



Wikinflection:

Massive semi-supervised generation of multilingual inflectional corpus from Wiktionary

Eleni Metheniti and Günter Neumann

December 14, 2018

Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI)

Breaking down the title

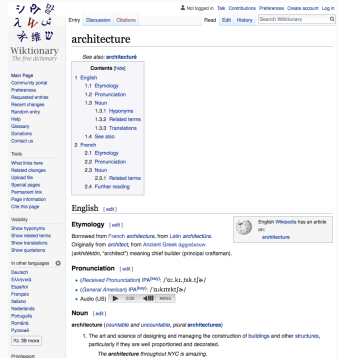
- **Massive:** large-scale
- **Semi-supervised generation:** generation with minimal human interference/labour
- **Multilingual inflectional corpus:** a corpus of the inflectional paradigms of nouns, adjectives, verbs from 140+ languages
- **from Wiktionary:** using the English version of Wiktionary (5k source languages, target language is English)

Introduction

A dictionary named Wiktionary

WIKTIONARY is a multilingual dictionary, where every lemma has:

- Sections per source language
- Pronunciation
- Etymology
- Definition
- Inflection
- Derivatives
- Translations
- Semantic information
- etc.



The screenshot shows the Wiktionary page for the lemma 'architecture'. The page is in English and includes a navigation bar at the top with links for 'Entry', 'Discussion', and 'Citations'. Below the navigation bar, the word 'architecture' is displayed in a large font. A 'Contents' box lists various sections: 1. English (1.1 Etymology, 1.2 Pronunciation, 1.3 Noun, 1.3.1 Homonyms, 1.3.2 Related terms, 1.3.3 Translations), 1.4 See also, 2. French (2.1 Etymology, 2.2 Pronunciation, 2.3 Noun, 2.3.1 Related terms, 2.4 Further reading). The 'Etymology' section states that the word is borrowed from French 'architecture', from Latin 'architectura', originally from 'architect', from Ancient Greek 'ἀρχιτέκτων' (architekton, 'architect') meaning chief builder (principal craftsman). The 'Pronunciation' section shows the IPA [ˌɑːrkiˈtɛktʃər] and provides audio and video links. The 'Noun' section defines 'architecture' as (countable and uncountable, plural 'architectures') the art and science of designing and managing the construction of buildings and other structures, particularly if they are well proportioned and decorated. A note mentions 'The architecture throughout NYC is amazing.'

Page for lemma 'architecture'.

The two sides of the same dictionary

WEB VERSION:

- human-readable information
- for human access
- static
- generated by XML files and server
- accessible online

Portuguese [edit]

Alternative forms [edit]

- *fallar* (obsolete)
- *falá* (apocopic or eye dialect)

Etymology [edit]

From Old Portuguese *falar*, from Vulgar Latin **fāb(u)lāre*, from Latin *fābulārī*, present infinitive of *fābulor* (“chat, converse”). Doublet of the borrowing *fabular*.

Pronunciation [edit]

- (Portugal) IPA^(nɐv): /feˈlaɾ/
- (Brazil) IPA^(nɐv): /faˈla(ɾ)/
- (Nordestino) IPA^(nɐv): /faˈla(h)/
- (Su) IPA^(nɐv): /faˈlaɹ/, /faˈlaɻ/

This web page for ‘falar’...

XML DUMP VERSION:

- machine-readable information
- for server and experts
- dynamic
- can generate web pages
- accessible offline

```
==Portuguese==
```

```
===Alternative forms===
```

```
* {{l|pt|fallar}} {{qualifier|obsolete}}
* {{l|pt|falá}} {{qualifier|apocopic or eye dialect}}
```

```
===Etymology===
```

```
From {{etyl|roa-opt|pt}} {{m|roa-opt|falar}}, from
{{etyl|la|pt}} {{m|la|fābulāri}}, present infinitive of
{{m|la|fābulor|chat, converse}}.
```

```
===Pronunciation===
```

```
* {{a|PT}} {{IPA|/feˈlaɾ/|lang=pt}}
* {{a|BR}} {{IPA|/faˈla(ɾ)/|lang=pt}}
* {{a|Nordestino}} {{IPA|/faˈla(h)/|lang=pt}}
* {{a|Su}} {{IPA|/faˈlaɹ/|/faˈlaɻ/|lang=pt}}
```

...Fis generated by this XML page.

