Wikinflection: Massive semi-supervised generation of multilingual inflectional corpus from Wiktionary

Eleni Metheniti and Günter Neumann

DFKI
Stuhlsatzenhausweg 3
66123 Saarbrücken

`eleni.metheniti@dfki.de`, `neumann@dfki.de`

Abstract

Wiktionary is an open- and crowd-sourced dictionary which has been an important resource for natural language processing/understanding/generation tasks, but a big portion of the available information, such as inflection, is hard to retrieve and has not been widely utilized. In this paper, we are describing our efforts to generate inflectional paradigms for lemmata of the English Wiktionary, by using both the dynamic links of the XML dump file and the static information of the web version. Our system can generate inflectional paradigms for 225K lemmata, with almost 8,5M forms from 1.708 inflectional templates, for over 150 languages, and after evaluating the generation, 216K lemmata and around 6M forms are of high quality. In addition, we retrieve morphological features, affixes and stem allomorphs for each paradigm and form. The system can produce a structured inflectional corpus from any version of the English Wiktionary XML dump file, and could also be adapted for other language versions. The first version of the source code is currently available online.

KEYWORDS: wiktionary, metadata, inflection, corpus, computational morphology.

1 Introduction

1.1 Motivation

Wiktionary is a multilingual, open-sourced project, part of the Wikimedia foundation, which hosts multilingual dictionaries in many target languages. Every lemma in the Wiktionary is sectioned per source language, and contains pronunciation, etymology, definition, derivatives, translations, semantic and inflectional information (if the entry is complete and such information is available). The free and open access and sourcing of this project has established it as a vastly used resource for natural language processing tasks, especially with the use of the domain's XML dump files. These are the source files which are used to dynamically generate the content of a static HTML page per request from the browser; however, the XML file which generates this page only shows links to other lemma pages and utility pages. Providing the XML files allows for natural language processing experts and enthusiasts to quickly and offline extract lexicographic information for many tasks (semantic, phonological, etc.), however, the structured nature of the data can impede or even prevent the extraction of some information, such as inflectional tables.

In this paper, we are describing our attempt to create a multilingual inflectional corpus, with morphological information of affixes and stem allomorphs. We use both the XML dump file and information pulled from the web version of the English Wiktionary, in order to decode machine-readable information (in this case, the dynamic link to an inflectional template) into a human-readable structured file, divided per lemma and per language template. The goal is to generate the inflectional corpus with as little supervision as possible, in order to ensure reproducibility for other users, and extensibility, so that it would be possible in the future to generate dictionaries with updated information or from different editions of the Wiktionary.

1.2 Why inflection

Inflection is the set of morphological processes that occur in a word, so that the word acquires certain grammatical features which either create syntactic dependencies in a phrase (e.g. agreement between nouns and adjectives) or add to the meaning but not change it (e.g. tense in verbs). Inflectional languages have different choices as to which and how these grammatical features will be expressed, for example, most English nouns have four possible forms (singular number, singular number in possessive form, plural number, plural number in possessive form), while nouns in German have eight possible forms (in four cases and two numbers) which vary depending on the gender and the way the noun is declined. The different forms of a word in inflectional languages may be formed by *affixation* (e.g. plural in English nouns), by changes in the stem of the word which will produce a stem *allomorph* (for example, *reduplication*, e.g. plural in Samoan verbs by duplicating part of the stem), or both (e.g. plural in German nouns by *ablaut* and affixation).

$$\text{English: } house_{[+singular]} \rightarrow house - s_{[+plural]}$$
$$\text{Samoan: } savali_{[+singular]} \rightarrow sa - va - vali_{[+plural]}$$
$$\text{German: } Haus_{[+singular]} \rightarrow H\ddot{a}us\text{—}er_{[+plural]}$$

Inflection has been an ongoing challenge for natural language processing, because of the different levels of morphological richness of every language, the extensiveness of some inflectional paradigms, the low frequency of some forms, the ambiguity when forms are homonyms but have different grammatical properties, to name a few reasons. However, it could prove useful to tasks that require identification, lemmatization, semantic relations, text generation or use of low-frequency words, because inflection shows the intrinsic bond between forms of the same lemma, and can identify or provide the form with the correct morphosyntactic properties, in tasks such as machine translation, natural language generation and semantic analysis.

2 Previous Work

As a valuable resource for a multitude of natural language processing tasks, there are many available tools to parse the Wiktionary and extract relevant information; the Wikimedia foundation provides the MediaWiki Action API and client libraries in many programming languages, so that users can parse dump files and access information in machine-readable or human-readable ways (MediaWiki, 2018). Concerning the Wiktionary dump files, a discussion page in the domain states the difficulty of parsing a Wiktionary dump file for all its information, because of the presence of dynamic links (Wikipedia contributors, 2017). Most parsing tools, under the auspice of the MediaWiki project or independently developed, either splice the dump file in individual XML page files for easier access (Roland, 2011), or parse and extract specific information which is explicitly stated in the dump file(s), e.g. translations (Acs et al., 2013) or lexical-semantic information (Zesch et al., 2008).

