pass all verification, they are seamlessly delivered to end users.

Research data too involves two separate environments. On the one hand, the researcher's own data is a constant work in progress. On the other hand, its published representation (a website, archive, or pdf document) is meant to be more stable and is often generated through extensive additional manual work. Thereafter, it may become infeasible to propagate updates from the researcher's data: the two environments become entirely separate. As we saw in Section 2.4, over time this can lead to data loss. By adapting practices from continuous deployment, we can seamlessly generate public presentations of our research data (following the revisable principle, see Section 4.3). Because this comes with the risk of introducing errors in the public facing representations, a crucial step is to introduce automated verification to guarantee data quality.

Figure 3 illustrates the steps of such a pipeline. In order to synchronize and save all changes, researchers can use versioning systems. The most extensively used such versioning software is *git*, which works by labelling every incremental change, and synchronizing sets of changes across personal copies and servers. Using versioning systems provides much more precise and powerful ways to save work than cloud synchronization. Moreover, these systems can easily be

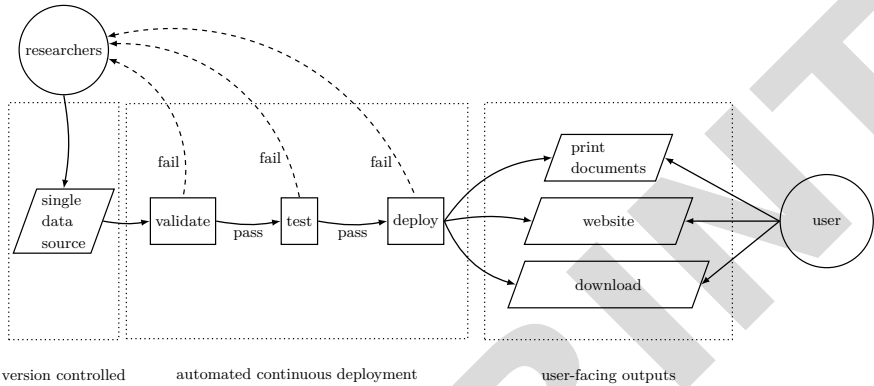


Fig. 3: Continuous deployment for research data: the researcher edits a single data source, which is automatically validated, tested, and deployed to user-facing outputs. Failure of any step rejects the update, and requires intervention from the researcher. Validation, testing and deployment can be run automatically each time the data are updated.

integrated with automated verification and deployment.⁸ In the context of research data, verifications can be broadly conceived as a two step process: validation and testing. **Validation** consists in checking the formal integrity of the data, and its conformity with what is declared by the metadata. For example, all rows of a table must have the same number of columns, ISO 639 codes must be current, official ones. If validation fails, the update is rejected with a message explaining why, and the researcher can make a new update. Validation can be done with generic tools associated with the standard(s) used. Sometimes, it is useful to add dataset-specific verifications. This is **testing**, for which dataset creators need to write their own custom scripts. Again, if the changes do not pass the tests, they are rejected with a message why and are back in the hands of the researcher. If they pass the tests, then the update can be seamlessly **deployed** to user-facing elements such as websites, or downloadable archives.

Usage of metadata and standards, normalization of the data and continuous deployments are all means to a single end: producing high quality, future-proof data with low maintenance costs, on a shoestring.

⁸ See for example gitlab pipelines or github actions.

6 The Paralex standard for inflected lexicons

We now present Paralex, a data standard we have developed for inflected lexicons. Prior to the establishment of Paralex, an ad-hoc format had emerged, with paradigm tables written as Excel spreadsheets, comma-separated value (CSV) or tab-separated values (TSV) tables, shared between researchers mostly by email. This practice led to poor documentation, little formal consistency, and obstacles to the findability, accessibility, interoperability or reusability of our data. In response, we decided to collectively devise a standard.

With the Paralex standard, we strive to provide data which is FAIR, so it can be used automatically (Section 3.2), CARE, so it respects and empowers language communities (Section 3.3), and DeAR, so it supports a good data ecosystem for dataset creators (Section 4). Paralex draws inspiration from the Cross-Linguistic Data Formats standard (CLDF, Section 3.5), and adheres to a similar philosophy and design principles, but is not based on this platform. The Paralex standard is formally documented in full at <https://www.paralex-standard.org>.

In this section, we discuss inflected lexicons of paradigms (Section 6.1) and possible formats for their tabular representation (Section 6.2). We then describe the component tables in a Paralex dataset (Section 6.3), the accompanying metadata (Section 6.4) and documentation (Section 6.5), as well as a Linked Data ontology which increases inter-operability (Section 6.6), and the Paralex ecosystem of automated tools (Section 6.7).

6.1 Inflected lexicons

An inflected lexicon is a dataset that contains words not just in citation form but in all of their inflected forms, providing a full inflectional paradigm for each lexeme. Inflected lexicons may document just a few lexemes, or as many as several thousand. The inflectional paradigms they contain are invaluable for investigations of inflectional morphology, and serve as a foundational resource for data-driven studies. They can support both traditional and computational research methods, for investigations focused both on individual systems and on cross-linguistic comparison. For language description, inflected lexicons are a powerful complement to the ‘Boasian Trilogy’ of grammar, lexicon, and text collection (Evans & Dench 2006), as they document morphological patterns with significantly more detail and completeness than typically is feasible in

descriptive grammars or dictionaries. Created by expert linguists, they provide high-resolution data capturing regularities, exceptions and contradictions in detail. This degree of completeness, although notionally possible to include in a grammar or dictionary, is most often omitted due to practical constraints.

A prominent set of inflected lexicons which exists already is the UniMorph lexicons (Batsuren et al. 2022). These lexicons for the most part have been extracted automatically from entries in Wiktionary (www.wiktionary.org). They have been used primarily in the field of Natural Language Processing, and to some extent in linguistic typology. However, the UniMorph lexicons are limited. For each inflected word form, they provide its lexeme, paradigm cell (following a standardized tagset, Sylak-Glassman et al. 2015) and orthographic form. For sophisticated linguistic research, it is desirable to add at minimum a phonological representation. More ambitiously, a comprehensive resource should take us beyond the basic triad of lexeme, cell description and inflected wordform, to include phenomena such as defectivity and overabundance, suppletion, inherent features, inflection classes, usage patterns, frequency, variation, and analytical choices that stand behind the data. By providing conventions for each of these dimensions, the Paralex standard aims to meet this ambition.

6.2 Tabular formats for paradigms

Under the hood, an inflected lexicon is composed of a set of tables. An important design decision for inflected lexicons is the choice of how to represent paradigms. As linguists will be aware, paradigms are very often presented as tables, and these can be organized according to a variety of conventional formats (Corbett 2013).

When presenting the paradigm of a single lexeme, linguists will often use a format as illustrated in Table 5a, in which rows represent the values of one morphosyntactic feature (such as *CASE*) and columns the values of another (such as *NUMBER*). This basic design plan takes advantage of the multi-dimensional structure of paradigms and has the benefits of offering an intuitive, easy-to-read layout that lends itself well to the printed page.

The single-paradigm format is effective for the task of presenting a small amount of data to the human eye, but suffers drawbacks as a format for data storage and sharing. First, if there are multiple lexemes to be described, a separate table is needed for each. This rapidly becomes unwieldy. Second, there is no obvious strategy for indicating additional information about each form: How is it pronounced? How frequent is it? Was it provided by a specific

(a) Single paradigm tables

	SINGULAR	PLURAL
NOM	rosa	rosae
VOC	rosa	rosae
ACC	rosam	rosās
GEN	rosae	rosārum
DAT	rosae	rosīs
ABL	rosā	rosīs

	SINGULAR	PLURAL
NOM	dominus	dominī
VOC	domine	dominī
ACC	dominum	dominōs
GEN	dominī	dominōrum
DAT	dominō	dominīs
ABL	dominō	dominīs

(c) Long form table

form_id	cell	lexeme	orth_form	gender
f1	NOM.SG	rosa	rosa	F
f2	VOC.SG	rosa	rosa	F
f3	ACC.SG	rosa	rosam	F
f4	GEN.SG	rosa	rosae	F
f5	DAT.SG	rosa	rosae	F
f6	ABL.SG	rosa	rosā	F
f7	NOM.PL	rosa	rosae	F
f8	VOC.PL	rosa	rosae	F
f9	ACC.PL	rosa	rosās	F
f10	GEN.PL	rosa	rosārum	F
f11	DAT.PL	rosa	rosīs	F
f12	ABL.PL	rosa	rosīs	F
f13	NOM.SG	dominus	dominus	M
f14	VOC.SG	dominus	domine	M
f15	ACC.SG	dominus	dominum	M
...	

