Paralex: a DeAR standard for rich lexicons of inflected forms.

Sacha Beniamine[1], *Cormac Anderson*[2], *Mae Carroll*[3], *Matías Guzmán Naranjo*[4], *Borja Herce*[5], *Matteo Pellegrini*[6], *Erich Round*[1], *Helen Sims-Williams*[1], *Tiago Tresoldi*[7]

[1]University of Surrey; [2]Max Planck Institute EVA; [3]Australian National University;
[4]Albert-Ludwigs-Universität Freiburg; [5]University of Zurich;
[6]CIRCSE Research Centre, Università Cattolica del Sacro Cuore, Milan; [7]Uppsala Universitet

1 Introduction

We present Paralex[1], a new technical standard for inflected lexicons in tabular format. Inflected lexicons document the inflected forms of words, such as the conjugations of verbs and the declensions of nouns. Such datasets are crucial to support morphological investigation using both computational and traditional methodologies, and constitute a necessary foundation for data-driven studies of individual systems as well as large scale typological works.

Many existing morphological datasets are not published durably. Some use proprietary formats, others exist solely as web portals (see Maiden et al. 2010, standardised by Beniamine et al. 2019). Resources which were intended for manual exploration are often not machine-readable. Those which are, often use their own sets of conventions, and are not inter-operable (see eg. Bonami et al., 2014; Pellegrini & Passarotti, 2018; Feist & Palancar, 2015). The Unimorph datasets (McCarthy et al., 2020) do provide inter-operable lists of inflected forms, but their usefulness for linguistic investigation is limited both by their automatic extraction and an exclusive focus on orthography (Malouf et al., 2020).

The Paralex standard aims to bring about high quality resources, which can be richly annotated, are machine-readable, inter-operable and durable. It describes lexicons constituted of `csv` tables in long format forming a relational database, accompanied by metadata in `json` format (§ 2). The standard is devised to promote good data practices and abides by the FAIR principles (Wilkinson et al., 2016), as well as our own set of principles (DeAR, § 3).

2 Data and metadata formats

Paradigms are conventionally written as tables in a variety of formats (Corbett, 2013). Authors often present single paradigms as in Table 1.a., where rows and columns represent morphosyntactic features. Such tables are impractical for presenting many lexemes, as this would require multiple tables. Thus, a more common format for this purpose (see e.g. Flexique, Bonami et al. 2014) is the Plat (Stump & Finkel, 2013), which arranges paradigm cells in columns, and lexemes in rows, as in Table 1.b. This format is more generally known as a *wide form* table. The major draw-back of this format is that it can only ever express a single piece of information per cell/lexeme intersection, making it impossible to cleanly record overabundant forms (Thornton, 2012) or multiple pieces of information for each form, including but not limited to its phonological form, its frequency, source, analysis, etc. Thus, we adopt instead the *long form* (For more discussion on wide vs long form for linguistic data, see Forkel et al., 2018), in which each inflected wordform is given its own row, as shown in Table 1.c. Rows have unique identifiers, and columns for forms, cells, and lexemes, and any further information. Overabundant word forms lead to multiple rows.

[1] The full standard specifications and documentation can be found at `https://www.paralex-standard.org`

(a) Single paradigm table

	SINGULAR	PLURAL
NOMINATIVE	rosa	rosae
VOCATIVE	rosa	rosae
ACCUSATIVE	rosam	rosās
GENITIVE	rosae	rosārum
DATIVE	rosae	rosīs
ABLATIVE	rosā	rosīs

(c) Long form table

form_id	cell	lexeme	orth_form
f1	NOM.SG	rosa	rosa
f2	VOC.SG	rosa	rosa
f3	ACC.SG	rosa	rosam
f13	NOM.SG	dominus	dominus
f14	VOC.SG	dominus	domine
...	...	...	...

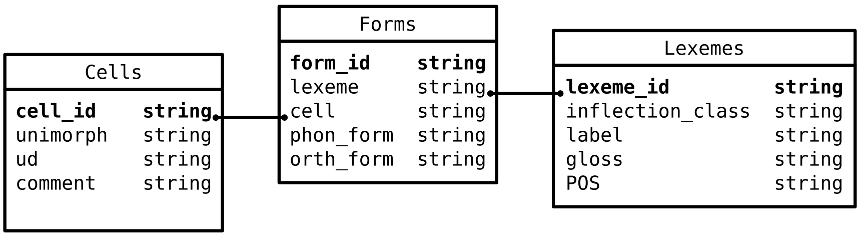
(b) Wide form table

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

Table 1: Paradigm formats, illustrated on two Latin nouns (Pellegrini & Passarotti, 2018).