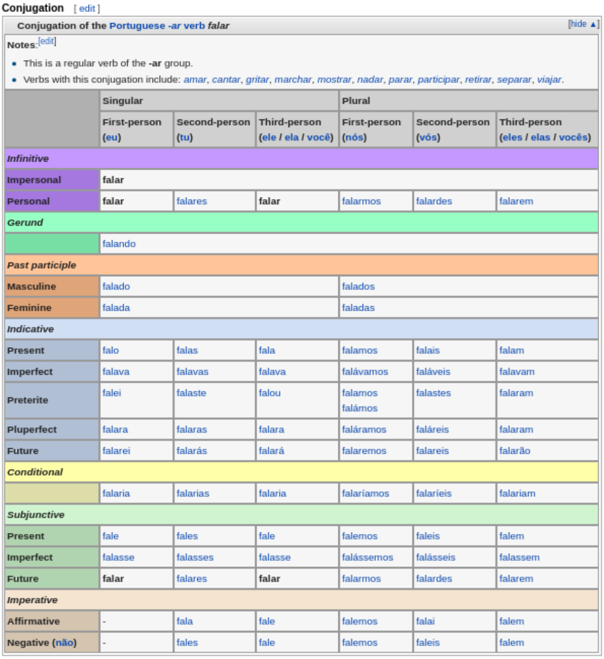
The potential of using Wiktionary as a source of inflectional information has not been untapped, however. Liebeck and Conrad (2015) have created `IWNLP`, a parser for the German edition of the Wiktionary which, with the Lemmatizer module, can produce a mapping from an inflected form to a lemma. First of all, they analyzed the inflectional templates used to dynamically generate inflections, and have re-implemented them in `C#` from the original `Lua`. Then, the tool uses the dynamic link in the page of a lemma, which points to an inflectional template, to generate the inflections (which are then used for the lemmatization task). The accuracy and quality of `IWNLP` is very high, however, so far they have only implemented templates for German nouns, adjectives and most frequently used templates for verbs, and are only using the German edition of Wiktionary.

Kirov et al. (2016) followed a radically different approach to gathering inflectional information from the Wiktionary; instead of using the XML dump file, they relied on the static HTML file and used the already generated tables, in order to pull the inflected forms of a lemma. They ensured that their parsing method would yield both the forms and the appropriate features, as noted in the table, and used this information to create a corpus of inflected words with annotated features, using their previously created *UniMorph* annotation schema. Their corpus is of high quality and includes almost 1M entries. However, the practice of pulling information from the web version of Wikimedia pages is highly discouraged, as it could put strain on the servers. In addition, their corpus lacks the information that is included in the dynamic links for a template (presented in detail in Section 3.1) and only includes the lemma and word features.

3 Methodology

3.1 Preliminary work

As mentioned in Section 1.1, a Wiktionary XML dump file contains all the pages available in the Wiktionary website for a specific target language, not in HTML format, but with dynamic links. Whenever there is a request to access a web page, the server runs a script for every component that is encoded, and decodes it into the relevant information. For example, as seen in Figure 1, the web page for the word *'falar'* ('to talk') contains a table with the verb conjugation for the word in Portuguese. This web page is generated by the XML page for *'falar'* (also included in the dump file in XML format, as seen in Figure 2), which contains links to other word pages (in multiple brackets) and utility pages (in multiple curly brackets). To generate the verb conjugation, this formula is used as a link: `{{pt-conj|fal|ar}}`. Its first parameter refers to the template which should be used, the second parameter is the stem of the word (`fal`) and the third parameter is conjugation information (the suffix `ar`).



Conjugation [edit]

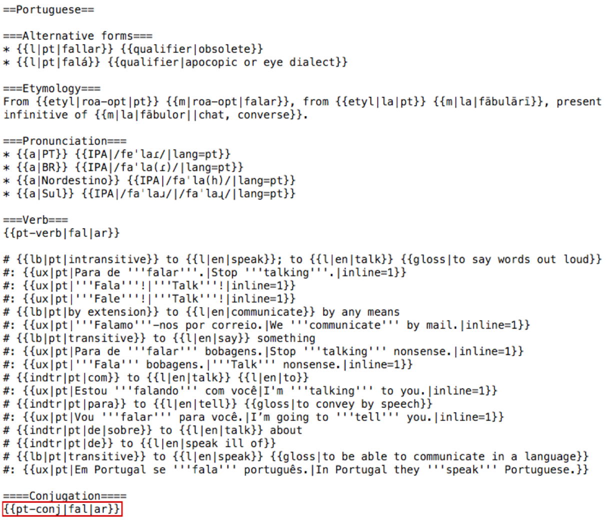
Conjugation of the Portuguese *-ar* verb *falar* [hide ▲]

Notes:[edit]

- This is a regular verb of the **-ar** group.
- Verbs with this conjugation include: *amar*, *cantar*, *gritar*, *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						
Impersonal	falar					
Personal	falar	falares	falar	falarmos	falardes	falarem
Gerund						
	falando					
Past participle						
Masculine	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 falámos	falastes	falaram
Pluperfect	falara	falaras	falara	faláramos	faláreis	falaram
Future	falarei	falarás	falará	falaremos	falareis	falarão
Conditional						
	falaria	falarias	falaria	falaríamos	falaríeis	falariam
Subjunctive						
Present	fale	fales	fale	falemos	faleis	falem
Imperfect	falasse	falasses	falasse	falássemos	falásseis	falassem
Future	falar	falares	falar	falarmos	falardes	falarem
Imperative						
Affirmative	-	fala	fale	falemos	falai	falem
Negative (não)	-	fales	fale	falemos	faleis	falem

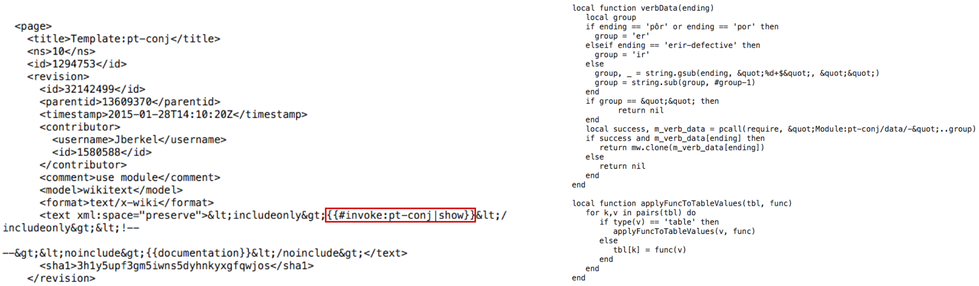
Figure 1: Web page (excerpt) of the lemma *'falar'* in the English Wiktionary.



```
==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|fābulārī}}, present
infinitive of {{m|la|fābulor||chat, converse}}.

===Pronunciation===
* {{a|PT}} {{IPA|/fɐˈlaɾ/|lang=pt}}
* {{a|BR}} {{IPA|/faˈla(ɾ)/|lang=pt}}
* {{a|Nordestino}} {{IPA|/faˈla(h)/|lang=pt}}
* {{a|Sul}} {{IPA|/faˈlaɹ/|/faˈlaɻ/|lang=pt}}

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

# {{lb|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}}
# {{lb|pt|by extension}} to {{l|en|communicate}} by any means
#: {{ux|pt|'''Falamo'''-nos por correio.|We '''communicate''' by mail.|inline=1}}
# {{lb|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}}
# {{indtr|pt|com}} to {{l|en|talk}} {{l|en|to}}
#: {{ux|pt|Estou '''falando''' com você|I'm '''talking''' to you.|inline=1}}
# {{indtr|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}}
# {{indtr|pt|de|sobre}} to {{l|en|talk}} about
# {{indtr|pt|de}} to {{l|en|speak ill of}}
# {{lb|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|fal|ar}}
```

