University of Surrey · Australian National University · universität freiburg · University of Zurich UZH · Università Cattolica del Sacro Cuore · Uppsala Universitet

Paralex

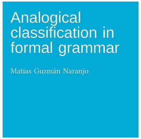
a DeAR standard for rich lexicons of inflected forms.

Sacha Beniamine &

Cormac Anderson

Mae Carroll

Matías Guzmán-Naranjo



Analogical classification in formal grammar

Matías Guzmán Naranjo

Borja Herce

Matteo Pellegrini

Erich Round

Helen Sims-Williams

Tiago Tresoldi

Introduction

> *"You can also collect butterflies and make many observations. If you like butterflies, that's fine; but such work must not be confounded with research, which is concerned to discover explanatory principles of some depth and fails if it does not do so."*
>
> *Chomsky, 1979, p. 57*

- We should not underestimate the importance of data
- Good data require good, usable standards
- Paralex: a standard for lexicons of inflected forms (paradigms)

Principles

Principles

Open Data ***1958***

Available
Editable
Re-Distributable

For the common good

Principles

Open Data *1958*

Available
Editable
Re-Distributable

For the common good

FAIR *2016*

Findable
Accessible
Inter-operable
Reusable

For machines

Principles

Open Data *1958*

Available
Editable
Re-Distributable

For the common good

FAIR *2016*

Findable
Accessible
Inter-operable
Reusable

For machines

CARE *2020*

Collective benefit
Authority to control
Responsibility
Ethics

For communities

Principles

Open Data *1958*

Available
Editable
Re-Distributable

For the common good

FAIR *2016*

Findable
Accessible
Inter-operable
Reusable

For machines

CARE *2020*

Collective benefit
Authority to control
Responsibility
Ethics

For communities

DeAR *2023*

Decentralized
Automated validation
Revisable

For researchers

Decentralized

- Centralized standardization is not long-lasting
- The standard should be usable by all,
- Its use can be incentivized by useful tools

Automated validation

- Manual curation of large datasets is necessary but error-prone

- Performing automated validation can:
  - Check the format and structure
  - Check constraints on content
  - Check references to other data

- We get this “for free” by using the standard !

Revisable

- We should not confuse **databases** with their presentations
- Presentations should be updated seamlessly when data changes
- **Solution:** (re)generate them from a single data source:
  - For print (pdf documents)
  - For the web (static sites)

Solutions

To publish high quality, easily citable, scientifically impactful data, useful for the long term:

Creation

- Metadata FAIR DeAR
- Standards Inter-operable Reusable DeAR
- Linked data Interoperable Reusable
- Validation DeAR

Publication

- Documentation Reusable
- DOIs Findable Accessible
- License Open Reusable
- Archived downloads FAIR
- Continuous pipelines DeAR

The standard

Table notations for paradigms

1- wide format

	PRS.1SG	PRS.2SG	PRS.3SG	PRS.1PL	PRS.2PL	PRS.3PL	...
CHANTER	ʃãt	ʃãt	ʃãt	ʃãtɔ̃	ʃãte	ʃãt	...
PELER	pɛl	pɛl	pɛl	pəlɔ̃	pəle	pɛl	...
MENER	mɛn	mɛn	mɛn	mənɔ̃	məne	mɛn	...
...	...	...	...	...	...	...	

2- Single paradigm table

	PRS	PST	FUT	...
1SG	mãʒ	mãʒE	mãʒəʁE	...
2SG	mãʒ	mãʒa	mãʒəʁa	...
3SG	mãʒ	mãʒa	mãʒəʁa	...
1PL	mãʒɔ̃	mãʒam	mãʒəʁɔ̃	...
2PL	mãʒe	mãʒat	mãʒəʁe	...
3PL	mãʒ	mãʒEʁ	mãʒəʁɔ̃	...

3- long format

LEXEME	CELL	form
CHANTER	PRS.1SG	ʃãt
CHANTER	PRS.2SG	ʃãt
CHANTER	PRS.3SG	ʃãt
CHANTER	PRS.1PL	ʃãtɔ̃
PELER	PRS.1SG	pɛl
...	...	...