Wiktionary and NLP

- Wiktionary is a widely used resource for NLP/NLG/NLU applications
- Advisable by Wiktionary to use XML files to avoid server load
- Many available tools to parse the XML file [5, 1, 7, 4]
- Easy to access some info, e.g. phonology...
- ... but what about inflection? [6]

```
==Portuguese==

===Alternative forms===
* {{\|pt|fallar}} {{qualifier|obsolete}}
* {{\|pt|falá}} {{qualifier|apocopic or eye dialect}}

===Etymology===
From {{etyl|roa-opt|pt}} {{m|roa-opt|falar}}, from {{etyl|la|pt}}
infinitive of {{m|la|fabulor|chat, converse}}.

===Pronunciation===
* {{a|PT}} {{IPA|/feˈlaɾ/|lang=pt}}
* {{a|BR}} {{IPA|/faˈla(ɾ)/|lang=pt}}
* {{a|Nordestino}} {{IPA|/faˈla(h)/|lang=pt}}
* {{a|Sul}} {{IPA|/faˈla.ɫ/|faˈla.ɫ/|lang=pt}}

===Verb===
{{pt-verb|fal|ar}}

# {{\|pt|intransitive}} to {{\|en|speak}}; to {{\|en|talk}}
#: {{ux|pt|Para de ''falar''.|Stop ''talking''.|inline=1}}
#: {{ux|pt|''Fala''!|''Talk''!|inline=1}}
#: {{ux|pt|''Fale''!|''Talk''!|inline=1}}
# {{\|pt|by extension}} to {{\|en|communicate}} by any means
#: {{ux|pt|''Falamo''-nos por correio.|We ''communicate'' by n}}
# {{\|pt|transitive}} to {{\|en|say}} something
#: {{ux|pt|Para de ''falar'' bobagens.|Stop ''talking'' nonsense.}}
#: {{ux|pt|''Fala'' bobagens.|''Talk'' nonsense.}}
#: {{ux|pt|Estou ''falando'' com você|I'm ''talking'' to you.}}
#: {{indtr|pt|para}} to {{\|en|tell}} {{gloss|to convey by speech}}
#: {{ux|pt|Vou ''falar'' para você.|I'm going to ''tell'' you.}}
# {{indtr|pt|de|sobre}} to {{\|en|talk}} about
# {{indtr|pt|de}} to {{\|en|speak ill of}}
# {{\|pt|transitive}} to {{\|en|speak}} {{gloss|to be able to con}}
#: {{ux|pt|Em Portugal se ''fala'' português.|In Portugal they ''
```



Liebeck and Conrad (2015) [3]: **IWNLP**

- Parser for German Wiktionary
- Re-implement templates from **Lua** to **C#**
- Inflection for some classes of nouns, adjectives, verbs

PROS:

- Very high quality
- Able to generate inflectional paradigms
- Uses only offline XML dump file

CONS:

- Only for German language and Wiktionary
- A lot of manual labour, hard to extend
- Not all templates are adapted

Kirov et al. (2016) [2]: UniMorph

- Multilingual corpus of inflected wordforms
- Pulling information from en.wiktionary.org, no XML dump file
- (2016) ~1M inflected forms, (2018) ~9M inflected forms
- Tagging with *UniMorph* schema

PROS:

- Very large, multilingual
- Includes tags
- Biggest open-source inflectional resource available, still growing

CONS:

- Pulling from online is bad practice
- Tagged wordforms but not organized in paradigms
- Non-reproducible

Our research questions

- Can we reverse-engineer the Wiktionary like Liebeck and Conrad...
- ... but in a large-scale and multilingually like Kirov et al. ...
- ... and in a reproducible, extendable, unsupervised way?