Figure 2: XML page (excerpt) of the lemma *'falar'* in the English Wiktionary.

The formula links the lemma to a utility page, which contains the template `Template:pt-conj` to generate the inflectional paradigm of *'falar'*. However, upon inspecting the web page of the template[1] and the corresponding XML page (Figure 3a), it is observed that the template is also generated and not explicitly written; as declared with a link, it calls for a module page, `Module:pt-conj` which will generate the template, which in turn will generate the inflectional paradigm of the verb. The module is written in `Lua` (an excerpt can be seen in Figure 3b and the full script is online[2]) and uses the other parameters provided in the inflectional formula for the lemma, `fal` and `ar`.



```
  <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;{{#invoke:pt-conj|show}}&lt;/
includeonly&gt;&lt;!--

--&gt;&lt;noinclude&gt;{{documentation}}&lt;/noinclude&gt;</text>
      <sha1>3h1y5upf3gm5iwns5dyhnkyxgfqwjos</sha1>
    </revision>
```

(a) XML page for the template.

```
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, &quot;%d+$&quot;, &quot;&quot;)
     group = string.sub(group, #group-1)
   end
   if group == &quot;&quot; then
        return nil
   end
   local success, m_verb_data = pcall(require, &quot;Module:pt-conj/data/-&quot;..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
```

(b) Module script in `Lua` (excerpt).

Figure 3: The template and module to generate `pt-conj`.

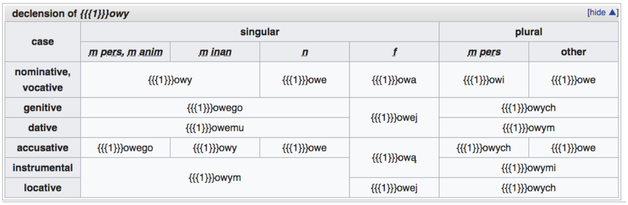
Our initial attempt at generating inflections was to replicate the process in which a module generates an inflectional paradigm in a web page, but this approach proved to be unsuccessful. The steps followed were: (a) reading the XML dump file and extracting the lemma pages, the template pages and the module pages, (b) saving the module pages in their individual `Lua` script file, (c) finding the inflectional formula(s) in each lemma (i.e. the template(s) and the required parameters), (d) finding the template(s) linked to the lemma, (e) finding the module linked to the template(s), and (f) run the module with the parameters of the inflectional formula(s). Unfortunately, this approach proved to be unsustainable, because the `Lua` scripts need more data than the given; they require the template pages on how to generate an HTML table, some of the scripts require multiple inflectional templates or data dictionaries that are not available in the XML file etc. The source code of a static HTML page mentions in a comment all the templates and modules that were used to generate the dynamic content, however, this information is not in all cases explicitly stated in the XML files. Therefore, it was not possible to produce significant results without querying every lemma page from the web edition for its comments, and it would be impossible to edit every script without supervision.

Our second attempt focused, at first, on templates; they are available online in static HTML pages, and they are generated by the respective modules. As seen in Figure 4, the generated web page for a template is quite comprehensible: the table includes all the forms of the inflectional paradigm with their respective labels, in the same way that they appear in the lemma web page (Figure 5) – the only difference is that, in place of stem, it includes the link `{{{1}}}`. From our previous attempt, we understood that a dynamic link to a template, for example, `{{pl-decl-adj-owy|róż}}` for the lemma *'różowy'* ('pink'), includes as first parameter the template name, as second parameter the stem of the paradigm, etc., therefore it was easy to

[1] `https://en.wiktionary.org/wiki/Template:pt-conj`

[2] `https://en.wiktionary.org/wiki/Module:pt-conj`

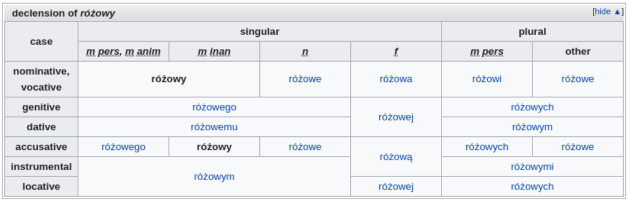
understand that `{{{1}}}` refers to the second parameter of the link[3]; when the module is called, the second parameter will replace the placeholder `{{{1}}}` and form the declension.



declension of *{{{1}}}owy* [hide ▲]

case	singular				plural	
	m pers, m anim	m inan	n	f	m pers	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	{{{1}}}owych	

Figure 4: Table from template `pl-decl-adj-owy`.