(b) Wide form table (Stump & Finkel 2013's *plat*)

lemma	NOM.SG	VOC.SG	ACC.SG	GEN.SG	DAT.SG	ABL.SG	NOM.PL	...
ROSA	rosa	rosa	rosam	rosae	rosae	rosā	rosae	...
DOMINUS	dominus	domine	dominum	dominī	dominō	dominō	dominī	...

Tab. 5: Paradigm formats, illustrated with two Latin nouns (Pellegrini & Passarotti 2018).

consultant? Was it generated automatically? In common practice, a certain amount of additional information can be conveyed in publications, through the shading of cells or with other annotations, but this is at best a strategy for data presentation, rather than for the rigorous cataloging of information. Third, the single-paradigm format assumes the paradigm to have a canonical structure, with each cell (i.e., each combination of morphosyntactic feature values) represented by one and only one inflected form. However, departures from canonicity are frequent. Lexemes may be defective (having missing forms, see Corbett 2015, Sims 2015), or cells may be overabundant (having more than one form, see Thornton 2019). In the typology of inflectional systems, non-canonicity is far from exceptional, and within individual systems it can be pervasive (for instance in the plural forms of Estonian nouns), but even if it were not, it is important for our datasets to be capable of recording exceptional cases with fidelity.

A second format for presenting paradigm data is the *wide format*, also called a *plat* (Stump & Finkel 2013), illustrated in Table 5b. In this format, each column has a label, presented in Table 5b at the very top. Each row then represents a lexeme. The first column, *lemma*, contains a distinct label for each lexeme. The lemma label functions as a unique identifier (Section 3.4) which can be used to refer to the lexeme easily and unambiguously. The remainder of the columns contain the lexeme’s paradigm. This plat format has enjoyed some success among morphologists as a practical structure for machine-readable inflected lexicons, in part because filling in a lexeme’s paradigm is easy. It is the input format required by the Principal Parts Analyzer (Stump & Finkel 2013) and early versions of Qumín (Beniamine 2018), and a number of lexicons had their first releases in this format (see e.g. Flexique version 1, Bonami, Caron & Plancq 2014). Using a plat addresses the first of our three problems identified above, since multiple lexemes can be represented in the rows of a single table. However, the two other issues (inability to add further information; assumption of canonicity) remain unresolved.

These problems are solved by the use of a third format, the *long format*, illustrated in Table 5c. This is the format mandated by the Paralex standard, in conformity with the principles of data normalization (Section 5.3). In long format, each row is a record corresponding to a single inflected wordform. The first column, *form_id*, is a unique identifier for the wordform. This is followed at a minimum by columns labeled *cell* and *lexeme*, and one of *orth_form* (as in Table 5c) or *phon_form*, so that the representation of a wordform is minimally a triad of a cell specification, a lexeme and a form (as in datasets from UniMorph), plus a unique identifier. However, this is only the minimum. In long format, the information recorded about each wordform can be freely ex-

panded by adding more columns: for example, Table 5c also contains gender information. The long format also readily accommodates paradigms which are not canonical. For example, an overabundant paradigm cell can be straightforwardly represented on multiple rows, with each of those rows representing one of the cell's wordforms. Finally, unlike the plat format, rows do not become excessively long. Thus, all four of the problems flagged above are solved through the use of the long form. For further discussion on wide versus long form for linguistic data, see Forkel et al. (2018). Finally, long format is often assumed by software designed to work with paradigm data (like modern versions of Qumín, Beniamine & Bouton 2025).

The advantages just mentioned make the long format well suited for storing and exchanging paradigm data. For human readers, however, it can be less intuitive than the plat or single-paradigm formats. This is a good reason for regarding Paralex primarily as a standard for data exchange and archiving, not one intended directly for data display. On the other hand, because Paralex is well suited to computational tasks, it provides a basis from which to generate additional formats automatically that are more conducive to human viewing (see Section 6.7, on the Paralex tool ecosystem, below).

6.3 Component tables in a Paralex dataset

The beating heart of a Paralex dataset is a central table, described just above in Section 6.2, which represents information about paradigms in long format, with one record per wordform: this is the *forms* table. Supporting that is a network of additional tables, which, together with the *forms* table, comprise an overall *tabular dataset*. The component tables of this tabular dataset are linked to one another through the use of unique identifiers. In this section, we describe each table and how it can link to others. For a full, technical account, we refer the reader to the online documentation and specifications at <https://www.paralex-standard.org>.

6.3.1 Relational schema

Each table in a Paralex dataset is dedicated to a distinct function. That function is to document one kind of entity which is involved in inflection: the paradigm forms themselves, the lexemes which the inflected forms belong to, the paradigm cells for which lexemes inflect, the *feature-values* which define the paradigm cells, the sounds which comprise phonological forms, and

the graphemes which comprise orthographic forms. A frequencies table can record nuanced frequency measurements, and a tags table can record flexibly defined category labels, giving researchers the capacity to extend the information they wish to document. For each of these tables, the Paralex standard provides a set of pre-defined columns. Figure 4 shows a few of these for each table. The Paralex structure is designed to be adaptable to researchers' needs, and to be lightweight. In the design of standards, it is common to distinguish between components that are mandatory, recommended or optional (Bradner 1997). In Paralex, only the forms table is mandatory. Only the sounds, cells and feature-values tables are recommended. All other tables are optional. Custom tables and custom columns can be added on a case-by-case basis as researchers find necessary.

The tables in a Paralex dataset are linked together through unique identifiers, which is to say, a Paralex dataset follows a relational model (Section 3.5). These identifiers are used whenever a record in one table needs to refer to a record in another (recall that in each table, a record is stored as one of the table's rows). To introduce some technical terminology, an identifier like *form_id* in the forms table is called a *primary key*. The function of a primary key is to label the row that it appears on. A second use of identifiers is as a *foreign key*, to refer to a record somewhere else in the dataset, perhaps in the same table or in another table. The foreign key makes this reference by citing the primary key of the record it is referring to. In this way, a primary key and a matching foreign key function together to establish two ends of a tether which links two records to one another, within or across tables. This kind of relation is called a *foreign key* relation, and it is indicated in Figure 4 by a solid line. For example, in the forms table, the values in the lexeme column are foreign keys which refer to primary key identifiers, *lexeme_id*, in the lexeme table.⁹

A second type of relation enables us to declare that the values in one column (for instance, the phonologically represented wordforms in the *phon_form* column of the forms table) are comprised of sequences of entities from another table (for instance, the sounds stored in the sounds table). We call these relations *vocabulary* relations (the metaphor being that one table provides the 'vocabulary' from which items in some other column are composed). Vocabulary relations are particularly useful for automatically validating (Section 4.2) that the contents of a column are well-formed. Vocabulary relations are indicated in Figure 4 by dashed lines.

⁹ Note that this naming pattern is distinct from that of CLDF, where ID in LanguageTables are pointed at by *Language_ID* in other tables.

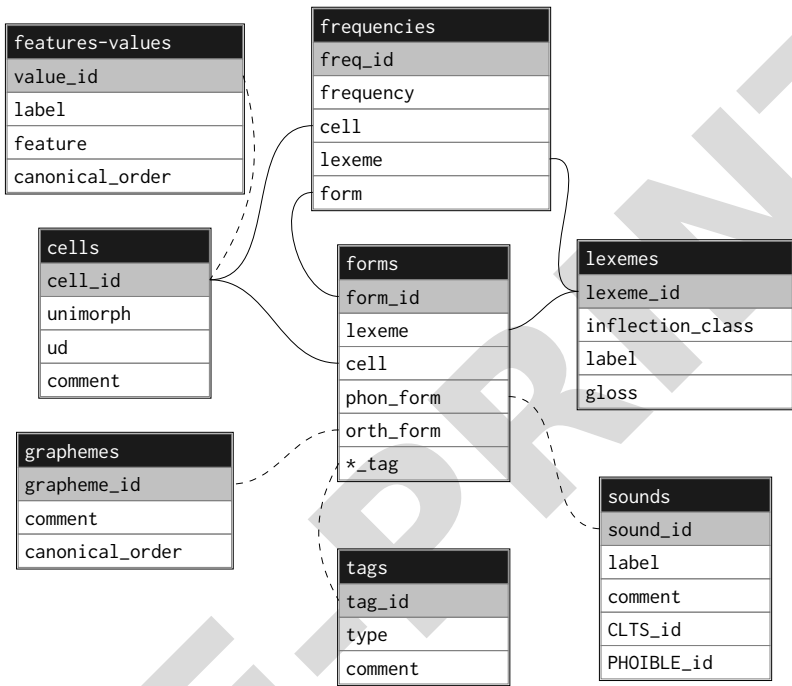


Fig. 4: Relation between tables of a Paralex dataset. Only some main columns are given. Shading indicate *primary* keys (identifiers), plain lines indicate *foreign* key relationships, dashed lines indicate *vocabulary* relationships.