Reverse-engineering the (English) Wiktionary

Web page generation

This part of the XML file...

```
==Portuguese==
===Alternative forms===
* {{l|pt|falar}} {{qualifier|obsolete}}
* {{l|pt|falá}} {{qualifier|apocopic or eye dialect}}

===Etymology===
From {{etyl|roa-opt|pt}} ({{n|roa-opt|falar}}, from {{etyl|la|pt}} ({{n|la|fabulári}}, present infinitive of {{n|la|fabulari}}|chat, converse)).

===Pronunciation===
* {{a|PT}} ({{IPA|/feˈla.z/|lang=pt}})
* {{a|BR}} ({{IPA|/faˈla.zi/|lang=pt}})
* {{a|Nordestino}} ({{IPA|/faˈla(h)/|lang=pt}})
* {{a|Sul}} ({{IPA|/faˈla.zi/|lang=pt}})

===Verb===
{{pt-verb|falar}}

# {{l|pt|intransitive}} to {{l|en|speak}}; to {{l|en|talk}} ({{gloss|to say words out loud}})
#: {{ux|pt|Para de ''falar''|.Stop ''talking''|.inline=1}}
#: {{ux|pt|''Fala''|''|''|''Talk''|''|''|.inline=1}}
#: {{ux|pt|''Fale''|''|''|''Talk''|''|''|.inline=1}}
# {{l|pt|by extension}} to {{l|en|communicate}} by any means
#: {{ux|pt|''Falano''|-nos por correio.|We ''communicate'' by mail|.inline=1}}
# {{l|pt|transitive}} to {{l|en|say}} something
#: {{ux|pt|Para de ''falar''|''bobagens|.Stop ''talking''|''nonsense|.inline=1}}
#: {{ux|pt|''Fala''|''bobagens|.''Talk''|''nonsense|.inline=1}}
# {{l|pt|com}} to {{l|en|talk}} ({{l|en|to}})
#: {{ux|pt|Estou ''falando''|'' com você|I'm ''talking''|'' to you|.inline=1}}
# {{l|pt|para}} to {{l|en|tell}} ({{gloss|to convey by speech}})
#: {{ux|pt|Vou ''falar''|'' para você.|I'm going to ''tell''|'' you|.inline=1}}
# {{l|pt|de|sobre}} to {{l|en|talk}} about
# {{l|pt|de}} to {{l|en|speak ill of}}
# {{l|pt|transitive}} to {{l|en|speak}} ({{gloss|to be able to communicate in a language}})
#: {{ux|pt|Em Portugal se ''fala''|'' português.|In Portugal they ''speak''|'' Portuguese.}}
```

```
====Conjugation====
{{pt-conj|falar}}
```



```
====Conjugation====
{{pt-conj|falar}}
```

... generates this conjugational table on the web page.

Conjugation [[edit](#)]

Conjugation of the Portuguese -ar verb falar [info]						
Notes: [info]						
• This is a regular verb of the -ar group.						
• Verbs with this conjugation include: amar, cantar, girar, marchar, mostrar, nadar, parar, participar, retirar, separar, viajar.						
	Singular			Plural		
	First-person (eu)	Second-person (tu)	Third-person (ele / ela / você)	First-person (nós)	Second-person (vós)	Third-person (eles / elas / vocês)
Infinitive	falar					
Impersonal	falar					
Personal	falar	falares	falar	falamos	falades	falarem
Gerund	falando					
Past participle	falado					
Maximale	falado			falados		
Feminine	falada			faladas		
Indicative						
Present	falo	falas	fala	falamos	falais	falam
Imperfect	falava	falavas	falava	falávamos	faláveis	falavam
Preterite	falei	falaste	falou	falamos	falastes	falaram
Pluperfect	falara	falaras	falara	faláramos	faláreis	falaram
Future	falarei	falarás	falará	falaremos	falareis	falarão
Conditional						
	falaria	falarías	falaria	falaríamos	falaríeis	falaríam
Subjunctive						
Present	fale	fales	fale	falemos	faleis	falem
Imperfect	falasse	falasses	falasse	falássemos	falásseis	falássem
Future	falar	falares	falar	faláremos	faláreis	falárem
Imperative						
Affirmative	-	fala	fale	falemos	falai	falem
Negative (não)	-	fales	fale	falemos	faleis	falem

Templates

`{{pt-conj|fal|ar}}`

is a **dynamic link** to a **template** with its required parameters.