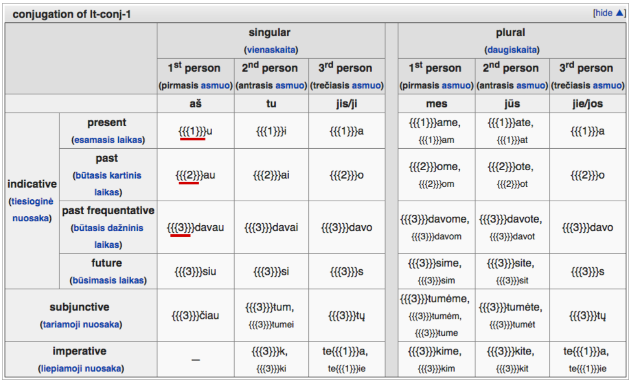
declension of *różowy* [hide ▲]

case	singular				plural	
	m pers, m anim	m inan	n	f	m pers	other
nominative, vocative	różowy		różowe	różowa	różowi	różowe
genitive	różowego			różowej	różowych	
dative	różowemu				różowym	
accusative	różowego	różowy	różowe	różową	różowych	różowe
instrumental	różowym				różowymi	
locative				różowej	różowych	

Figure 5: Table from lemma *różowy*.

The same process applies for templates which require more than two parameters, for example the Lithuanian verb *'gauti'* ('to get') has the dynamic link `{{lt-conj-1|gaun|gav|gau}}` because the template requires four parameters to create the conjugation table (Figures 6, 7) – the last three exist because this inflectional paradigm requires stem allomorphs. This proves to be very interesting, because the template link provides information that sometimes is not mentioned in the entry of a lemma, i.e. the type of inflectional paradigm that the word follows, and the changes that happen to the form beyond affixation.

For our approach, we decided that template information, both dynamic and static, was going to be an integral part of our corpus. In the following section, we will explain how we explored and used it.

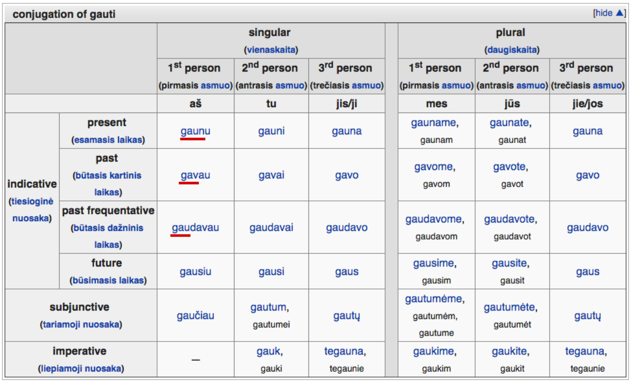


conjugation of lt-conj-1 [hide ▲]

		singular (vienaskaita)			plural (daugiskaita)		
		1st person (pirmasis asmuo)	2nd person (antrasis asmuo)	3rd person (trečiasis asmuo)	1st person (pirmasis asmuo)	2nd person (antrasis asmuo)	3rd person (trečiasis asmuo)
		aš	tu	jis/ji	mes	jūs	jie/jos
indicative (tiesioginė nuosaka)	present (esamasis laikas)	{{{1}}}u	{{{1}}}i	{{{1}}}a	{{{1}}}ame, {{{1}}}am	{{{1}}}ate, {{{1}}}at	{{{1}}}a
	past (būtasis kartinis laikas)	{{{2}}}au	{{{2}}}ai	{{{2}}}o	{{{2}}}ome, {{{2}}}om	{{{2}}}ote, {{{2}}}ot	{{{2}}}o
	past frequentative (būtasis dažninis laikas)	{{{3}}}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

Figure 6: Table from template `lt-conj-1`.

[3] Traditionally, programming languages start indexing elements at zero, therefore the first parameter has position [0], the second [1], and so on.



conjugation of gauti [hide ▲]

		singular (vienaskaita)			plural (daugiskaita)		
		1st person (pirmasis asmuo)	2nd person (antrasis asmuo)	3rd person (trečiasis asmuo)	1st person (pirmasis asmuo)	2nd person (antrasis asmuo)	3rd person (trečiasis asmuo)
		aš	tu	jis/ji	mes	jūs	jie/jos
indicative (tiesioginė nuosaka)	present (esamasis laikas)	gaunu	gauni	gauna	gauname, gaunam	gaunate, gaunat	gauna
	past (būtasis kartinis laikas)	gavau	gavai	gavo	gavome, gavom	gavote, gavot	gavo
	past frequentative (būtasis dažninis laikas)	gaudavau	gaudavai	gaudavo	gaudavome, gaudavom	gaudavote, gaudavot	gaudavo
	future (būsimasis laikas)	gausiu	gausi	gaus	gausime, gausim	gausite, gausit	gaus
subjunctive (tariamoji nuosaka)		gaučiau	gautum, gautumei	gautų	gautumėme, gautumėm, gautume	gautumėte, gautumėt	gautų
imperative (liepiamoji nuosaka)		–	gauk, gauki	tegauna, tegaunie	gaukime, gaukim	gaukite, gaukit	tegauna, tegaunie

Figure 7: Table from lemma *gauti*.

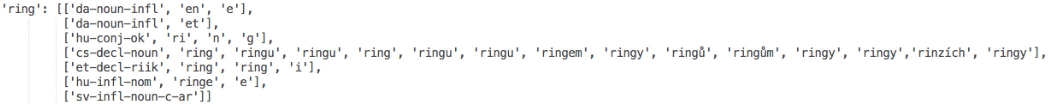
3.2 Creating inflectional templates and paradigms

In the process of extracting inflectional templates to later use them on lemmata, it is already explained why we could not easily extract them from the XML dump file; thus our other option was to extract them from the respective web pages. While querying the Wiktionary website is not the recommended practice, the number of requests we would have to perform would be significantly lower than the ones performed by Kirov et al., because we would look for templates and not all web pages.