form_id	lexeme	cell	orth_form	phon_form	overabundance_tag
f1	dream	pst	dreamt	d r ɛ m t	t-form;irreg
f2	dream	pst	dreamed	d r iː m d	d-form
f3	learn	pst	learned	l ɜː n d	d-form
f4	learn	pst	learnt	l ɜː n t	t-form

Tab. 6: Small example of overabundance in English past tenses.

tag_id	tag_column_name	comment
d-form	overabundance_tag	past in -ed
t-form	overabundance_tag	past in -t
irreg	overabundance_tag	irregular past with stem alternation

Tab. 7: A short tags table for the forms in Table 6

6.3.2 The forms table

The forms table was introduced briefly in Section 6.2 above. Here we provide more detail. Each row of the forms table documents a single inflected form. Only four columns are mandatory: a primary key identifier (`form_id`); foreign key identifiers (`cell`, `lexeme`) that refer to a cell and a lexeme; and some representation of the wordform, either orthographic (`orth_form`) or phonological (`phon_form`). Table 6 constitutes a valid small forms table. Phonological forms are given as sequences of phonological sounds separated by spaces, e.g. “*d r ɛ m t*”.¹⁰

To support the documentation of non-canonical phenomena, some additional conventions are provided. Defective entries (Sims 2015) must be explicitly specified (they are coded `#DEF#` following an already existing de-facto standard, see Bonami, Caron & Plancq 2014, Pellegrini & Passarotti 2018). Declaring a cell to be defective is distinct from simply not documenting a form. (Choosing not to document a form indicates merely that the data is missing, without making any statement regarding the existence of the form.) Paralex also enables documentation of overabundance (Thornton 2019). Take, for example, the overabundant English verbs `DREAM` and `LEARN`; both *dreamt* and *dreamed* and *learnt* and *learned*, respectively, are possible past-tense forms. In the forms table, each possible past-tense form is documented on its own row, as shown in Table 6. Taking this further, it is notable that the forms *dreamt* and *learnt* on the one hand, and *dreamed* and *learned* on the other hand, have something in common: they use the same affixal strategy. It can often be useful to annotate regularities such as these (e.g. long and short variants of the plurals in Estonian declension, see Aigro & Vihman 2023), so that users can systematically filter them. The Paralex standard provides a method to carry out this kind of annotation, using *tags*. In Paralex, tags are researcher-defined category labels. Definitions of the tags are recorded in a single dedicated `tags` table, and the tags themselves are used to flag records in other tables, in specially reserved columns. A short example of a `tags` table is given in Table 7. Note that the primary key identifier there is `tag_id`, and the identifiers in that `tag_id` column are visible being used as foreign keys in Table 6, in the column `overabundance_tag`. Thus, the `overabundance_tag` column allows us to tag the wordforms in Table 6, while Table 7 documents the researcher’s intended meaning of the tags, in its `comment` column.

¹⁰ This convention is adopted from the CLDF standard, and ensures unambiguous segmentation of forms into phonemes.

The need to annotate sets of forms as having a common property may prove useful beyond just the case of overabundance. The Paralex standard provides a series of optional tag columns, in addition to `overabundance_tag`. A `defectiveness_tag` may mark sets of defective forms (for example, indicating pluralia tantum), a `variants_tag` may indicate speaker-level variation (geographical, dialectal, etc.), an `epistemic_tag` may indicate the epistemic status of a form (e.g. `manually_checked`, `controversial`, `uncertain`, or `attested`), an `analysis_tag` may serve to distinguish among competing analytic choices (such as when analyses have been taken from different sources).

6.3.3 The lexemes table

The lexemes table provides information at the level of whole lexemes. Each row documents a separate lexeme.¹¹ The sole mandatory column is `lexeme_id`. The Paralex standard specifies optional columns to document a citation form (`label`), inflection class, lexical frequency, etc. Table 8 gives a very short lexemes table for the two lexemes of Table 5c. Note that while a frequency column is available, more complex frequency relationships can be provided using instead the flexible frequency table described below.

<code>lexeme_id</code>	<code>label</code>	<code>inflection_class</code>	<code>POS</code>	<code>meaning</code>	<code>frequency</code>
lex-dominus	dominus	2	noun	master	10000
lex-rosa	rosa	1	noun	rose	6000

Tab. 8: Small example of a lexemes table for two Latin nouns.

6.3.4 The cells table and feature-values table

The rows of the `cells` table provide the full inventory of feature-value combinations for which lexemes can inflect (the paradigm cells), with each row documenting one cell. The one mandatory column is `cell_id`. Paralex speci-

¹¹ A significant scientific question for dataset creators is the decision of what exactly constitutes a lexeme. On this issue, we refer to discussions found in Fradin & Kerleroux (2003), Thornton (2018), and Pellegrini (2023).

fies optional columns for providing a reader-friendly `label`, and to map the cell onto other vocabularies (Section 3.4). Table 9 provides an example with mappings to universal dependencies (`ud`) and UniMorph tagsets.

<code>cell_id</code>	<code>label</code>	<code>ud</code>	<code>unimorph</code>
<code>nom.pl</code>	nominative plural	NOUN:Nom+Plur	N;NOM;PL
<code>nom.sg</code>	nominative singular	NOUN:Nom+Sing	N;NOM;SG
<code>voc.pl</code>	vocative plural	NOUN:Voc+Plur	N;VOC;PL
<code>voc.sg</code>	vocative singular	NOUN:Voc+Sing	N;VOC;SG
<code>acc.pl</code>	accusative plural	NOUN:Acc+Plur	N;ACC;PL
<code>acc.sg</code>	accusative singular	NOUN:Acc+Sing	N;ACC;SG

Tab. 9: Excerpt of a cells table for Latin nominal inflection.

The Paralex standard does not dictate how the structure of a paradigm should be analysed. This is a decision for the creator of the dataset. Rather, Paralex provides a standard for documenting that structure and increasing its transparency by mapping it onto other, existing ontologies.

Following a convention formalized in the Leipzig Glossing Rules (Comrie, Haspelmath & Bickel 2008), the `cell_id` identifiers for cells are combinations of feature-values separated by dots. This allows flexibility in the definition of cells, while ensuring the labels are fundamentally compositional. To prevent ambiguity, Paralex forbids the use of capitalization to distinguish feature values (e.g. S for subject but s for singular).¹²

Optionally, the feature-values themselves may be documented in a feature-values table. In addition to a column `value_id`, this table must contain the columns `label` (e.g. “singular”) and `feature` (e.g. “number”), which provide human-readable labels for the morphosyntactic value and feature respectively. Additional columns can be added, to link to other ontologies. An example is given in Table 10.

Some readers may wonder about our choice of terminology for the columns in the feature-values table. Note that what morphologists call a ‘value’, is termed a ‘feature’ in UniMorph (Batsuren et al. 2022), and what morphologists call a ‘feature’ is termed a ‘dimension’ in UniMorph. Paralex follows the morphologist convention.

¹² Paralex does not require these cells to be uppercase in the raw data, contrary to rule 5 of the Leipzig Glossing Rules. This is because transformation to uppercase in any human readable presentation is trivial.

value_id	label	feature	ud	unimorph	canonical_order
sg	singular	number	Sing	SG	1
pl	plural	number	Plur	PL	2
nom	nominative	case	Nom	NOM	1
voc	vocative	case	Voc	VOC	2
acc	accusative	case	Acc	ACC	3
gen	genitive	case	Gen	GEN	4
dat	dative	case	Dat	DAT	5
abl	ablative	case	Abl	ABL	6

Tab. 10: Feature-values table for Latin nominal inflection.

6.3.5 The sounds table and graphemes table

A sounds table can be used to describe the full inventory of sounds which appear in phonological forms. This table has one mandatory column, `sound_id`, which contains identifiers that are identical to the symbols used in the phonological forms. Predefined optional columns include a `label`, `comment`, and identifiers that link to other ontologies (`CLTS_id`, `PHOIBLE_id`). A graphemes table may serve to document orthographic graphemes in a similar manner.