- `pt-conj`: template for verb conjugation in Portuguese
- `fal`: stem of the word
- `ar`: conjugation class

A **template** has its own XML page.

But where is the conjugational information in the page?

```
<page>
  <title>Template:pt-conj</title>
  <ns>10</ns>
  <id>1294753</id>
  <revision>
    <id>32142499</id>
    <parentid>13609370</parentid>
    <timestamp>2015-01-28T14:10:20Z</timestamp>
    <contributor>
      <username>Jberkel</username>
      <id>1580588</id>
    </contributor>
    <comment>use module</comment>
    <model>wikitext</model>
    <format>text/x-wiki</format>
    <text xml:space="preserve">&lt;includeonly&gt;
      <div style="border: 1px solid orange; padding: 2px;">
        <code>{{#invoke:pt-conj|show}}</code>
      </div>
      &lt;/includeonly&gt;&lt;
    </text>
    <sha1>3h1y5upf3gm5iwns5dyhnkyxgfqwjos</sha1>
  </revision>
</page>
```

Modules

```
{{#invoke:pt-conj|show}}
```

is a **dynamic link** to a **module**.

The required parameters are passed by the template parameters. The module also requires additional info (other conjugational classes, table templates etc).

Language: **Lua**

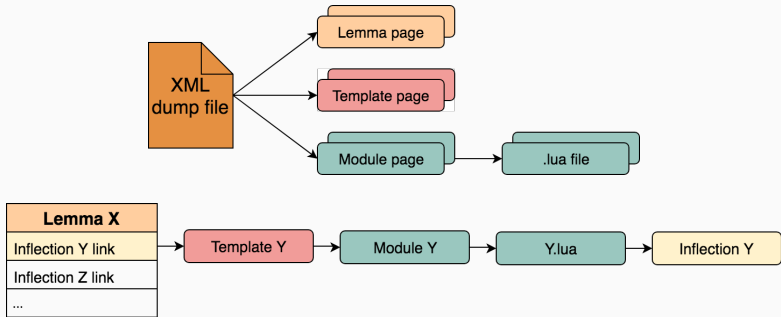
```
local exports = {}

local function verbData(ending)
    local group
    if ending == 'pôr' or ending == 'por' then
        group = 'er'
    elseif ending == 'erir-defective' then
        group = 'ir'
    else
        group, _ = string.gsub(ending, "%d+$", "")
        group = string.sub(group, #group-1)
    end
    if group == "" then
        return nil
    end
    local success, m_verb_data = pcall(require, "Module:pt-conj/data/.."..group)
    if success and m_verb_data[ending] then
        return mw.clone(m_verb_data[ending])
    else
        return nil
    end
end

local function applyFuncToTableValues(tbl, func)
    for k,v in pairs(tbl) do
        if type(v) == 'table' then
            applyFuncToTableValues(v, func)
        else
            tbl[k] = func(v)
        end
    end
end
```

— *stem* (required if applicable): beginning of the verb. All characters of the
— *ending* (required): Ending of the verb. The last characters chosen specifica
— *compound* (required if applicable): Compound words. Text to be added after t
function **exports.infect**(stem, ending, compound)
local data = verbData(ending)

Reverse-engineering the Wiktionary: Attempt 1



Unsupervised generation is **not possible**.
Missing information to run `.lua` script successfully.

Back to the drawing block...

How do generated templates look like online?

en.wiktionary.org/wiki/Template:pt-conj

This template generates a navigation box for [Portuguese verb](#) conjugation entries.
The actual work is done by [Module:pt-conj](#).

en.wiktionary.org/wiki/Template:pl-decl-adj-owy

declension of {{{1}}}owy [hide ▲]						
case	singular				plural	
	<i>m pers, m anim</i>	<i>m inan</i>	<i>n</i>	<i>f</i>	<i>m pers</i>	other
nominative, vocative	{{{1}}}owy		{{{1}}}owe	{{{1}}}owa	{{{1}}}owi	{{{1}}}owe
genitive	{{{1}}}owego			{{{1}}}owej	{{{1}}}owych	
dative	{{{1}}}owemu				{{{1}}}owym	
accusative	{{{1}}}owego	{{{1}}}owy	{{{1}}}owe	{{{1}}}ową	{{{1}}}owych	{{{1}}}owe
instrumental	{{{1}}}owym				{{{1}}}owymi	
locative					{{{1}}}owej	