The first step started with the XML dump file; as mentioned, it contains all the pages of the web domain, so we would be able to extract the pages of lemmata and the pages of templates; the goal was to create a dictionary of all the word entries in the Wiktionary which (a) are lemmata and not pages of a form and (b) contain at least one dynamic link to an inflectional template. For example, the words *'falar'*, *'różowy'* and *'gauti'* are added to the dictionary of entries, because they fulfill the needed criteria, but the words *'houses'*[4] or *'architecture'*[5] are not (the former is not a lemma, and the latter does not have any links to inflectional information on Wiktionary). Multiple template links for a single lemma means that the word exists as a lemma in multiple languages, or that it adheres to many inflectional paradigms (e.g. in Figure 8, the lemma *'ring'* in Danish may follow two different noun inflectional paradigms.). In this step, we also create a list of all template names which are used for inflectional paradigms. The template names extracted also had to fulfill some criteria; they needed to contain the words 'noun', 'verb', etc. or the words 'decl', ' conj', etc. in order to not be confused with templates for other linguistic information (e.g. phonology) or utility templates (e.g. 'table' templates which generate the format of an HTML table). In this step, we extracted 454.470 pages of lemmata with inflectional template links (out of the 5.740.594 words in the latest edition of the English Wiktionary dump file) and 7.068 inflectional templates.

We then had to perform a request for the web page of every template; in order to perform these requests only once, we opted to download the HTML files and use the local files to generate the templates. In Python3, we used the BeautifulSoup library to parse the HTML code and find an HTML inflectional table, and the pandas library to convert the table to a data frame. We had to overcome some formatting issues, for example, merged cells which contained labels

[4] https://en.wiktionary.org/wiki/houses
[5] https://en.wiktionary.org/wiki/architecture



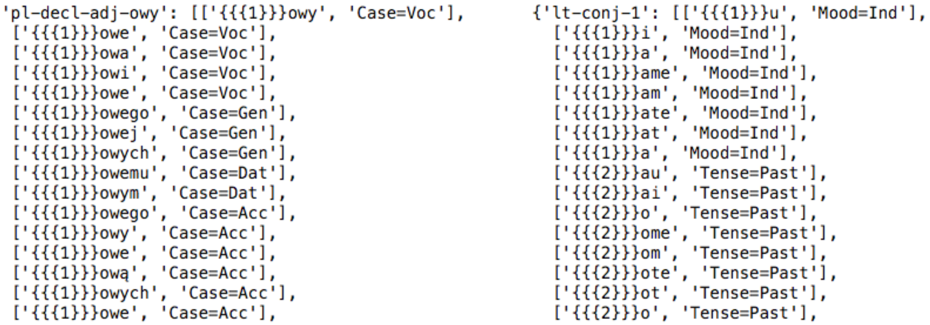
```
'ring': [['da-noun-infl', 'en', 'e'],
         ['da-noun-infl', 'et'],
         ['hu-conj-ok', 'ri', 'n', 'g'],
         ['cs-decl-noun', 'ring', 'ringu', 'ringu', 'ring', 'ringu', 'ringu', 'ringem', 'ringy', 'ringů', 'ringům', 'ringy', 'ringy','rinzích', 'ringy'],
         ['et-decl-riik', 'ring', 'ring', 'i'],
         ['hu-infl-nom', 'ringe', 'e'],
         ['sv-infl-noun-c-ar']]
```

Figure 8: The entry *'ring'* in the word dictionary. The first digits of every template, before the first dash, refer to the language's ISO 639-5 code.

for multiple rows or columns had to be unmerged and duplicated (Ricco, 2017), and some cells contained multiple forms, either phonetic transcriptions in the Latin alphabet or different possible forms – we decided that as much of the available information should be preserved, so duplicate entries had to be made in the inflectional template.

In addition, we decided to convert the extracted morphological features from text to Universal Dependencies morphological feature tags (Nivre et al., 2018); this was done for the sake of uniformity, because authors of different templates had made different choices in the way features were written (e.g. 'singular' vs. 'sing.'). Universal Dependencies were specifically chosen, because they have been used across many NLP applications, from treebanks and annotated data to syntactic parsers, morphological taggers etc. and are extensively documented, therefore our corpus could prove useful for a variety of applications. In order to convert the arbitrary Wiktionary tags to UD tags, we built a database of all the available Universal Dependencies tags, and as many Wiktionary tags and their variations as we could possibly gather. This database however is not exhaustive, as in some cases tags are written abbreviated (e.g. 'masculine' can appear as 'm', 'm.', 'male', 'masc.' etc) or are not in English (e.g. Figure 7).

The greatest challenge, however, came with templates such as `pt-conj` (see Section 3.1); such templates' web pages have no inflectional tables and rely solely on the respective module to generate the inflections. Since we have made the decision not to use any module information, these templates unfortunately could not be parsed, and would not be included in our corpus. Out of the 7.068 saved HTML pages, 2.927 templates had inflectional information in table format that could be parsed. The output of our template reading script produces a dictionary of templates, where every template includes a list of all possible inflected forms, each with their morphological features. Examples of this dictionary's entries can be seen in Figures 9a, 9b.



```
'pl-decl-adj-owy': [['{{{1}}}owy', 'Case=Voc'],
 ['{{{1}}}owe', 'Case=Voc'],
 ['{{{1}}}owa', 'Case=Voc'],
 ['{{{1}}}owi', 'Case=Voc'],
 ['{{{1}}}owe', 'Case=Voc'],
 ['{{{1}}}owego', 'Case=Gen'],
 ['{{{1}}}owej', 'Case=Gen'],
 ['{{{1}}}owych', 'Case=Gen'],
 ['{{{1}}}owemu', 'Case=Dat'],
 ['{{{1}}}owym', 'Case=Dat'],
 ['{{{1}}}owego', 'Case=Acc'],
 ['{{{1}}}owy', 'Case=Acc'],
 ['{{{1}}}owe', 'Case=Acc'],
 ['{{{1}}}ową', 'Case=Acc'],
 ['{{{1}}}owych', 'Case=Acc'],
 ['{{{1}}}owe', 'Case=Acc'],
```

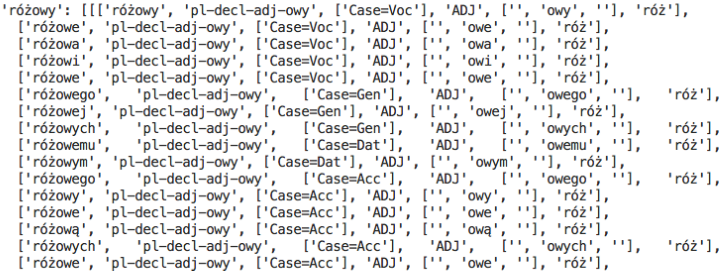
(a) Parsed template: `pl-decl-adj-owy`.

```
{'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'],
```

(b) Parsed template: `lt-conj-1`.