The forms table

lexeme	cell	phon_form	orth_form
CHANTER	PRS.1SG	ʃ ɑ̃ t	chante
CHANTER	PRS.2SG	ʃ ɑ̃ t	chantes
CHANTER	PRS.3SG	ʃ ɑ̃ t	chante
CHANTER	PRS.1PL	ʃ ɑ̃ t ɔ̃	chantons
CHANTER	PRS.2PL	ʃ ɑ̃ t e	chantez
CHANTER	PRS.3PL	ʃ ɑ̃ t	chantent
PELER	PRS.1SG	p ɛ l	pèle
PELER	PRS.2SG	p ɛ l	pèles
…	…	…	

- Rows: inflected forms
- Form-level info:
    - Other forms
    - Identifiers
    - Analyses
    - Comments
    - & any ad-hoc columns
- csv format

The forms table

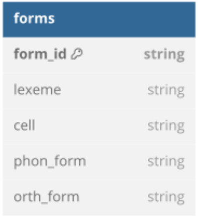
forms.csv

```
lexeme,cell,phon_form,orth_form
chanter,prs.1sg,ʃ ɑ̃ t,chante
chanter,prs.2sg,ʃ ɑ̃ t,chantes
chanter,prs.3sg,ʃ ɑ̃ t,chante
chanter,prs.1pl,ʃ ɑ̃ t ɔ̃,chantons
chanter,prs.2pl,ʃ ɑ̃ t e,chantez
chanter,prs.3pl,ʃ ɑ̃ t,chantent
peler,prs.1sg,p ɛ l,pèle
peler,prs.2sg,p ɛ l,pèles
```

- Rows: inflected forms
- Form-level info:
  - Other forms
  - Identifiers
  - Analyses
  - Comments
  - & any ad-hoc columns
- csv format

Relational schema

form_id	lexeme	cell	phon_form	orth_form
f1	CHANTER	PRS.IND.1.SG	ʃ ɑ̃ t	chante
f4	CHANTER	PRS.IND.1.PL	ʃ ɑ̃ t ɔ̃	chantons
f60	PELER	PRS.IND.1.SG	p ɛ l	pèle
f64	PELER	PRS.IND.1.PL	p ø l ɔ̃	pelons
f90	FINIR	PRS.IND.1.SG	f i n i	finis
f94	FINIR	PRS.IND.1.PL	f i n i s ɔ̃	finis

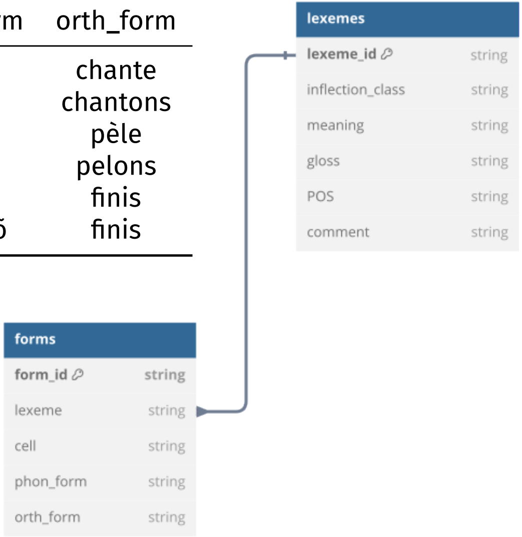


forms	
form_id 🔑	string
lexeme	string
cell	string
phon_form	string
orth_form	string

Relational schema

form_id	lexeme	cell	phon_form	orth_form
f1	CHANTER	PRS.IND.1.SG	ʃ ɑ̃ t	chante
f4	CHANTER	PRS.IND.1.PL	ʃ ɑ̃ t ɔ̃	chantons
f60	PELER	PRS.IND.1.SG	p ɛ l	pèle
f64	PELER	PRS.IND.1.PL	p ø l ɔ̃	pelons
f90	FINIR	PRS.IND.1.SG	f i n i	finis
f94	FINIR	PRS.IND.1.PL	f i n i s ɔ̃	finis