Generated templates

XML file > lemma 'rózowy' > $\{\{p\ell\text{-decl-adj-owy} \mid r\acute{o}\acute{z}\}\}$
generates this:

declension of <i>różowy</i> [hide ▲]						
case	singular				plural	
	<i>m pers, m anim</i>	<i>m inan</i>	<i>n</i>	<i>f</i>	<i>m pers</i>	other
nominative, vocative	rózowy		różowe	różowa	różowi	różowe
genitive	rózowego			różowej	różowych	
dative	różowemu				różowym	
accusative	różowego	różowy	różowe	różową	różowych	różowe
instrumental	różowym				różowymi	
locative	różowym			różowej	różowych	

0: template name, 1: stem

Generated templates

en.wiktionary.org/wiki/Template:lt-conj-1

conjugation of lt-conj-1		singular (vienaskaita)			plural (daugiskaita)		
		1 st person (pirmasis asmuo)	2 nd person (antrasis asmuo)	3 rd person (trečiasis asmuo)	1 st person (pirmasis asmuo)	2 nd person (antrasis asmuo)	3 rd person (trečiasis asmuo)
		aš	tu	jis/ji	mes	jūs	jie/jos
indicative (tiesioginė nuosaka)	present (esamasis laikas)	<u>{{{1}}}</u> u	{{{1}}}i	{{{1}}}a	{{{1}}}ame, {{{1}}}am	{{{1}}}ate, {{{1}}}at	{{{1}}}a
	past (būtašis kartinis laikas)	<u>{{{2}}}</u> au	{{{2}}}ai	{{{2}}}o	{{{2}}}ome, {{{2}}}om	{{{2}}}ote, {{{2}}}ot	{{{2}}}o
	past frequentative (būtašis dažninis laikas)	<u>{{{3}}}</u> davau	{{{3}}}davai	{{{3}}}davo	{{{3}}}davome, {{{3}}}davom	{{{3}}}davote, {{{3}}}davot	{{{3}}}davo
	future (būsimasis laikas)	{{{3}}}siu	{{{3}}}si	{{{3}}}s	{{{3}}}sime, {{{3}}}sim	{{{3}}}site, {{{3}}}sit	{{{3}}}s
subjunctive (tariamoji nuosaka)		{{{3}}}čiau	{{{3}}}tum, {{{3}}}tumei	{{{3}}}tų	{{{3}}}tumėme, {{{3}}}tumėm, {{{3}}}tume	{{{3}}}tumėte, {{{3}}}tumėt	{{{3}}}tų
imperative (liepiamoji nuosaka)		—	{{{3}}}k, {{{3}}}ki	te{{{1}}}a, te{{{1}}}ie	{{{3}}}kime, {{{3}}}kim	{{{3}}}kite, {{{3}}}kit	te{{{1}}}a, te{{{1}}}ie

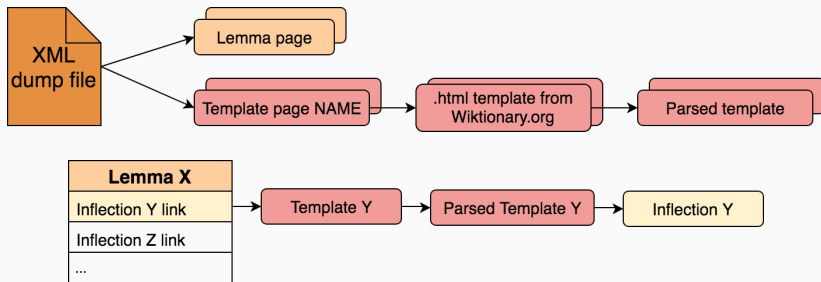
Generated templates

XML file > lemma 'gauti' > {{lt-conj-1|gaun|gav|gau}}
generates this:

		singular (vienaskaita)			plural (daugiskaita)		
		1 st person (pirmasis asmuo)	2 nd person (antrasis asmuo)	3 rd person (trečiasis asmuo)	1 st person (pirmasis asmuo)	2 nd person (antrasis asmuo)	3 rd person (trečiasis asmuo)
		aš	tu	jis/ji	mes	jūs	jie/jos
indicative (tiesioginė nuosaka)	present (esamasis laikas)	<u>ga</u> nu	gauni	gauna	gauname, gaunam	gaunate, gaunat	gauna
	past (būtasīs kartinis laikas)	<u>gav</u> au	gavai	gavo	gavome, gavom	gavote, gavot	gavo
	past frequentative (būtasīs dažninis laikas)	<u>gav</u> avau	gavavai	gavavo	gavavome, gavavom	gavavote, gavavot	gavavo
	future (būsimasis laikas)	gau <u>si</u>	gausi	gaus	gausime, gausim	gausite, gausit	gaus
subjunctive (tariamoji nuosaka)		gau <u>čia</u>	gautum, gautumei	gautų	gautumėme, gautumėm, gautume	gautumėte, gautumėt	gautų
imperative (liepiamoji nuosaka)		—	gauk, gauki	tegauna, tegaunie	gaukime, gaukim	gaukite, gaukit	tegauna, tegaunie

0: template name, 1-3: stem allomorphs

Reverse-engineering the Wiktionary: Attempt 2



Reverse-engineering the Wiktionary: Attempt 2

Goal: create an annotated corpus with:

- morphological information: stem allomorphs, prefixes, suffixes
- morphosyntactic information: UD tags

gauti

Lithuanian:

gaunu	∅	gaun	u	Mood=Ind Number=Sing Person=1 Tense=Pres VerbForm=Fin
gauni	∅	gaun	i	Mood=Ind Number=Sing Person=2 Tense=Pres VerbForm=Fin
gauna	∅	gaun	a	Mood=Ind Number=Sing Person=3 Tense=Pres VerbForm=Fin
gauname	∅	gaun	ame	Mood=Ind Number=Plur Person=1 Tense=Pres VerbForm=Fin
gaunam	∅	gaun	am	Mood=Ind Number=Plur Person=1 Tense=Pres VerbForm=Fin
gaunate	∅	gaun	ate	Mood=Ind Number=Plur Person=2 Tense=Pres VerbForm=Fin
gaunat	∅	gaun	at	Mood=Ind Number=Plur Person=2 Tense=Pres VerbForm=Fin
gauna	∅	gaun	a	Mood=Ind Number=Plur Person=3 Tense=Pres VerbForm=Fin
gavau	∅	gav	au	Mood=Ind Number=Sing Person=1 Tense=Past VerbForm=Fin
gavai	∅	gav	ai	Mood=Ind Number=Sing Person=2 Tense=Past VerbForm=Fin
gavo	∅	gav	o	Mood=Ind Number=Sing Person=3 Tense=Past VerbForm=Fin
gavome	∅	gav	ome	Mood=Ind Number=Plur Person=1 Tense=Past VerbForm=Fin
gavom	∅	gav	om	Mood=Ind Number=Plur Person=1 Tense=Past VerbForm=Fin
[...]				

Inflection Generation

Extracting the lemmata

1. **Find pages** in XML dump file where:
 - there is content (e.g. not template pages)
 - content is a lemma and not an inflected word (e.g. 'houses')
 - content is a lemma and at least one **dynamic link** to a template
2. **Process** the XML code to extract lemma and its dynamic link(s) to templates

5.740.594 word pages → 454.470 lemmata with inflection

Extracting the templates

1. **Find pages** in XML dump file where:
 - there is a template
 - related to inflection and not other linguistic information (e.g. phonology) or utilities (e.g. table generation)
2. Collect the **template titles**
3. Pull the html page for en.wiktionary.org/wiki/<template_name>
4. Process **<template_name>.html** with Python

Template processing

- BeautifulSoup, pandas etc. to parse HTML table
- Custom table tags → UD tags conversion
- Only significant and correctly parsed templates are kept