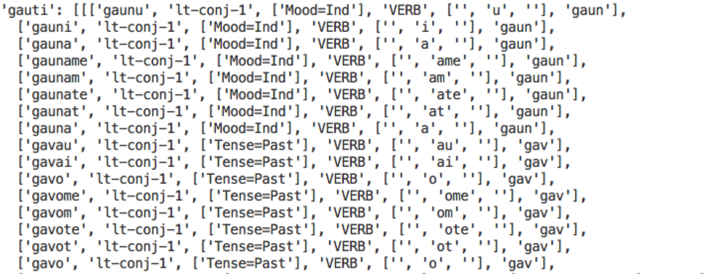
Figure 9: Entries from the template dictionary.

After the dictionary of templates is made, we used the dictionary of word entries to iterate over every word, and for every template link that the word had, we generated an inflectional paradigm, as seen in Figures 10, 11. Each form has information for its corresponding inflectional template, morphological data, the part-of-speech tag (depending on the template), the stem used by the template to generate this form, and a list of prefixes, suffixes and infixes if available. The resulting dictionary has 225453 lemmata, which have been matched with 1708 templates in order to generate 8426480 forms, in a total of 199 languages.



```
'różowy': [[['różowy', 'pl-decl-adj-owy', ['Case=Voc'], 'ADJ', ['', 'owy', ''], 'róż'],
  ['różowe', 'pl-decl-adj-owy', ['Case=Voc'], 'ADJ', ['', 'owe', ''], 'róż'],
  ['różowa', 'pl-decl-adj-owy', ['Case=Voc'], 'ADJ', ['', 'owa', ''], 'róż'],
  ['różowi', 'pl-decl-adj-owy', ['Case=Voc'], 'ADJ', ['', 'owi', ''], 'róż'],
  ['różowe', 'pl-decl-adj-owy', ['Case=Voc'], 'ADJ', ['', 'owe', ''], 'róż'],
  ['różowego',   'pl-decl-adj-owy',   ['Case=Gen'],   'ADJ',   ['', 'owego', ''],   'róż'],
  ['różowej', 'pl-decl-adj-owy', ['Case=Gen'], 'ADJ', ['', 'owej', ''], 'róż'],
  ['różowych',   'pl-decl-adj-owy',   ['Case=Gen'],   'ADJ',   ['', 'owych', ''],   'róż'],
  ['różowemu',   'pl-decl-adj-owy',   ['Case=Dat'],   'ADJ',   ['', 'owemu', ''],   'róż'],
  ['różowym', 'pl-decl-adj-owy', ['Case=Dat'], 'ADJ', ['', 'owym', ''], 'róż'],
  ['różowego',   'pl-decl-adj-owy',   ['Case=Acc'],   'ADJ',   ['', 'owego', ''],   'róż'],
  ['różowy', 'pl-decl-adj-owy', ['Case=Acc'], 'ADJ', ['', 'owy', ''], 'róż'],
  ['różowe', 'pl-decl-adj-owy', ['Case=Acc'], 'ADJ', ['', 'owe', ''], 'róż'],
  ['różową', 'pl-decl-adj-owy', ['Case=Acc'], 'ADJ', ['', 'ową', ''], 'róż'],
  ['różowych',   'pl-decl-adj-owy',   ['Case=Acc'],   'ADJ',   ['', 'owych', ''],   'róż'],
  ['różowe', 'pl-decl-adj-owy', ['Case=Acc'], 'ADJ', ['', 'owe', ''], 'róż'],
```

Figure 10: Paradigm (excerpt) for *różowy*.

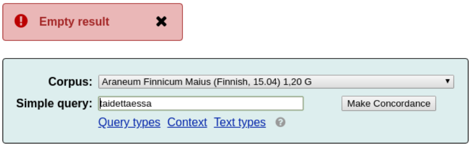


```
'gauti': [[['gaunu', 'lt-conj-1', ['Mood=Ind'], 'VERB', ['', 'u', ''], 'gaun'],
  ['gauni', 'lt-conj-1', ['Mood=Ind'], 'VERB', ['', 'i', ''], 'gaun'],
  ['gauna', 'lt-conj-1', ['Mood=Ind'], 'VERB', ['', 'a', ''], 'gaun'],
  ['gauname', 'lt-conj-1', ['Mood=Ind'], 'VERB', ['', 'ame', ''], 'gaun'],
  ['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'],
  ['gavai', 'lt-conj-1', ['Tense=Past'], 'VERB', ['', 'ai', ''], 'gav'],
  ['gavo', 'lt-conj-1', ['Tense=Past'], 'VERB', ['', 'o', ''], 'gav'],
  ['gavome', 'lt-conj-1', ['Tense=Past'], 'VERB', ['', 'ome', ''], 'gav'],
  ['gavom', 'lt-conj-1', ['Tense=Past'], 'VERB', ['', 'om', ''], 'gav'],
  ['gavote', 'lt-conj-1', ['Tense=Past'], 'VERB', ['', 'ote', ''], 'gav'],
  ['gavot', 'lt-conj-1', ['Tense=Past'], 'VERB', ['', 'ot', ''], 'gav'],
  ['gavo', 'lt-conj-1', ['Tense=Past'], 'VERB', ['', 'o', ''], 'gav'],
```

Figure 11: Paradigm (excerpt) for *gauti*.