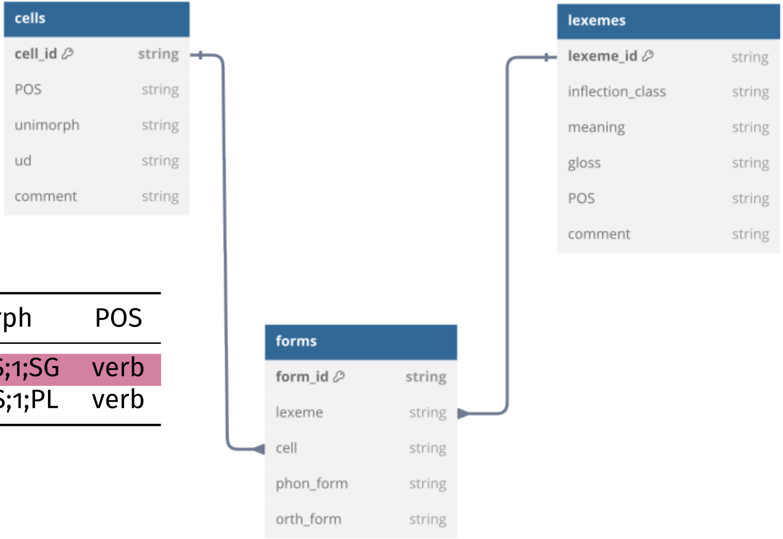
lexeme_id	inflection_class	gloss
CHANTER	groupe-1	to eat
PELER	groupe-1	to peel
FINIR	groupe-2	to end



lexemes	
lexeme_id 🔑	string
inflection_class	string
meaning	string
gloss	string
POS	string
comment	string

forms	
form_id 🔑	string
lexeme	string
cell	string
phon_form	string
orth_form	string

Relational schema



cells	
cell_id	string
POS	string
unimorph	string
ud	string
comment	string

lexemes	
lexeme_id	string
inflection_class	string
meaning	string
gloss	string
POS	string
comment	string

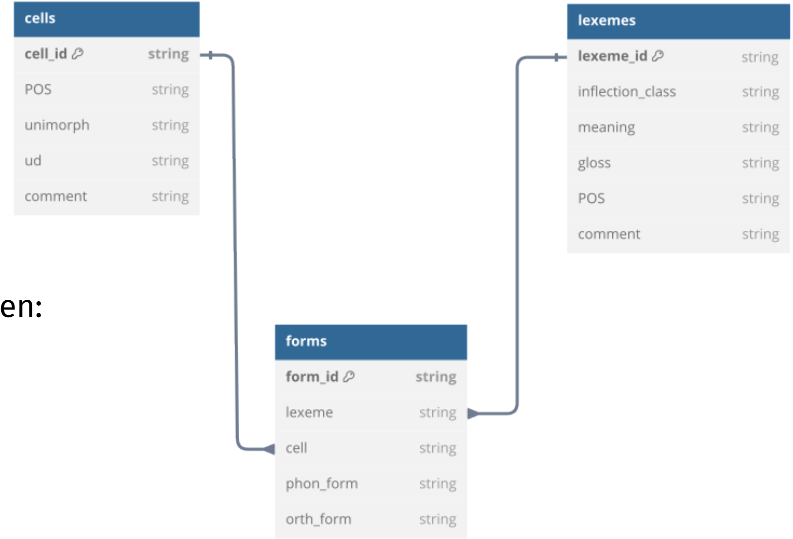
forms	
form_id	string
lexeme	string
cell	string
phon_form	string
orth_form	string

cell_id	unimorph	POS
IND.PRS.1.SG	V;IND;PRS;1;SG	verb
IND.PRS.1.PL	V;IND;PRS;1;PL	verb