A paralex lexicon is minimally constituted of a simple `forms` table (see Table 1.c), associating forms (orthographic or phonological) with paradigm cells, lexemes, and unique identifiers. The standard further describes tables to document entities from the `forms` table: `lexemes`, `cells`, `feature-values`, `sounds`, and `graphemes`. A `tags` table declares user-defined properties of forms and a very flexible `frequencies` table records frequency measurements. A set of columns is pre-defined for each table. Paralex lexicons may use pre-defined tables and columns, adding any additional ones as needed. The tables are linked by two types of relationships. Foreign key relations allow direct references between tables rows: For example, "NOM.SG" in the cell column of the `forms` table refers to the row with the identifier "NOM.SG" in the `cells` table. The foreign key relations between the three main tables are illustrated in Figure 1. Moreover, elements from some tables are composed of identifiers from other tables. For example, cells (e.g. NOM.SG) are composed of feature-values separated by dots (NOM, SG), orthographic forms are composed of graphemes, phonological forms are composed of sounds symbols, etc.

Beyond relations between tables, references to linked vocabularies greatly increase the value of datasets, and are encouraged. Languages can be denoted by glottocodes or ISO-639-2 codes; cells and features can refer to the Universal Dependencies and/or to the Unimorph conventions, sounds may refer to CLTS' BIPA (Anderson et al., 2018), etc. To further enhance interoperability with different resources, the standard is coupled with an ontology, where RDF classes and properties are introduced, corresponding to tables and columns defined in the standard, respectively. Their relation to existing standard vocabularies – such as the General Ontology for Linguistic Description (GOLD; Farrar & Langendoen 2003) and the Lexicon Model for Ontologies (OntoLex; McCrae et al. 2017) – is expressed by means of sub-class (`subClassOf`) and sub-property (`subPropertyOf`) relations, as defined in the RDF Schema vocabulary. This allows the conversion of Paralex data into ontolex-compliant lexicons in RDF, guaranteeing semantically richer interoperability not only with other morphological lexicons, but also with lexical resources of other kinds.



Cells	
cell_id	**string**
unimorph	string
ud	string
comment	string

Forms	
form_id	**string**
lexeme	string
cell	string
phon_form	string
orth_form	string

Lexemes	
lexeme_id	**string**
inflection_class	string
label	string
gloss	string
POS	string

Figure 1: Relations between the three main Paralex tables.

Metadata are any information about the dataset that are not directly part of the data. A first type of metadata is global information about a dataset, such as its author(s), name, identifier, license, etc. This information is usually provided in landing pages, articles or documentation files, in a way that is easy to understand for humans, but often not machine readable. Furthermore, many other pieces of information about the data itself are often left implicit, such as: what does each table document? How are tables related? What values are expected in each column? This is neither future-proof (the context is likely to be lost) nor machine-readable. Thus, Paralex lexicons explicitly encode metadata in a `json` file following the frictionless standard (Fowler et al., 2018). Its creation is facilitated by a Paralex python package which can fill in all conventional information from the standard.

3 Philosophy

Paralex datasets adhere to the FAIR principles (Wilkinson et al., 2016), which focus on data users: they ensure that datasets be readable by both machines and by humans across subfields, disciplines and time. Focusing on data creators, we introduce our own set of principles for scientific data: **De**centralisation, **A**utomated verification and **R**evisable pipelines (DeAR).

The acronym FAIR stands for Findable, Accessible, Interoperable and Reusable. Findability relies on persistent global identifiers (F1), rich metadata (F2) referring to the identifier (F3), and indexation in searchable resources (F4). Paralex addresses F2 through the metadata file, recommends using DOIs (F1,F3), and archiving lexicons in dedicated repositories (F4). These measures also ensure that the data is Accessible. Inter-operability consists in using a formal, accessible, shared, broadly applicable language for knowledge representation (I1), FAIR vocabularies (I2) and reference to other (meta)data (I3). The formats chosen for Paralex fit the descriptions in I1. Compliance with I2 and I3 rely on the use of linked identifiers, and the Paralex ontology. Finally, rich metadata also addresses reusability, by ensuring well-described data (R1) which can be re-used and combined in other contexts.