3.3 Evaluating and correcting the paradigms

The generation process was successful, however, it is necessary that we check the quality of the produced paradigms. The first option would be to examine if these words exist in a corpus, and therefore are grammatical formations. However, it would be impossible to find corpora for all 199 languages, and corpora large enough to include forms that are grammatical but might be of very low frequency. We conducted a small-scale experiment to prove that the use of corpora would be neither feasible nor fruitful; we used the inflectional paradigm of the Finnish verb '*taitaa*' and queried some of its forms in *Araneum Finnicum Maius* (Benko, 2016), a Finnish corpus of over 1,2B tokens. forms such as '*taidettaessa*' (second passive infinitive in inessive case) were not found (Figure 12).



Corpus: Araneum Finnicum Maius (Finnish, 15.04) 1,20 G

Simple query: taidettaessa

Figure 12: Querying the word '*taidettaessa*' returned no results from *Finnicum Maius*.

Our second option to evaluate our generated inflections would be to look inside the source, the Wiktionary, and check if the generated forms exist in the page of a lemma. This approach would ensure that our generating process yields the same results as the modules generating templates and lemmata. However, for reasons already stated, we decided to not check every single inflectional paradigm; instead, we created a script which randomly selects one lemma from each template (i.e. 1.708 unique lemmata), finds the web page of the lemma and looks up the presence of all the inflected forms of the lemma. In Table 1 the results of two random evaluations are presented. [6]

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

Table 1: Evaluation results. 'All' refers to the number of forms in the paradigm.

After a few evaluation runs, we first noticed that some templates and template links behaved irregularly compared to others; for example, the lemma *'tocar'* contains the link to the Spanish conjugation template as `{{es-conj-ar|to}}`, but the second parameter does not provide a correct stem to complete the template. Also, as previously seen in Figure 8, the authors of the Danish links did not include stem information in the links, because their templates use the lemma for inflections, but the authors of the Estonian and Hungarian links include the stem and stem allomorphs, and the Czech authors have opted to not generate any inflections with a template, but to input all forms as word allomorphs. We decided to perform a quick revision to templates per language, and create a few exceptions for languages with different dynamic link formatting, because Wiktionary authors of the same language tend to adhere to the same formatting rules. It would be impossible to manually check all the templates, and it would also not be reproducible in case of new or updated templates in Wiktionary.

We also noticed that our evaluation might produce false negatives, because of decoding problems when parsing an HTML page (a problem which appeared in languages such as Serbo-Croatian), or because of irregularities in the inflection of a random word which are caused by extraneous factors (for example, as seen in Table 1, the word *Herostratus* for `la-decl-2nd` produced false negatives because the word is a common noun and does not have a plural form), or because in lemmata with too many generated information the server sometimes fails to execute all Lua scripts and generate all content.

Using the Wiktionary as means of evaluation is not optimal, but it is the best available source so far for looking up full inflectional paradigms and words that are grammatical but may not exist in a language corpus. After running the evaluation, our script stores in memory the indices of the false forms per paradigm, and then updates the templates by removing these false forms (or the entire template if it is incorrect) and re-generates the paradigms, which some of them now may be partial but should have good quality forms. The number of generated paradigms and forms may vary per generation, but for random evaluation No. 1, the total number of paradigms was 216.378, generated by 1.537 templates, and yielded 5.970.799 forms, and for random evaluation No. 3, the total number of paradigms was 210.172, generated by 1.521 templates, and yielded 6.024.077 forms. A table for random evaluation No. 3 can be found in Table 2.[6]

[6] Full tables are available online at `https://tinyurl.com/wikinflection`

			Before evaluation and correction		After random eval. No.3 and correction	
Language	***Template***	***Lemmas per temp.***	***Inflections per lemma***	***No. forms***	***Inflections per lemma***	***No. forms***
Finnish	`fi-decl-valo-koira`	4	58	232	58	232
Romanian	`ro-noun-f-ea`	23	10	230	10	230
Assamese	`as-proper noun`	58	7	406	7	406
Lower Sorbian	`dsb-decl-noun-17`	62	18	1116	18	1116
Gujarati	`*gu-conj-v`	1	7	7	-	0
Finnish	`*fi-decl-kala-koira`	2	63	126	-	0

Table 2: Randomly selected templates from random evaluation 3. Note that the last two templates have been removed after the evaluation process, and are noted with an asterisk.

Comparing our system's results to the most recent version of Kirov et al. *UniMorph* project corpus, it is noted that according to the available resources online, the *UniMorph* project currently has a corpus of 8.8M words, compared to our 6 to 6.5M words.[7] In addition, the *UniMorph* corpus has significantly more forms for high frequency languages, however, a lot of languages mentioned in the 'annotated languages' section do not have an available corpus (languages which are not available yet are listed in a different section in the webpage). As Table 3 demonstrates, *UniMorph* has many more forms for Arabic, but is lacking when it comes to low frequency languages such as the dialects of Alemannic German. Our system's lacking may occur either because the language experts for certain languages have created modules or badly-formatted templates to generate inflectional paradigms (and our system is not capable of processing either of those), but *UniMorph* has access to this information from accessing directly the lemmata's webpages.

Language		**UniMorph**		**Wikinflection**	
Name	***ISO***	***Forms***	***Paradigms***	***Forms before evaluation***	***Forms after evaluation***
Adyghe	ady	n/a	n/a	440	440
Albanian	sqi	33483	589	8767	8767
Alemannic German	gsw	0	0	232	232
Ancient Greek	grc	0	0	3312	3312
Arabic	ara	140003	4134	36	36
Aragonese	an	0	0	448	448
Armenian	hye	338461	7033	2824	59
Assamese	as	0	0	13790	13790
Asturian	ast	n/a	n/a	23599	23329
Avestan	ae	0	0	6	6
SUM		**8850395**	**309083**	**6518762**	**6024077**

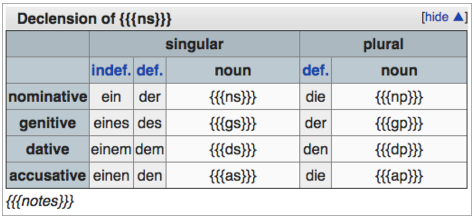
Table 3: A list of the fist 10 languages (in alphabetical order) and the number of their forms, from the *UniMorph* project and our system, *Wikinflection*. The full table is available online.[6] 'n/a' refers to languages that (according to the *UniMorph* project website) have been annotated, but no number has been published. '0' refers to the absence of the language from the corpus. For *Wikinflection* statistics, Random Evaluation No. 3 was used.