Relational schema

cell_id	GRACE	flexique	unimorph	ud	ftb
COND.PRS.1.PL	Vmcp1p-	cond.1pl	V;COND;1;PL	Mood=Cnd|Number=Plur|Person=1|Tense=Pres	m=cond|n=p|p=1|t=pst
COND.PRS.1.SG	Vmcp1s-	cond.1sg	V;COND;1;SG	Mood=Cnd|Number=Sing|Person=1|Tense=Pres	m=cond|n=s|p=1|t=pst
COND.PRS.2.PL	Vmcp2p-	cond.2pl	V;COND;2;PL	Mood=Cnd|Number=Plur|Person=2|Tense=Pres	m=cond|n=p|p=2|t=pst
COND.PRS.2.SG	Vmcp2s-	cond.2sg	V;COND;2;SG	Mood=Cnd|Number=Sing|Person=2|Tense=Pres	m=cond|n=s|p=2|t=pst
COND.PRS.3.PL	Vmcp3p-	cond.3pl	V;COND;3;PL	Mood=Cnd|Number=Plur|Person=3|Tense=Pres	m=cond|n=p|p=3|t=pst
COND.PRS.3.SG	Vmcp3s-	cond.3sg	V;COND;3;SG	Mood=Cnd|Number=Sing|Person=3|Tense=Pres	m=cond|n=s|p=3|t=pst
IND.FUT.1.PL	Vmif1p-	fut.1pl	V;IND;FUT;1;PL	Mood=Ind|Number=Plur|Person=1|Tense=Fut	m=ind|n=p|p=1|t=fut
IND.FUT.1.SG	Vmif1s-	fut.1sg	V;IND;FUT;1;SG	Mood=Ind|Number=Sing|Person=1|Tense=Fut	m=ind|n=s|p=1|t=fut
IND.FUT.2.PL	Vmif2p-	fut.2pl	V;IND;FUT;2;PL	Mood=Ind|Number=Plur|Person=2|Tense=Fut	m=ind|n=p|p=2|t=fut
IND.FUT.2.SG	Vmif2s-	fut.2sg	V;IND;FUT;2;SG	Mood=Ind|Number=Sing|Person=2|Tense=Fut	m=ind|n=s|p=2|t=fut
IND.FUT.3.PL	Vmif3p-	fut.3pl	V;IND;FUT;3;PL	Mood=Ind|Number=Plur|Person=3|Tense=Fut	m=ind|n=p|p=3|t=fut
IND.FUT.3.SG	Vmif3s-	fut.3sg	V;IND;FUT;3;SG	Mood=Ind|Number=Sing|Person=3|Tense=Fut	m=ind|n=s|p=3|t=fut
IMP.PRS.1.PL	Vmmp1p-	imp.1pl	V;POS;IMP;1;PL	Mood=Imp|Number=Plur|Person=1|Tense=Pres	m=Imp|n=p|p=1|t=pst
IMP.PRS.2.PL	Vmmp2p-	imp.2pl	V;POS;IMP;2;PL	Mood=Imp|Number=Plur|Person=2|Tense=Pres	m=Imp|n=p|p=2|t=pst
IMP.PRS.2.SG	Vmmp2s-	imp.2sg	V;POS;IMP;2;SG	Mood=Imp|Number=Sing|Person=2|Tense=Pres	m=Imp|n=s|p=2|t=pst
INF	Vmn—-	inf	V;NFIN	VerbForm=Inf	m=inf
IND.IPFV.1.PL	Vmii1p-	ipfv.1pl	V;IND;PST;1;PL;IPFV	Mood=Ind|Number=Plur|Person=1|Tense=Imp	m=ind|n=p|p=1|t=Imp
IND.IPFV.1.SG	Vmii1s-	ipfv.1sg	V;IND;PST;1;SG;IPFV	Mood=Ind|Number=Sing|Person=1|Tense=Imp	m=ind|n=s|p=1|t=Imp
IND.IPFV.2.PL	Vmii2p-	ipfv.2pl	V;IND;PST;2;PL;IPFV	Mood=Ind|Number=Plur|Person=2|Tense=Imp	m=ind|n=p|p=2|t=Imp
IND.IPFV.2.SG	Vmii2s-	ipfv.2sg	V;IND;PST;2;SG;IPFV	Mood=Ind|Number=Sing|Person=2|Tense=Imp	m=ind|n=s|p=2|t=Imp
...	...	...	...	...	