6.3.6 The frequencies table

Although the tables forms, cells, and lexemes all permit a frequency column, this is sometimes not sufficient to express the full nuance of frequency relations. For example, there may be multiple sources of frequencies (multiple corpora or multiple measurement methodologies), or the measures might not be taken per-form, per-cell or per-lexeme (for example, they might be per orthographic token, without a cell distinction). A frequencies table supports the flexibility of expressing arbitrary frequency measurements, using any available identifier of any other table.

6.4 Metadata

Every Paralex dataset has its associated metadata (Section 5.1). The metadata for a Paralex dataset is stored as a JSON file following the Frictionless standard (Fowler, Barratt & Walsh 2018). This file declares descriptive and administrative metadata for the dataset, and structural metadata describing

each table and the set of columns that comprise it.¹³ Compiling this information formally for every column of every table of the dataset takes space; the JSON file is typically quite long, and is not intended to be read by or written by a person. Instead, it is best created using the automated tools discussed in Section 6.7 below.

The metadata serves as a body of documentation as well as a guide for validating, manipulating and revising the data (Section 4). It also promotes the longevity of the dataset itself, thus mitigating the problem of dataset loss (Section 2.4). The metadata specifications of Paralex have been designed to promote so-called *graceful degradation*, so that well in the future, even as conventions are forgotten, the dataset remains interpretable for as long as possible:

- As long as the Paralex standard and its automated tool ecosystem are maintained, the JSON file can be read, written and manipulated in user-friendly, domain-specific ways.
- If the Paralex ecosystem is partially or completely forgotten, the JSON file is still entirely interpretable as it respects the Frictionless conventions, and can still be manipulated by domain-general tools compliant with the Frictionless standard. Frictionless comprises libraries to manipulate data and metadata using Python, Javascript, R, Ruby, PHP, Java, Swift, go and Julia.
- If the Frictionless standard ceases to be available, the file can still be read as a structured set of attributes and values, the meaning of which, although based on a lost convention, have been designed to be relatively transparent.

6.5 Documentation

Even very good metadata will not cover everything that researchers might want to know about a dataset (but see Mosel 2012).

In modern practice, data statements (Bender & Friedman 2018) or data sheets (Geburu et al. 2021) provide information such as the process by which data were obtained or created, analytic and modeling choices, and its intended use. The Paralex standard provides a template file, `data_sheet.md`, with useful questions assembled from the data statements and data sheets. In addition, datasets typically include a `README.md` which summarizes the main information

¹³ In Frictionless terminology, the dataset is called the *package*, tables are *resources* and table columns are described by resource *schemas*.

about a dataset. Datasets may contain a number of additional documentation files. Additional Paralex documentation can be given in more plaintext files, using the markdown format (.md). A tool is provided to render the documentation in a showcase website (Section 6.7).

6.6 An RDF ontology to support Linked Data

The Paralex standard provides a means to solve the problem of inconsistent data presentation (Section 2.3) for inflectional datasets, and to ensure that the datasets are interoperable with one another. Another desideratum, however, is that inflectional datasets be interoperable with resources of other kinds. This is supported to a degree by the use of references, within Paralex tables, to external standards. However, further compatibility can be achieved through integration with the web of Linked Data (Section 3.4). This enables complex queries to be formulated across datasets, for example connecting entries from an inflected lexicon to lexicographic information in dictionaries or to attestations in a corpus.

To support integration between Paralex datasets and other Linked Data resources, we have adopted a strategy comparable to what was done in the CLDF project: we have supplemented Paralex with a Resource Description Framework (RDF) ontology that defines RDF classes and RDF properties that correspond respectively to tables and columns in Paralex. This, when coupled with the mandatory primary key identifiers that appear in each row of every Paralex table, is technically sufficient to enable the conversion of Paralex lexicons into RDF statements (see full specification in the Paralex documentation at https://www.paralex-standard.org/paralex_ontology.xml).

A challenge for any set of systems that aspire to interoperability is to successfully cope with minor departures in the use of similar concepts by different parties. In the framework of Linked Data, this is achieved using the mechanism of inheritance, according to which a more specific class or property *inherits* its traits from a more generic one. The classes and properties of Paralex inherit primarily from classes and properties of OntoLex (McCrae et al. 2017), and to some extent from other general-purpose ontologies like the General Ontology for Linguistic Description (GOLD, Farrar & Langendoen 2003) and LexInfo (Cimiano et al. 2011). Technically, a relationship of inheritance is established using the RDF predicate `subClassOf`. To illustrate how this can work in a concrete example, we take a short excerpt from PrinParLat, a Paralex lexicon

that has already been converted to RDF.¹⁴ Figure 5 presents a single row of the Paralex forms table, and next to it, the equivalent in RDF statements.

(a) Paralex forms table			(b) Corresponding RDF statements	
form_id	orth_form	cell		
191	ablauare	prs.act.inf	1	paralex:Form rdfs:subClassOf ontolex:Form .
			2	ppl:form_191 rdf:type paralex:Form .
			3	ppl:form_191 paralex:orth_form "ablauare" .
			4	ppl:form_191 paralex:cell ppl:cell_prs.act.inf .

Fig. 5: A row of the forms table of PrinParLat, and its expression in RDF. In the RDF statements, we use a compact serialization format, in which we replace the unvarying part of URIs with a shorthand, here: ppl → <http://lila-erc.eu/data/lexicalResources/prinparlat/id/>, rdf → <http://www.w3.org/1999/02/22-rdf-syntax-ns#>, paralex → <https://www.paralex-standard.org/paralex-ontology.xml#>, rdfs → <http://www.w3.org/2000/01/rdf-schema#>, ontolex → <http://www.w3.org/ns/lemon/ontolex#>.

Our Paralex RDF ontology defines a Form class. Line 1 in Figure 5b states that this class inherits from the more generic Form class in OntoLex. This inheritance relationship ensures interoperability between the forms in the Paralex dataset and forms provided in the numerous RDF-native lexicons based on OntoLex that already exist. RDF lines 2–4 in Figure 5b then declare information about the single Paralex record shown in Figure 5a, whose Paralex form_id identifier is 191. Line 2 states that the record whose RDF identifier is form_191¹⁵ is a record of a form (as opposed to a lexeme, cell, or grapheme, etc.). Lines 3 and 4 state that this record has an orth_form “ablauere” and a cell value cell_prs.act.inf.

Following the principles of Linked Data, data points in the RDF conversion of a Paralex dataset are assigned Uniform Resource Identifiers (URI) whenever possible. This makes it possible to define relations between data points across different Paralex lexicons, for instance if we wished to declare relations such as cognacy across languages.

Our ontology allows for conversion in both directions between Ontolex and Paralex lexicons, resulting in an enrichment of the amount of information available in both formats, and opening new opportunities for addressing problems of data coverage (Section 2.1).

¹⁴ <https://zenodo.org/records/10658159>.

¹⁵ In the RDF version, the string “form_” has been added to the id of forms to distinguish between numerical identifiers for different items (e.g., lexemes), in line with good practices for URI design.

6.7 Tools for a researcher-friendly Paralex ecosystem

To facilitate the work of linguists to develop Paralex datasets, we provide an ecosystem of accompanying tools, designed with two aims in mind: lowering the level of technical skills required to produce datasets; and reducing the number of repetitive steps, so dataset developers can focus on the content of the dataset rather than on the technical aspects. Each tool can be used without the need for skills in programming, database management, or the like. Our tools work in conjunction with data tables of the kind described in Section 6.3. The researcher should first create these tables and save them in the widely-used CSV format (this can be done, for example, in a wide range of software applications, including Excel or LibreOffice Calc). The tools themselves are run by entering simple commands in a command line interface (e.g. in the Windows or Mac OS terminal), to execute tasks such as creating the metadata JSON file (Section 6.7.1); validating the dataset table (Section 6.7.2), and generating visualizations (Section 6.7.3). These tools are implementations of our solutions (Section 5), applying the DeAR principles (Section 4) to the development of a Paralex-compliant dataset.

6.7.1 Metadata generation

```

1  title: "LatInFlexi: Latin Inflected Lexicon"
2  languages_iso639:
3    - lat
4  tables:
5    forms:
6      path: "latin_forms.csv"
7  name: latinflexi

```

Fig. 6: A minimal configuration file in YAML format, used to produce metadata for a dataset of Latin morphology, containing the only mandatory Paralex table for forms.