7.068 downloaded templates → 2.927 parsed templates

```
<th rowspan="4" style="background:#EFEFEF">indicative  
<br /><small><a href="/wiki/tiesioginė%C4%97_nuosaka"  
title="tiesioginė nuosaka">tiesioginė <br />nuosaka/<a  
>/</small>  
</th>  
<th style="background:#EFEFEF">present <br /><small><a  
href="/wiki/esamasis_laikas" title="esamasis  
laikas">esamasis laikas</a></small>  
</th>  
<td><span class="Latn" lang="lt">{{1}}u</span>  
</td>  
<td><span class="Latn" lang="lt">{{1}}i</span>  
</td>  
<td><span class="Latn" lang="lt">{{1}}a</span>  
</td>  
<td><span class="Latn" lang="lt">{{1}}ame</span>,&br/><br /><small>{{1}}am</small>  
</td>  
<td><span class="Latn" lang="lt">{{1}}ate</span>,&br/><br /><small>{{1}}at</small>  
</td>  
<td><span class="Latn" lang="lt">{{1}}a</span>  
</td></tr>
```


```
{'lt-conj-1': [{'1'}u', 'Mood=Ind'],  
  [{'1'}i', 'Mood=Ind'],  
  [{'1'}a', 'Mood=Ind'],  
  [{'1'}ame', 'Mood=Ind'],  
  [{'1'}am', 'Mood=Ind'],  
  [{'1'}ate', 'Mood=Ind'],  
  [{'1'}at', 'Mood=Ind'],  
  [{'1'}a', 'Mood=Ind'],  
  [{'2'}au', 'Tense=Past'],  
  [{'2'}ai', 'Tense=Past'],  
  [{'2'}o', 'Tense=Past'],  
  [{'2'}ome', 'Tense=Past'],  
  [{'2'}om', 'Tense=Past'],  
  [{'2'}ote', 'Tense=Past'],  
  [{'2'}ot', 'Tense=Past'],  
  [{'2'}o', 'Tense=Past'],
```


Templates that didn't make the cut...

Some examples:

- **pt-conj**: does not contain template tables
- **de-decl-noun-m**: not well-written
- **io-conj**: requires external information

	singular			plural	
	indef.	def.	noun	def.	noun
nominative	ein	der	{{{ns}}}	die	{{{np}}}
genitive	eines	des	{{{gs}}}	der	{{{gp}}}
dative	einem	dem	{{{ds}}}	den	{{{dp}}}
accusative	einen	den	{{{as}}}	die	{{{ap}}}

	present	past	future
infinitive	Template:io-coar	Template:io-coir	Template:io-coor
tense	Template:io-coas	Template:io-cois	Template:io-coos
conditional	Template:io-coous		
imperative	Template:io-coez		
adjective active participle	Template:io-coanta	Template:io-cointa	Template:io-coonta

Time to generate the inflection!

1. Use lemma's dynamic link(s) to find the appropriate template, the stem allomorphs, other parameters
2. **Generate** inflected forms, with UD tags, prefixes, suffixes and infixes (null if none), and stem (allomorph)

Results

⇒ 225.453 lemmata, matched with 1.708 templates to generate 8.426.480 inflected forms, in 199 languages

```
'gauti': [[['gaunu', 'lt-conj-1', ['Mood=Ind'], 'VERB', ['', 'u', ''], 'gaun'],  
  ['gauni', 'lt-conj-1', ['Mood=Ind'], 'VERB', ['', 'i', ''], 'gaun'],  
  ['gauna', 'lt-conj-1', ['Mood=Ind'], 'VERB', ['', 'a', ''], 'gaun'],  
  ['gauname', 'lt-conj-1', ['Mood=Ind'], 'VERB', ['', 'ame', ''], 'gaun'],  
  ['gaunam', 'lt-conj-1', ['Mood=Ind'], 'VERB', ['', 'am', ''], 'gaun'],  
  ['gaunate', 'lt-conj-1', ['Mood=Ind'], 'VERB', ['', 'ate', ''], 'gaun'],  
  ['gaunat', 'lt-conj-1', ['Mood=Ind'], 'VERB', ['', 'at', ''], 'gaun'],  
  ['gauna', 'lt-conj-1', ['Mood=Ind'], 'VERB', ['', 'a', ''], 'gaun'],  
  ['gavau', 'lt-conj-1', ['Tense=Past'], 'VERB', ['', 'au', ''], 'gav']]
```

Evaluation

- **Human evaluation** was not possible (or desired), because of volume and non-reproducibility
- Using **corpora** was not possible, because inflected forms are rare/low-frequency
- Our choice: use the English Wiktionary!