Relational schema



cells	
cell_id	string
POS	string
unimorph	string
ud	string
comment	string

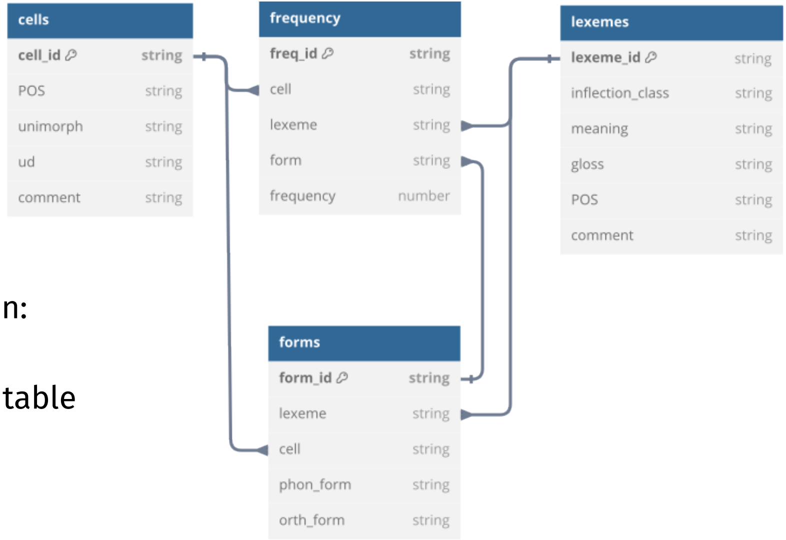
lexemes	
lexeme_id	string
inflection_class	string
meaning	string
gloss	string
POS	string
comment	string

forms	
form_id	string
lexeme	string
cell	string
phon_form	string
orth_form	string

frequency can be given:

- Directly in tables

Relational schema



cells	
cell_id	string
POS	string
unimorph	string
ud	string
comment	string

frequency	
freq_id	string
cell	string
lexeme	string
form	string
frequency	number

lexemes	
lexeme_id	string
inflection_class	string
meaning	string
gloss	string
POS	string
comment	string

forms	
form_id	string
lexeme	string
cell	string
phon_form	string
orth_form	string

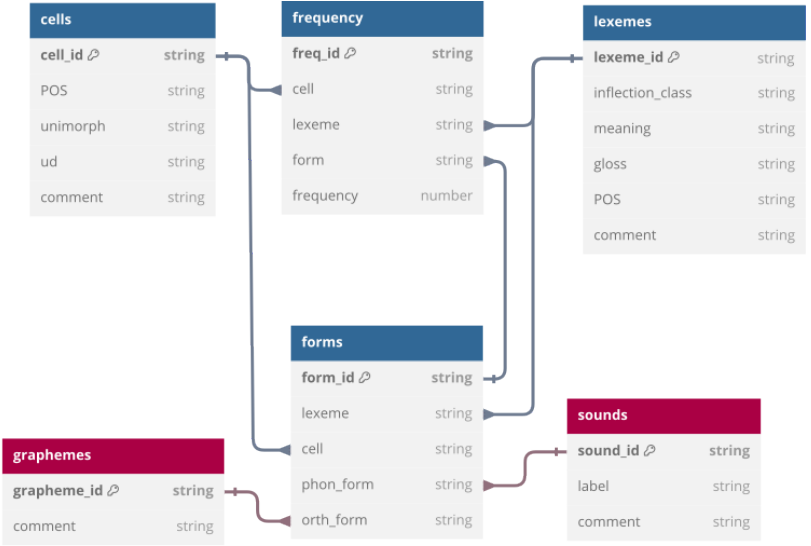
frequency can be given:

- Directly in tables
- or by a dedicated table

Vocabulary relations

form_id	lexeme	cell	phon_form
form_158742	mésuser	ptcp.pst.f.pl	m E z y z e
form_819	aboyer	ind.prs.1.pl	a b w a j ɔ̃
form_41745	chroniquer	cond.prs.1.sg	k ʁ O n i k ə ʁ E
form_91334	détricoter	ind.pst.3.pl	d E t ʁ i k O t E ʁ
form_197935	refleurir	imp.prs.2.sg	ʁ ə f l Ø ʁ i
form_122951	galvaniser	ind.pst.3.sg	g a l v a n i z a
form_11785	anoblir	ind.prs.2.pl	a n O b l i s e
form_99328	encourager	ind.prs.2.pl	ɑ̃ k u ʁ a ʒ e
form_237143	surprendre	sbjv.prs.1.pl	s y ʁ p ʁ ə n j ɔ̃
...	...	...	...