Paralex datasets are described by metadata in the Frictionless format (Section 6.4), stored as a JSON file. Creating this Frictionless JSON is best

performed automatically, not by hand. The `paralex` Python package¹⁶ provides a simple tool to do so, given a collection of Paralex-compliant tables and a simple parameter file that indicates where the files are located, and states the main properties of the dataset. The tool inspects each table it has been told about, matches it with definitions provided in the standard, and infers as much information about it as possible. Following the Frictionless standard, it writes the resulting metadata in a file ending in `.package.json`. Figure 6 provides an example of a minimal configuration file, and Figure 7 an excerpt of the resulting JSON package.

```

1  { "name": "latinflexi",
2    "title": "LatInfLexi: Latin Inflected Lexicon",
3    "languages_iso639": ["lat"],
4    "profile": "data-package",
5    "resources": [ {
6      "name": "forms",
7      "type": "table",
8      "title": "Inflected forms",
9      "path": "latin_forms.csv",
10     "scheme": "file",
11     "format": "csv",
12     "mediatype": "text/csv",
13     "encoding": "utf-8",
14     "...": "..."
15   }, "..."], "... }

```

Fig. 7: The beginning of a JSON file produced by the YAML configuration described above, using the command line `paralex meta latinflexi.yml`.

6.7.2 Metadata validation

The metadata file (Section 6.7.1) can be used to check the integrity of the data, delivering on the promise of the Automated validation part of the DeAR principles (Section 4.2). Carrying out validation regularly is the surest way to

¹⁶ The package can be installed using the generic python installer, `pip`, with the command: `pip install paralex`

ensure that a dataset remains well-formed and Paralex-compliant throughout its development process and subsequent life-cycle. Since Paralex follows the Frictionless metadata standard, much about a Paralex dataset can be validated using a generic tool provided by Frictionless. Additional checks can be run using our custom tool:

```
1 frictionless validate latinflexi.package.json
2 paralex validate latinflexi.package.json
```

These tools report back any infelicities encountered during validation and provide hints for improving the dataset.

6.7.3 Revisable dataset visualization

In Section 2.4, we cautioned that websites should not be confused with the data they are based on, yet they do remain a useful tool for viewing and exploring such data. To generate user-friendly websites for browsing a Paralex dataset, we provide a simple tool, `mkdocs-paralex`.¹⁷ The website it produces provides human-readable paradigm tables for each lexeme, automatically generated in a tabular layout. The arrangement of features into columns or rows can be customized interactively by the website user. Figure 8 shows an example of such human-readable paradigms. We take advantage of foreign keys (Section 6.3) to create hyperlinks that facilitate navigation. The website also contains all available documentation, including a readable representation of the JSON metadata. For details on how the tool is used, we refer the reader to the online tutorial.¹⁸

The websites that are generated by `mkdocs-paralex` are static. That is, they are self-contained and have no reliance on the presence of a database running in the background. Static websites minimize maintenance demands and constitute a secure way to showcase a dataset over the long term, helping to mitigate the problem of data loss (Section 2.4).

¹⁷ `mkdocs-paralex` relies on the `mkdocs` static website management tool and can be installed through `pip`.

¹⁸ <https://paralex-standard.org/tutorial/>

	lexeme_id	label	inflection_class	POS	frequency
+ sõnā_20502		sõnā	83	noun	30

	SG	PL
NOM	sõnā /suona/	sõnād /suonā d/
GEN	sõnā /suona/	sõnād /suonā d/
PART	sõnō /suon'u/	sõņđi /suon'di/
DAT	sõnēn /suonā n/	sõnādōn /suonā dōn/
INS	sõnāks /suonā ks/	sõnādōks /suonā dōks/
ILL	sõnō /suon'u/	sõņđi /suon'di/
ELAT	sõnāst /suonā st/	sõņđi /suon'di/
INESS	sõnās /suonā s/	sõņđi /suon'di/

Fig. 8: Screenshot of the lexemes table from the website of the *ParaLiv* dataset (Bouton 2024). The example shows the beginning of the paradigm of the Livonian noun *sõnā* ‘word’. Other tables can be accessed from the navigation bar.

6.7.4 Dataset dissemination

Publishing online visualizations is not sufficient to disseminate machine-readable datasets. In addition, datasets created in a **Decentralized** fashion still need to be referenced together, so that they are discoverable and accessible. For this, we recommend publishing them in the Paralex community¹⁹ of the Zenodo repository. Zenodo also ensures archival longevity and can make the dataset citable by providing a Digital Object Identifier (DOI). The paralex package provides shortcuts to list (`paralex list`) and download (`paralex get`) any dataset published in the Zenodo community. Appendix A summarizes Paralex datasets currently available on Zenodo.

6.7.5 Dataset revision

The four steps described above—metadata generation (Section 6.7.1), dataset validation (Section 6.7.2), visualization (Section 6.7.3) and dissemination (Section 6.7.4)—are likely to be performed repeatedly, both during the initial development of the dataset and in subsequent updates after its release. To

¹⁹ <https://zenodo.org/communities/paralex>

perform those steps, it suffices to run the tools via the command line at each iteration. Successive versions of the same dataset can be published on the same Zenodo deposit, maintaining separate DOIs for each version and for the entire deposit. For researchers with access to more advanced know-how, we also encourage the use of continuous deployment (Section 5.4), that is, user-defined workflow pipelines that run automatically whenever a change is made, and implemented for instance in repositories such as GitLab or GitHub. We emphasize though, that the continuous deployment approach is not a necessity, and that advanced skills are not needed for running our tools via the command line.

7 Conclusion

The use of machine-readable data resources is becoming integral to linguistic typology. Here we have put the spotlight on how those resources are created and maintained. We began by identifying problems that confront the field: problems of coverage, commensurability, consistency of presentation, durability, and more generally, impediments concerning time-limited funding and the technical skills required. We reviewed steps that have already been taken elsewhere and identified, in the principles of Open Data, FAIR, CARE and Linked Data, a path towards partial solutions. We found, however, that these initiatives leave a gap regarding the needs of dataset creators. To address this, we developed the DeAR principles, which recommend that data creation be **D**ecentralized (yet interoperable), so that expertise and responsibility is spread among participating partners; **A**utomatically verified, to support machine readability and high consistency; and **R**evisable, so that corrections, revisions and supplemental information systematically and reliably percolate from researcher data to public facing outputs.

We illustrated the practicalities and benefits of this approach by describing a language domain in which it has been successfully implemented, the **Paralex standard** for inflected lexicons. At its most basic, a Paralex lexicon is a simple, structured list of forms, accompanied with some metadata and useful documentation. At its most complex, it may be a network of tables documenting all entities involved (lexemes, inflected forms, paradigm cells, sounds, graphemes, etc.), and linking them through identifiers. In designing the standard, we implemented sufficient formal rigidity to enable machine-readability and to facilitate automatic verification and the generation of revisable outputs. Yet we understand that as linguists, we are often most interested specifically

in those parts of language that are very complex to analyse, and thus most likely to strain against rigid categories. Thus, while the standard upholds rigidity in terms of data formatting and arrangement, it retains flexibility around data content, allowing linguists to make project-specific choices about data content, while continuing to benefit from other aspects of standardization.

Our approach contributes to simplifying the technical skills required to create and use datasets. We provide ready-made tools for validating data against their metadata and for generating static websites from any standard dataset. The use of static websites (rather than dynamic ones) requires a minimum amount of technical infrastructure, lowering the costs of data presentation. Furthermore, we rely on simple text formats (CSV, JSON, markdown) which we believe will remain interpretable in the long term, thus addressing urgent issues regarding the preservation of data. This permits contemporary data, augmented with conventionally formatted modifications over time, to serve as the aggregate foundation for future empirical and theoretical research.

Well curated datasets, which remain available for the long term, are crucial to our work, and guaranteed to feed linguistic and typological work for decades.

A Datasets

Several Paralex datasets are already available on the Zenodo Paralex Community.²⁰ Table 11 lists them and provides some basic information about each.

²⁰ <https://zenodo.org/communities/paralex/records?q=&l=list&p=1&s=10&sort=newest>.