Process

1. **Random selection** of one lemma/template →
2. **Generate** inflection →
3. **Pull lemma page** from Wiktionary, check if forms exist →
4. **Delete** incorrect forms from template or entire template

Evaluation Results

- Every evaluation run is randomized and unique
- Results after Random Evaluation 1 and 2:

Template	Random evaluation No. 1			Random evaluation No. 2				
	Word	All	Correct	False	Word	All	Correct	False
la-decl-2nd	<i>campus</i>	12	12	0	<i>Herostratus</i>	12	8	4
de-decl-adj	<i>großbürgerlich</i>	48	48	0	<i>unmöglich</i>	48	48	0
ga-decl-m1	<i>gob</i>	16	12	4	<i>baneachlach</i>	16	12	4
ang-decl-noun-a-n	<i>bispell</i>	8	4	4	<i>gedal</i>	8	4	4
osx-decl-noun-a-n	<i>baluwerk</i>	8	8	0	<i>god</i>	8	8	0
pl-decl-noun-masc-ani	<i>palant</i>	15	15	0	<i>torbacz</i>	15	14	1

Results after Random Evaluation No. 1

⇒ 216.378 lemmata, matched with 1.537 templates to generate 5.970.799 inflected forms

Full tables for 3 random evaluations can be found here.

Comparison to UniMorph

- *UniMorph* is larger (+ high-frequency languages)
- *Wikiflection* covers more low-frequency languages
- *Wikiflection* has more morphological information
- *UniMorph* uses own tags, *Wikiflection* uses UD

<i>Language</i>	<i>UniMorph</i>	<i>Wikiflection</i> (after eval. 3)
<i>Adyge</i>	n/a	440
<i>Albanian</i>	33.483	8.767
<i>Alemannic German</i>	0	232
<i>Ancient Greek</i>	0	3.312
<i>Arabic</i>	140.003	36
<i>Aragonese</i>	0	448
<i>Armenian</i>	338.461	59
<i>Assamese</i>	0	13.790
<i>Asturian</i>	n/a	23.329
<i>Avestan</i>	0	6
SUM	8.850.395	6.024.077

Full table of comparison can be found [here](#).

Conclusion

Wikinflection is an approach to tap into **information previously unexploited**, and to generate an **inflectional corpus** with as little human supervision as possible (so that it can be **replicated** and **extended**).

Successful? Yes... but still not perfect.

- Non-unified style and syntax among contributors/languages
- Still missing information in inflection (**cs**, **rfinfl**, **grc** etc.)
- Wiktionary grows, evolves and revises all the time

- Try other Wiktionary target languages
- Improvement of template table parsing (missing tags)
- (Some) Human evaluation?
- Re-attempt to use modules for high-frequency languages

Code available at:

github.com/lenakmeth/Wikinflexion

Just add XML file!

Questions?



J. Acs, K. Pajkossy, and A. Kornai.

Building basic vocabulary across 40 languages.

In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 52–58, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.



C. Kirov, J. Sylak-Glassman, R. Que, and D. Yarowsky.

Very-large scale parsing and normalization of wiktionary morphological paradigms.

In *LREC*, 2016.



M. Liebeck and S. Conrad.

Iwnlp: Inverse wiktionary for natural language processing.

In *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)*, volume 2, pages 414–418, 2015.



MediaWiki.

Api:client code — mediawiki, the free wiki engine, 2018.

[Online; accessed 1-October-2018].



O. Roland.

Dictionary builder.

<https://github.com/newca12/dictionary-builder>, 2011.

[Online; accessed 1-October-2018].



Wikipedia contributors.

Wiktionary:parsing, 2017.

[Online; accessed 1-October-2018].



T. Zesch, C. Müller, and I. Gurevych.

Extracting lexical semantic knowledge from wikipedia and wiktionary.

In *LREC*, volume 8, pages 1646–1652, 2008.