Vocabulary relations



cells	
cell_id 🔑	string
POS	string
unimorph	string
ud	string
comment	string

frequency	
freq_id 🔑	string
cell	string
lexeme	string
form	string
frequency	number

lexemes	
lexeme_id 🔑	string
inflection_class	string
meaning	string
gloss	string
POS	string
comment	string

forms	
form_id 🔑	string
lexeme	string
cell	string
phon_form	string
orth_form	string

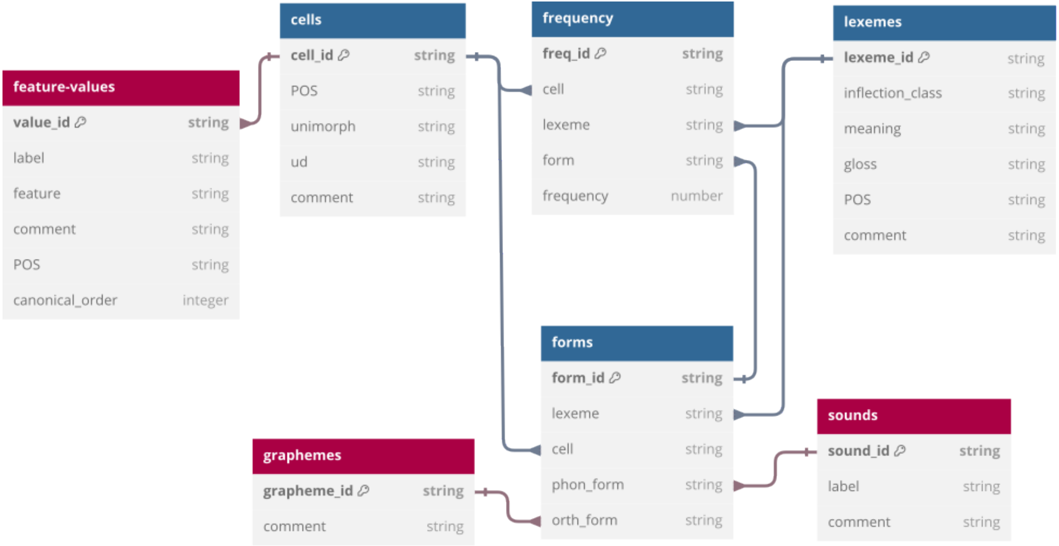
graphemes	
grapheme_id 🔑	string
comment	string

sounds	
sound_id 🔑	string
label	string
comment	string

Vocabulary relations

value_id	label	feature	POS	canonical_order
inf	infinitive	Mode	verb	1
ind	indicative	Mode	verb	2
sbjv	subjunctive	Mode	verb	3
cond	conditional	Mode	verb	4
imp	imperative	Mode	verb	5
ptcp	participle	Mode	verb	6
sg	singular	Number	verb	1
pl	plural	Number	verb	2
1	first person	Person	verb	1
2	second person	Person	verb	2
3	third person	Person	verb	3
prs	present	Tense	verb	1
fut	future	Tense	verb	2
...	...	...	...	

Vocabulary relations



feature-values	
value_id 🔑	**string**
label	string
feature	string
comment	string
POS	string
canonical_order	integer

cells	
cell_id 🔑	**string**
POS	string
unimorph	string
ud	string
comment	string

frequency	
freq_id 🔑	**string**
cell	string
lexeme	string
form	string
frequency	number

lexemes	
lexeme_id 🔑	string
inflection_class	string
meaning	string
gloss	string
POS	string
comment	string

forms	
form_id 🔑	**string**
lexeme	string
cell	string
phon_form	string
orth_form	string

graphemes	
grapheme_id 🔑	**string**
comment	string

sounds	
sound_id 🔑	**string**
label	string
comment	string

Overabundance

form_id	lexeme	cell	phon_form
f1	dream	pst	d r ɛ m t
f2	dream	pst	d r iː m d
f3	learn	pst	l ɜː n d
f4	learn	pst	l ɜː n t
f5	leap	pst	l ɛ p t
f6	leap	pst	l iː p t
f7	sweat	pst	s w ɛ t
f8	sweat	pst	s w ɛ t ɪ d