Name (Version)	Language	Family	Includes Orth. Freq.	Cells			Lexemes			DOI (Reference)	
				N	V	A	P	N	V	A	P
AraVeLex (1.0.1)	Standard Arabic	Afro-Asiatic	✓	✓	134	—	—	—	1046	—	10.5281/zenodo.10100677 (Beniamine 2018)
ParaKasem (1.0)	Kasem	Atlantic-Congo	✓	2	—	—	—	1909	—	—	10.5281/zenodo.14193534 (Beniamine & Guzmán Naranjo 2021)
VeLeRo (1.0)	Romanian	Indo-European	✓	—	39	—	—	—	7269	—	10.5281/zenodo.14202659 (Herce & Pricop 2024b)
VeLeCa (1.0)	Catalan	Indo-European	✓	—	50	—	—	—	3477	—	10.5281/zenodo.14203024 (Herce & Pricop 2024a)
Vlexique (2.0.2)	French	Indo-European	✓	✓	51	—	—	—	5274	—	10.5281/zenodo.10638681 (Beniamine, Coavoux & Bonami 2024)
LeFFI (2.0)	Italian	Indo-European	✓	—	53	—	—	—	2744	—	10.5281/zenodo.10522079 (Pellegriani & Cignarella 2020)
PrinParLatInFLexi (1.0.0)	Latin	Indo-European	✓	—	254	—	—	—	8014	—	10.5281/zenodo.17819183
PrinParLat (1.1)	Latin	Indo-European	✓	5	8	—	20935	8014	11051	—	10.5281/zenodo.8027826 (Pellegriani 2023)
LatInFLexi (2.0)	Latin	Indo-European	✓	✓	12	254	—	1038	3348	—	10.5281/zenodo.10522692 (Pellegriani 2020)
VeLePor (2.0.2)	Portuguese	Indo-European	✓	—	65	—	—	—	4992	—	10.5281/zenodo.5121543 (Beniamine, Bonami & Luis 2021)
Zallex (1.0.0)	Russian	Indo-European	✓	—	14	—	—	44833	—	—	10.5281/zenodo.15235589
VeLeSpa (1.0)	Spanish	Indo-European	✓	—	63	—	—	—	6513	—	10.5281/zenodo.14206007 (Herce 2024b)
ParaLatvian (1.0)	Standard Latvian	Indo-European	✓	14	—	—	—	3706	—	—	10.5281/zenodo.14217628 (Beniamine & Guzmán Naranjo 2021)
VeLePa (1.0)	Central Pame	Otomanguean	✓	—	58	—	—	—	216	—	10.5281/zenodo.14204262 (Herce 2024a)
Para-lipulankunyja (1.0)	Pitjantjatjara	Pama-Nyungan	✓	—	9	—	—	—	1137	—	10.5281/zenodo.14504218
PMST-Hyow (1.0.0)	Asho Chin	Sino-Tibetan	✓	—	341	11	—	18	—	1	10.5281/zenodo.17788529
PMST-Chiru (1.0.0)	Chiru	Sino-Tibetan	✓	—	156	10	—	10	—	1	10.5281/zenodo.17779437
Classical Tibetan (1.1)	Classical Tibetan	Sino-Tibetan	✓	—	4	—	—	265	—	—	10.5281/zenodo.12685122 (Sims-Williams & Zemp 2024)
PMST-Hmar (1.0.0)	Hmar	Sino-Tibetan	✓	—	156	7	—	2	—	1	10.5281/zenodo.17779055
PMST-Lamkang (1.0.0)	Lamkang	Sino-Tibetan	✓	—	152	6	—	2	—	1	10.5281/zenodo.17780049
PMST-Ranglong (1.0.0)	Ranglong	Sino-Tibetan	✓	—	136	6	—	2	—	1	10.5281/zenodo.17778036
ParaFin (2.0.0)	Finnish	Uralic	✓	✓	151	—	—	5000	—	—	10.5281/zenodo.13736131 (Bouton 2024)
ParaHungarian (1.0)	Hungarian	Uralic	✓	—	34	—	—	12729	—	—	10.5281/zenodo.14217969 (Beniamine & Guzmán Naranjo 2021)
ParaLiv (1.1.0)	Livonian	Uralic	✓	✓	16	—	—	6776	—	—	10.5281/zenodo.11391420 (Bouton 2024)
ParaKar (1.0.2)	Livi	Uralic	✓	✓	33	—	—	4975	—	—	10.5281/zenodo.13736170 (Bouton 2024)
Esthetic (1.0.5)	Standard Estonian	Uralic	✓	✓	28	51	—	5475	5076	—	10.5281/zenodo.8383522 (Beniamine et al. 2024)
Ngkolmpu (1.6)	Ngkolmpu	Yam	✓	—	1180	—	—	192	—	—	10.5281/zenodo.7049922 (Carroll 2025)

Tab. 11: Datasets currently archived on Zenodo in the Paralex format. N: nouns, V: verbs, P: pronouns, A: adjectives.

References

- Aigro, Mari & Virve-Anneli Vihman. 2023. Realised overabundance in Estonian noun paradigms: a corpus study. *Word Structure* 16(2–3). 154–175. <https://doi.org/10.3366/word.2023.0227>.
- Anderson, Cormac et al. 2018. A cross-linguistic database of phonetic transcription systems. *Yearbook of the Poznań Linguistic Meeting* 4(1). 21–53. <https://doi.org/10.2478/yplm-2018-0002>.
- Anderson, Cormac et al. 2023. Variation in phoneme inventories: quantifying the problem and improving comparability. *Journal of Language Evolution* 8(2). 149–168. <https://doi.org/10.1093/jole/lzad011>.
- Baker, Monya. 2016. 1,500 scientists lift the lid on reproducibility. *Nature* 533. 452–454.
- Batsuren, Khuyagbaatar et al. 2022. UniMorph 4.0: Universal Morphology. In Nicoletta Calzolari et al. (eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 840–855. Marseille, France: European Language Resources Association. <https://aclanthology.org/2022.lrec-1.89>.
- Bender, Emily M & Batya Friedman. 2018. Data statements for natural language processing: toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics* 6. 587–604. https://doi.org/10.1162/tacl_a_00041.
- Beniamine, Sacha. 2018. *Classifications flexionnelles: étude quantitative des structures de paradigmes*. Université Sorbonne Paris Cité - Université Paris Diderot dissertation. <https://tel.archives-ouvertes.fr/tel-01840448>.
- Beniamine, Sacha, Olivier Bonami & Ana R Luís. 2021. The fine implicative structure of European Portuguese conjugation. *Isogloss* 7(9). 1–35. <https://doi.org/10.5565/rev/isogloss.109>.
- Beniamine, Sacha & Jules Bouton. 2025. *Qumin*. <https://doi.org/10.5281/zenodo.15008373>.
- Beniamine, Sacha, Maximin Coavoux & Olivier Bonami. 2024. *Vlexique2.0: a rich lexicon of French verbal inflection with form-level frequencies*. Talk presented at the 21st International Morphology Meeting (IMM21), Vienna, Austria. 28–30 August 2024.
- Beniamine, Sacha & Matías Guzmán Naranjo. 2021. Multiple alignments of inflectional paradigms. *Proceedings of the Society for Computation in Linguistics* 4(21). 216–227. <https://doi.org/10.7275/ymc0-p491>.
- Beniamine, Sacha, Martin Maiden & Erich Round. 2019. *Romance verbal inflection dataset 2.0*. Version 2.0.1. <https://doi.org/10.5281/zenodo.3552367>.
- Beniamine, Sacha et al. 2024. Eesthetic: a Paralex lexicon of Estonian paradigms. In Nicoletta Calzolari et al. (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 5526–5537. Torino, Italia: ELRA & ICCL. <https://aclanthology.org/2024.lrec-main.491>.
- Berez-Kroeker, Andrea L. et al. (eds.). 2022. *The open handbook of linguistic data management*. The MIT Press.
- Berners-Lee, Tim. 2006. *Linked data-design issues*. <https://www.w3.org/DesignIssues/LinkedData.html>. Accessed: 2024 October 2024.
- Bickel, Balthasar. 2002. Autotypologizing databases and their use in fieldwork. In Peter K Austin, Helen Dry & Peter Wittenburg (eds.), *Proceedings of the International LREC Workshop on Resources and Tools in Field Linguistics, Las Palmas, Canary Islands*. 26–27