When faced with the task of creating a large number of standardised datasets, one solution is for a single team to retro-standardise large amounts of data into a single database. Unfortunately, compounded datasets tend to be cited at the expense of original resources, leading to loss of recognition for data creators. Moreover, centralisation concentrates power over indigenous and endangered languages into the hands of a few institutions, going contrary to the CARE principles (Carroll et al., 2020). Thus, Paralex rather aims to stimulate a **De**centralized adoption of the standard. Although there must of course be a single definition of the standard, we intend to make it easy and flexible to use, and to produce tools which incentivize its adoption. Creating large databases is difficult and error-prone. In order to improve data quality, we promote the **A**utomated validation of datasets. The statements contained in the metadata file can be verified automatically against the data using existing frictionless tools. This process can ensure perfect formatting, valid references across tables, and check expected properties of data content. Validation can be performed at each update of the data to maintain high data quality. Finally, it is crucial for data to be linked to its published presentations (such as websites) through **R**evisable pipelines. The inter-operability of standardized datasets makes it possible to create websites which can be re-generated whenever the data is updated.

4 Conclusion

The Paralex standard provides formal conventions for coding inflected lexicons and their metadata. It is suited to encoding inflectional systems across languages, for purposes ranging from lists of inflected forms to richly annotated lexicons. It describes mechanisms to handle phenomena such as overabundance, defectivity, and multiple types of variation. It is accompanied by a helper tool to generate the metadata with minimal effort, and a tool is in development to create static websites automatically. As linguists, we are most interested in the parts of language that are complex to analyze, and thus complex to code. Thus, this standard accommodates a great deal of flexibility regarding the exact content of the data, allowing linguists to make project-specific analytical choices about content, while reaping other benefits of standardization.

References

Anderson, Cormac, Tiago Tresoldi, Thiago Chacon, Anne-Maria Fehn, Mary Walworth, Robert Forkel & Johann-Mattis List. 2018. A cross-linguistic database of phonetic transcription systems. *Yearbook of the Poznan Linguistic Meeting* 4(1). 21–53. doi:10.2478/yplm-2018-0002.

Beniamine, Sacha, Martin Maiden & Erich Round. 2019. Romance verbal inflection dataset 2.0. doi:10.5281/zenodo.3552367.

Bonami, Olivier, Gauthier Caron & Clément Plancq. 2014. Construction d'un lexique flexionnel phonétisé libre du français. In Franck Neveu, Peter Blumenthal, Linda Hriba, Annette Gerstenberg, Judith Meinschaefer & Sophie Prévost (eds.), *Actes du quatrième congrès mondial de linguistique française*, 2583–2596.

Carroll, Stephanie Russo et al. 2020. The CARE principles for indigenous data governance. *Data Science Journal* 19. doi:10.5334/dsj-2020-043.

Corbett, Greville G. 2013. Paradigm conventions. Paper at the 46th Annual Meeting of the Societas Linguistica Europaea, Split, Croatia. 18-21 September 2013.

Farrar, Scott & D Terence Langendoen. 2003. A linguistic ontology for the semantic web. *GLOT international* 7(3). 97–100.

Feist, Timothy & Enrique L. Palancar. 2015. Oto-Manguean Inflectional Class Database. University of Surrey. doi:10.15126/SMG.28/1.

Forkel, Robert, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping & Russell D. Gray. 2018. Cross-linguistic data formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data* 5. 180205. doi: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. doi:10.2218/ijdc.v12i2.577.

Maiden, Martin et al. 2010. Oxford online database of romance verb morphology. Online website. Browsable database. `http://romverbmorph.clp.ox.ac.uk/`.

Malouf, Robert, Farrell Ackerman & Arturs Semenuks. 2020. Lexical databases for computational analyses: A linguistic perspective. In *Proceedings of the society for computation in linguistics 2020*, 446–456. New York: ACL. `https://aclanthology.org/2020.scil-1.52`.

McCarthy, Arya D. et al. 2020. UniMorph 3.0: Universal Morphology. In *Proceedings of the twelfth language resources and evaluation conference*, 3922–3931. Marseille, France: European Language Resources Association. `https://aclanthology.org/2020.lrec-1.483`.

McCrae, John P, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar & Philipp Cimiano. 2017. The ontolex-lemon model: development and applications. In *Proceedings of eLex 2017 conference*, 19–21.

Pellegrini, Matteo & Marco Passarotti. 2018. LatInfLexi: an Inflected Lexicon of Latin Verbs. In Elena Cabrio, Alessandro Mazzei & Fabio Tamburini (eds.), *Proceedings of the fifth italian conference on computational linguistics (clic-it 2018)*, vol. 2253 CEUR Workshop Proceedings, December. `http://ceur-ws.org/Vol-2253/paper23.pdf`.

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

Thornton, Anna M. 2012. Reduction and maintenance of overabundance. a case study on italian verb paradigms. *Word Structure* 5(2). 183–207.

Wilkinson, Mark D. et al. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific Data* 3(1). 160018. doi:10.1038/sdata.2016.18.