Overabundance

form_id	lexeme	cell	phon_form
f1	dream	pst	d r ɛ m t
f2	dream	pst	d r iː m d
f3	learn	pst	l ɜː n d
f4	learn	pst	l ɜː n t
f5	leap	pst	l ɛ p t
f6	leap	pst	l iː p t
f7	sweat	pst	s w ɛ t
f8	sweat	pst	s w ɛ t ɪ d

Overabundance

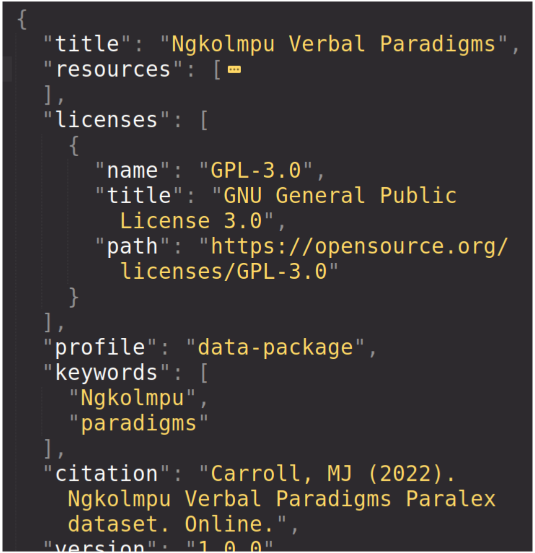
form_id	lexeme	cell	phon_form	overabundance_tag
f1	dream	pst	d r ɛ m t	irreg
f2	dream	pst	d r iː m d	d-form
f3	learn	pst	l ɜː n d	d-form
f4	learn	pst	l ɜː n t	t-form
f5	leap	pst	l ɛ p t	irreg
f6	leap	pst	l iː p t	t-form
f7	sweat	pst	s w ɛ t	irreg
f8	sweat	pst	s w ɛ t ɪ d	d-form

Metadata

Metadata are information

- About a dataset
  - Authors, Contributors
  - Title
  - Relation to other datasets
- About its structure
  - Conventions & coding
  - Meaning and expectations for each table, column

Frictionless metadata



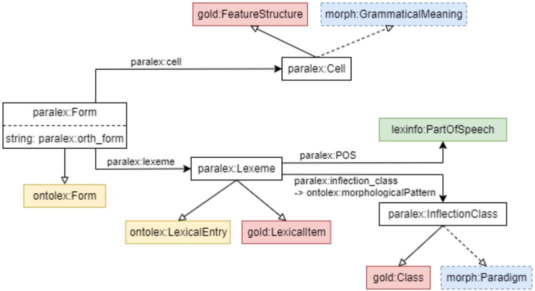
```
{
  "title": "Ngkolmpu Verbal Paradigms",
  "resources": [⋯
  ],
  "licenses": [
    {
      "name": "GPL-3.0",
      "title": "GNU General Public
       License 3.0",
      "path": "https://opensource.org/
       licenses/GPL-3.0"
    }
  ],
  "profile": "data-package",
  "keywords": [
    "Ngkolmpu",
    "paradigms"
  ],
  "citation": "Carroll, MJ (2022).
   Ngkolmpu Verbal Paradigms Paralex
   dataset. Online.",
  "version": "1.0.0"
```

FRICTIONLESS DATA

- Separate file, distributed with the data
- Machine readable format: `json`
- Can be generated
- Allows for rigorous validation

`https://frictionlessdata.io/`

Ontology



- RDF `classes` ↦ tables,
- RDF `properties` ↦ columns
- Links to GOLD; OntoLex; Lexinfo

- Enables conversion to OntoLex-compliant RDF lexicons.
  - Interoperable with lexical resources of different kinds (e.g. corpora, dictionaries)
  - See the **PrinParLat** poster

Tools

- **To generate metadata:** python ‘paralex‘ package (R package planned)
- **To validate data:** Frictionless software
- **To publish data websites:** a paralex plugin for mkdocs (currently being prepared)
- **To convert to ontolex:** Generic tools to come

Conclusion

Conclusion: Summary

- **Paralex standard** for tabular lexicons of inflected forms
    - Formal conventions on structure
    - High flexibility regarding analytic choices

- **Good data principles:** FAIR for machine readability; Open Data for science; CARE for minoritized communities and DeAR for researchers.