- May 2002. Nijmegen: Max Planck Institute for Psycholinguistics. <https://doi.org/10.5167/UZH-76860>.
- Bird, Steven & Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language* 79(3). 557–582. <https://doi.org/10.1353/lan.2003.0149>.
- Bonami, Olivier, Gauthier Caron & Clément Plancq. 2014. Construction d'un lexique flexionnel phonétisé libre du français. *SHS Web of Conferences, Actes du quatrième Congrès Mondial de Linguistique Française* 8. 2583–2596. <https://doi.org/10.1051/shsconf/20140801223>.
- Bouton, Jules. 2024. Towards standardized inflected lexicons for the Finnic languages. In Mika Hämmäläinen et al. (eds.), *Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages*, 59–66. Helsinki, Finland: Association for Computational Linguistics. <https://aclanthology.org/2024.iwclul-1.7/>.
- Bradner, Scott O. 1997. *Key words for use in RFCs to indicate requirement levels*. Tech. rep. RFC 2119. Internet Engineering Task Force. 3 pp. <https://doi.org/10.17487/RFC2119>. <https://www.rfc-editor.org/info/bcp47> (5 November, 2024).
- Carroll, Mae. 2025. *Ngkolmpu paralex v1.6*. Version 1.6. <https://doi.org/10.5281/zenodo.15016957>.
- Carroll, Mae & Sasha Beniamine. 2025. Exponence and the theory of discriminative information in paradigms. *Morphology* 35. 227–269. <https://doi.org/10.1007/s11525-025-09437-2>.
- Carroll, Stephanie Russo et al. 2020. The CARE principles for indigenous data governance. *Data Science Journal* 19(1). 1–12. <https://doi.org/10.5334/dsj-2020-043>.
- Casillas, Joseph V. et al. 2025. Opening open science to all: demystifying reproducibility and transparency practices in linguistic research. *Linguistics* 63(6). 1547–1575. <https://doi.org/10.1515/ling-2023-0249>.
- Chao, Yuen-Ren. 1934. The non-uniqueness of phonemic solutions of phonetic systems. *Proceedings of the Institute of History and Linguistics* 4. 363–397. <https://doi.org/10.6355/bihpas.193401.0363>.
- Cimiano, Philipp et al. 2011. LexInfo: a declarative model for the lexicon-ontology interface. *Journal of Web Semantics* 9(1). 29–51. <https://doi.org/10.1016/j.websem.2010.11.001>.
- Codd, Edgar F. 1970. A relational model of data for large shared data banks. *Communications of the Association for Computing Machinery* 13(6). 377–387. <https://doi.org/10.1145/362384.362685>.
- Comrie, Bernard, Martin Haspelmath & Balthasar Bickel. 2008. *The Leipzig Glossing Rules: conventions for interlinear morpheme-by-morpheme glosses*. <https://www.eva.mpg.de/lingua/resources/glossing-rules.php> (23 June, 2025).
- Corbett, Greville G. 2000. *Number* (Cambridge textbooks in linguistics). Cambridge: Cambridge University Press. 1358 pp.
- Corbett, Greville G. 2006. *Agreement* (Cambridge textbooks in linguistics). Cambridge [u.a.]: Cambridge Univ. Press. 328 pp.
- Corbett, Greville G. 2013. Paradigm conventions. Paper at the 46th Annual Meeting of the Societas Linguistica Europaea, Split, Croatia. 18–21 September 2013. https://www.academia.edu/9055930/Paradigm_conventions.
- Corbett, Greville G. 2015. Morphosyntactic complexity: a typology of lexical splits. *Language* 91(1). 145–193. <https://doi.org/10.1353/lan.2015.0003>.
- Creissels, Denis. 2006. *Creissels: Denis*. Vol. 1. Paris: Hermes Science / Lavoisier. 412 pp.

- Dryer, Matthew S & Martin Haspelmath (eds.). 2013. *WALS online*. Data set. Leipzig. <https://doi.org/10.5281/zenodo.13950591>.
- Elsner, Micha & Sacha Beniamine. 2024. Computational approaches to morphological typology. *Journal of Language Modelling* 12(2). 271–286. <https://doi.org/10.15398/jlm.v12i2.431>.
- Evans, Nicholas & Alan Dench. 2006. Introduction: catching language. In Felix K Ameka, Alan Dench & Nicholas Evans (eds.), *Catching language: the standing challenge of grammar writing*, 1–40. Berlin: De Gruyter Mouton. <https://doi.org/10.1515/9783110197693.1>.
- Farrar, Scott & D Terence Langendoen. 2003. A linguistic ontology for the semantic web. *GLOT International* 7(3). 97–100. <https://user.phil-fak.uni-duesseldorf.de/~bontcheva/WS0809OL/GLOT-LinguisticOntology.pdf>.
- Forkel, Robert, Sebastian Bank & Christoph Rzymiski. 2019. *clld/clld: clld - a toolkit for cross-linguistic databases*. Version 5.0.0. <https://doi.org/10.5281/zenodo.3437148>.
- Forkel, Robert & Johann-Mattis List. 2020. CLDFBench: give your cross-linguistic data a lift. eng. In Nicoletta Calzolari et al. (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 6995–7002. Marseille, France: European Language Resources Association. <https://aclanthology.org/2020.lrec-1.864/>.
- Forkel, Robert et al. 2018. Cross-linguistic data formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data* 5(180205). <https://doi.org/10.1038/sdata.2018.205>.
- Fowler, Dan, Jo Barratt & Paul Walsh. 2018. Frictionless data: making research data quality visible. *International Journal of Digital Curation* 12(2). 274–285. <https://doi.org/10.2218/ijdc.v12i2.577>.
- Fradin, Bernard & Françoise Kerleroux. 2003. Troubles with lexemes. In Geert Booij et al. (eds.), *Selected papers from the Third Mediterranean Morphology Meeting*, 177–196. Barcelona: IULA – Universitat Pompeu Fabra.
- Gebru, Timnit et al. 2021. Datasheets for datasets. *Communications of the Association for Computing Machinery* 64(12). 86–92. <https://doi.org/10.1145/3458723>.
- Hammarström, Harald & Robert Forkel. 2022. Glottocodes: identifiers linking families, languages and dialects to comprehensive reference information. *Semantic Web* 13(6). 917–924. <https://doi.org/10.3233/sw-212843>.
- Heath, Tom & Christian Bizer. 2011. *Linked data: evolving the web into a global data space*. Dordrecht: Springer. <https://doi.org/10.1007/978-3-031-79432-2>.
- Herce, Borja. 2024a. VeLePa: Central Pame verbal inflection in a quantitative perspective. *Morphology* 34(3). 281–319. <https://doi.org/10.1007/s11525-024-09426-x>.
- Herce, Borja. 2024b. VeLeSpa: an inflected verbal lexicon of Peninsular Spanish and a quantitative analysis of paradigmatic predictability. *Language Resources and Evaluation* 59. 1705–1718. <https://doi.org/10.1007/s10579-024-09776-2>.
- Herce, Borja & Bogdan Pricop. 2024a. VeLeCa: a verbal lexicon of Catalan with PCFP analysis. *Isogloss* 10(1). 1–17. <https://doi.org/10.5565/rev/isogloss.457>.
- Herce, Borja & Bogdan Pricop. 2024b. VeLeRo: an inflected verbal lexicon of standard Romanian and a quantitative analysis of morphological predictability. *Language Resources and Evaluation* 59(1). 621–637. <https://doi.org/10.1007/s10579-024-09721-3>.
- International Phonetic Association. 1999. *Handbook of the International Phonetic Association: a guide to the use of the International Phonetic Alphabet*. Cambridge: Cambridge University Press.
- Internet Engineering Task Force. 2022. *Best current practice* 47. Standard BCP 47. RFC Editor. <https://www.rfc-editor.org/info/bcp47> (5 November, 2024).