- **to come**: static sites; R package; sets of lexicons.

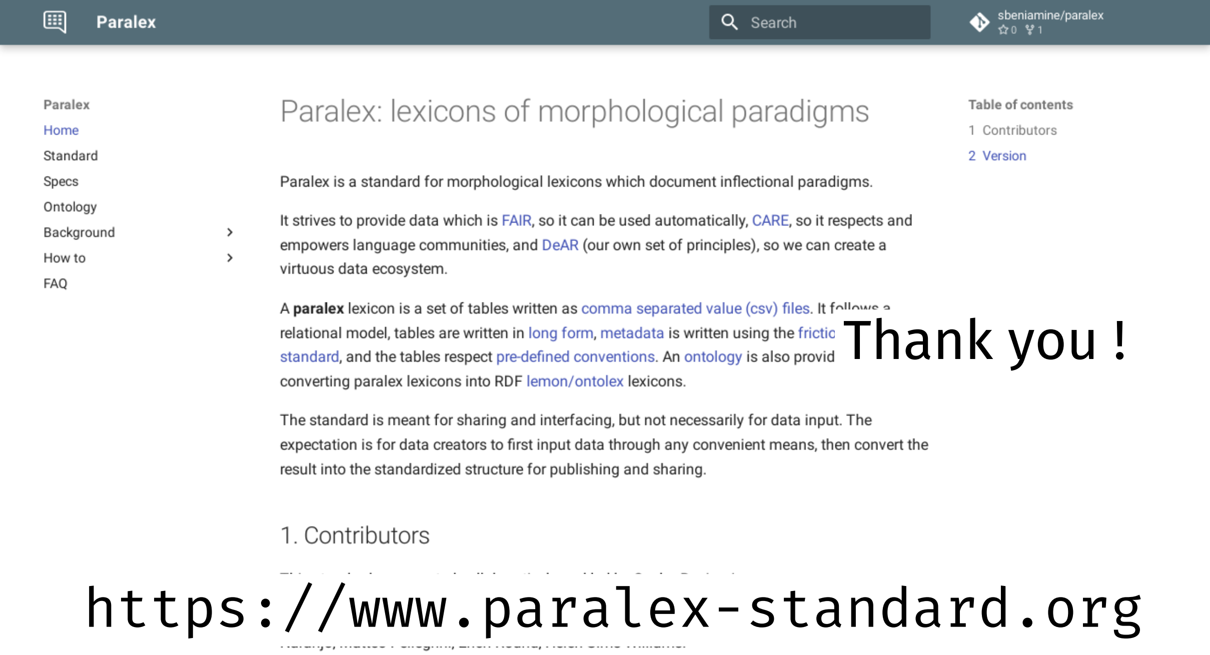
Conclusion: Benefits

For **data creators**:

- Helps us publish high quality data
- Minimize the cost of maintenance

For **data users**:

- Usable for both quantitative & qualitative work
- Long lasting and up to date data
- Easy to share & reuse , crucial to incremental science



Paralex: lexicons of morphological paradigms

Table of contents

Paralex is a standard for morphological lexicons which document inflectional paradigms.

It strives to provide data which is FAIR, so it can be used automatically, CARE, so it respects and empowers language communities, and DeAR (our own set of principles), so we can create a virtuous data ecosystem.

A **paralex** lexicon is a set of tables written as comma separated value (csv) files. It follows a relational model, tables are written in long form, metadata is written using the frictio standard, and the tables respect pre-defined conventions. An ontology is also provid converting paralex lexicons into RDF lemon/ontolex lexicons.

The standard is meant for sharing and interfacing, but not necessarily for data input. The expectation is for data creators to first input data through any convenient means, then convert the result into the standardized structure for publishing and sharing.

1. Contributors

Naranjo, Matteo Pellegrini, Erich Round, Helen Sims-Williams.

Thank you !

https://www.paralex-standard.org