- ISO Central Secretary. 2019a. *Date and time – representations for information interchange – part 1: basic rules*. Standard ISO 8601-1:2019. Geneva: International Organization for Standardization. <https://www.iso.org/standard/70907.html>.
- ISO Central Secretary. 2019b. *Geographic information — referencing by coordinates*. Standard ISO 19111:2019. Geneva: International Organization for Standardization. <https://www.iso.org/standard/74039.html>.
- ISO Central Secretary. 2020. *Codes for the representation of names of countries and their subdivisions – part 1: country codes*. Standard ISO 3166-1:2020. Geneva: International Organization for Standardization. <https://www.iso.org/standard/72482.html>.
- ISO Central Secretary. 2022. *Information and documentation – codes for the representation of names of scripts*. Standard ISO 15924:2022. Geneva: International Organization for Standardization. <https://www.iso.org/standard/81905.html>.
- ISO Central Secretary. 2023. *Code for individual languages and language groups*. Standard ISO 639:2023. Geneva: International Organization for Standardization. <https://www.iso.org/standard/74575.html>.
- List, Johann-Mattis, Michael Cysouw & Robert Forkel. 2016. Concepticon: a resource for the linking of concept lists. In Nicoletta Calzolari et al. (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, 2393–2400. Luxembourg: European Language Resources Association (ELRA). <http://www.lrec-conf.org/proceedings/lrec2016/summaries/127.html>.
- List, Johann-Mattis et al. 2022. Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. *Scientific Data* 9(1). 316. <https://doi.org/10.1038/s41597-022-01432-0>.
- List, Johann-Mattis et al. 2023. *Lexibank analysed [Data set]*. <https://doi.org/10.5281/zenodo.7836668>.
- List, Johann-Mattis et al. 2024. *CLTS. Cross-Linguistic Transcription Systems (v2.3.0) [Data set]*. <https://doi.org/10.5281/zenodo.10997741>.
- Maiden, Martin (ed.). 2010. *Oxford Online Database of Romance Verb Morphology*. Browsable database. <http://romverbmorph.clp.ox.ac.uk/> (23 June, 2025).
- McCrae, John P et al. 2017. The Ontolex-Lemon model: development and applications. In Iztok Kosem et al. (eds.), *Proceedings of eLex 2017 conference*, 19–21. <https://elex.link/elex2017/wp-content/uploads/2017/09/paper36.pdf>.
- Mosel, Ulrike. 2012. Advances in the accountability of grammatical analysis and description by using regular expressions. In Sebastian Nordoff (ed.), *Electronic grammaticography*, 235–250. Honolulu: University of Hawai'i Press.
- Mosteiro Romero, Pablo & Damián Blasi. 2025. Word boundaries and the morphology-syntax trade-off. In Sane Yagi et al. (eds.), *Proceedings of the New Horizons in Computational Linguistics for Religious Texts*, 86–93. Abu Dhabi, UAE: Association for Computational Linguistics. <https://aclanthology.org/2025.clrel-1.9/> (8 February, 2026).
- Namer, Fiammetta et al. 2023. Démonette-2, a derivational database for French with broad lexical coverage and fine-grained morphological descriptions. *Lexique* 33. 6–40. <https://doi.org/10.54563/lexique.1242>.
- Nivre, Joakim et al. 2016. Universal Dependencies v1: a multilingual treebank collection. In Nicoletta Calzolari et al. (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 1659–1666. Portorož, Slovenia: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2016/pdf/348_Paper.pdf.

- Nivre, Joakim et al. 2020. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In Nicoletta Calzolari et al. (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 4034–4043. Marseille, France: European Language Resources Association. <http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.497.pdf>.
- Pellegrini, Matteo. 2020. Patterns of interpredictability and principal parts in Latin verb paradigms: an entropy-based approach. *Journal of Latin Linguistics* 19(2). 195–229. <https://doi.org/10.1515/joll-2020-2014>.
- Pellegrini, Matteo. 2023. Flexemes in theory and in practice. *Morphology* 33. 361–395. <https://doi.org/10.1007/s11525-023-09414-7>.
- Pellegrini, Matteo & Alessandra Teresa Cignarella. 2020. (Stem and word) predictability in Italian verb paradigms: an entropy-based study exploiting the new resource LeFFI. In Felice Dell'Orletta, Johanna Monti & Fabio Tamburini (eds.), *Proceedings of the 7th Italian Conference on Computational Linguistics CLiC-it 2020*, 341–346. Accademia University Press. <https://doi.org/10.4000/books.aaccademia.8830>.
- Pellegrini, Matteo & Marco Passarotti. 2018. LatInFLexi: an inflected lexicon of Latin verbs. In Elena Cabrio, Alessandro Mazzei & Fabio Tamburini (eds.), *Proceedings of the 5th Italian Conference on Computational Linguistics CLiC-it 2018*, 324–329. Accademia University Press. <https://doi.org/10.4000/books.aaccademia.3582>.
- Phillips, Addison & Mark Davis. 2006. *Matching of Language Tags*. Tech. rep. RFC 4647. Internet Engineering Task Force. <https://doi.org/10.17487/RFC4647>. <https://www.rfc-editor.org/info/bcp47> (5 November, 2024).
- Phillips, Addison & Mark Davis. 2009. *Tags for Identifying Languages*. Tech. rep. RFC 5646. Internet Engineering Task Force. <https://doi.org/10.17487/RFC5646>. <https://www.rfc-editor.org/info/bcp47> (5 November, 2024).
- Round, Erich R. 2023. Canonical phonology and criterial conflicts: Relating and resolving four dilemmas of phonological typology. *Linguistic Typology* 27(2). 267–287. <https://doi.org/10.1515/lingty-2022-0032>.
- Rzysmski, Christoph et al. 2020. The database of cross-linguistic colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific Data* 7(1). 13. <https://doi.org/10.1038/s41597-019-0341-x>.
- Schweinberger, Martin & Michael Haugh. 2025. Reproducibility, replicability, and robustness in corpus linguistics: an introduction. *International Journal of Corpus Linguistics*.
- Sims, Andrea. 2015. *Inflectional defectiveness*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9781107053854>.
- Sims-Williams, Helen & Marius Zemp. 2024. *Classical Tibetan Verbal Paradigms*. Version 1.1. Zenodo. <https://doi.org/10.5281/zenodo.14170906>.
- Skirgård, Hedvig et al. 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances* 9(16). eadg6175. <https://doi.org/10.1126/sciadv.adg6175>.
- Spencer, Andrew. 2012. Identifying stems. *Word Structure* 5. 88–108. <https://doi.org/10.3366/word.2012.0021>.
- Stump, Gregory T. & Raphael Finkel. 2013. *Morphological typology: from word to paradigm*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9781139248860>.
- Swadesh, Morris. 1952. Lexico-statistic dating of prehistoric ethnic contacts: with special reference to North American Indians and Eskimos. *Proceedings of the American Philosophical Society* 96(4). 452–463. <http://www.jstor.org/stable/3143802>.

- Sylak-Glassman, John et al. 2015. A language-independent feature schema for inflectional morphology. In Chengqing Zong & Michael Strube (eds.), *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 674–680. Beijing, China: Association for Computational Linguistics (ACL). <https://doi.org/10.3115/v1/P15-2111>.
- Tallman, Adam J. R. & Sandra Auderset. 2023. Measuring and assessing indeterminacy and variation in the morphology-syntax distinction. *Linguistic Typology* 27(1). 113–156. <https://doi.org/10.1515/lingty-2021-0041>.
- TEI Consortium (ed.). 2025. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 4.9.0. <https://tei-c.org/guidelines/p5/> (23 June, 2025).
- Thornton, Anna M. 2018. Troubles with flexemes. In Oliver Bonami et al. (eds.), *The lexeme in descriptive and theoretical morphology*, 303–321. Berlin: Language Science Press. <https://doi.org/10.5281/zenodo.1407011>.
- Thornton, Anna M. 2019. Overabundance: A Canonical Typology. In Franz Rainer et al. (eds.), *Competition in Inflection and Word-Formation*, 223–258. Cham: Springer International Publishing.
- Vidra, Jonáš et al. 2019. DeriNet 2.0: towards an all-in-one word-formation resource. In Magda Ševčíková et al. (eds.), *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2019)*, 81–89. Praha, Czechia: ÚFAL MFF UK. <https://aclanthology.org/W19-8500.pdf>.
- Wickham, Hadley. 2014. Tidy data. *Journal of statistical software* 59. 1–23.
- Wilkinson, Mark D. et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3(1). 160018. <https://doi.org/10.1038/sdata.2016.18>.
- Wilson, Greg et al. 2017. Good enough practices in scientific computing. *PLoS computational biology* 13(6). e1005510.
- Windhouwer, Menzo & Alexis Dimitriadis. 2008. Sustainable operability: keeping complex resources alive. In A. Witt et al. (eds.), *Language Resources and Evaluation Conference 2008 workshop: sustainability of language resources and tools for natural language processing*, 9–18. http://www.lrec-%20conf.org/proceedings/lrec2008/workshops/W17_Proceedings.pdf.
- Witzlack-Makarevich, Alena et al. 2022. Managing AUTOTYP data: design principles and implementation. In Andrea L. Berez-Kroeker et al. (eds.), *The Open Handbook of Linguistic Data Management*, 631–642. The MIT Press. <https://doi.org/10.7551/mitpress/12200.003.0061>.
- Zeller, Britta, Jan Šnajder & Sebastian Padó. 2013. DERivBase: inducing and evaluating a derivational morphology resource for German. In Hinrich Schuetze, Pascale Fung & Massimo Poesio (eds.), *Proceedings of the Association for Computational Linguistics 2013*, 1201–1211. Sofia, Bulgaria: Association for Computational Linguistics. www.aclweb.org/anthology/P13-1118.pdf.