4 Discussion

Our approach to generate all inflectional paradigms from Wiktionary, on a large and multilingual scale, proved successful, but not as high-quality as we initially expected. First of all, we lost access to a big portion of inflectional information, because we only opted to use static template information and not any modules. Second, one of our greatest challenges was the different

[7] `https://unimorph.github.io/`, accessed November 21, 2018.

formatting of different languages; the dynamic links were different among languages as discussed in Section 3.3, and also the inflectional tables looked vastly different, with very inconsistent information flow, use of header rows, header columns and feature naming. This also caused loss of some morphological features when parsing inflectional tables, a problem which we are aware of and are in the process of solving. The inconsistencies span further in some cases, with templates which do not follow the Wiktionary-wide format of dynamic links (Figure 13a), or templates which are incomplete or contain links to other content, sometimes nonexistent (Figure 13b), and these templates were automatically rejected during parsing.

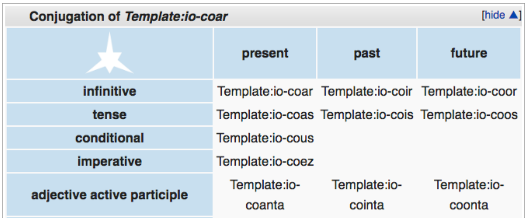


Declension of {{{ns}}} [hide ▲]

	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}}}

{{{notes}}}

(a) Table from template `de-decl-noun-m`.



Conjugation of *Template:io-coar* [hide ▲]

★	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-cous		
imperative	Template:io-coez		
adjective active participle	Template:io-coanta	Template:io-cointa	Template:io-coonta

(b) Table from template `io-conj`.

Figure 13: Examples of templates which cannot be parsed by our system.

Another reason why we were only able to retrieve inflectional information from half of the lemmata with inflectional links was the ever-changing and evolving nature of the Wiktionary. Many entries contain inflectional links in the format of `{{rfinfl|LANGUAGE|POS}}`, which are actually placeholders for templates that do not exist yet[8]. Concerning the generated paradigms, because words of the same inflectional schema tend to follow similar morphological processes, we are confident to believe that the quality of the generated forms is at least satisfactory. There are cases of false negatives as discussed in Section 3.3, but these are to be expected from semi-supervised generation. False positives are rare, but may occur in cases where the inflectional paradigm requires more than one template or additional information from modules in order to be generated; for example, the template `{{hu-infl-nom}}`[9] used for nouns such as *'ring'* calls for extra parameters and server data in order to produce stem allomorphs during declension. While this approach is effective for generating the web tables, it impedes our generation of a correct inflectional paradigm, and also is inconsistent with the way the verb inflectional tables were made for the same language (e.g. `{{hu-conj-szem-üd}}`[10] requires a third parameter to produce stem allomorphs, an approach which could have been used for nouns too).

Despite the issues we had to overcome, our system is able to generate an inflectional corpus, with information which would be hard to extract even with state-of-the-art tools, such as stem allomorphs and affixes. This information is usually sparsely available, especially for low-resource languages such as Crimean Tatar, Võro and Northern Sami, and could prove to be a useful source not only for natural language processing, but also for linguistic research and language learners. In addition, with the improvement of our morphological feature tagging, we aim to create a large resource of tagged tokens and types, which could improve performance on many natural language processing tasks, especially those who require the use of low-frequency words.

[8] `https://en.wiktionary.org/wiki/Template:rfinfl`

[9] `https://en.wiktionary.org/wiki/Template:hu-infl-nom`

[10] `https://en.wiktionary.org/wiki/Template:hu-conj-szem-üd`

5 Conclusion

Wiktionary has become an essential linguistic resource, and it is important to ensure that all its available information is accessible for research. Our attempts to parse the Wiktionary for inflectional information have allowed us to utilize data which has either been partially available (Kirov et al. (2016), Liebeck and Conrad (2015)) or has not been available so far (stem allomorphs). Although we were only able to access a fraction of the available inflectional information, we were able to construct paradigms for over 140 languages, some of which being low-frequency languages and previously did not have available inflectional corpora. Our project is available online on Github[11] and can be downloaded and used alongside an English Wiktionary XML dump file, to produce a local corpus of inflectional paradigms. While we had to tackle several difficulties and we are currently in the process of perfecting the output, our system is a new approach to parsing and providing multilingual linguistic resources for computational morphology.

Our future work will focus, primarily, on improving template parsing so that all possible morphological features are extracted, and on performing human evaluation on the produced output in order to ensure high quality. We aim to keep increasing the size and quality of the generated corpora, by exploring whether the use of modules could be possible in some cases, and we would like to soon release the corpus as a pre-made resource as well, in order to be directly used. Additionally, we will explore how easy it would be to adapt our code to generate inflectional paradigms from other editions of Wiktionary – if the other editions maintain the same data and link structure as the English dump file, it could be as simple as translating a few headings and features in the code.

6 Acknowledgements

This work was partially funded by the European Union's Horizon 2020 grant agreement No. 731724 (iREAD).

[11] `https://github.com/lenakmeth/Wikinflection`

References

Acs, J., Pajkossy, K., and Kornai, A. (2013). Building basic vocabulary across 40 languages. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 52–58, Sofia, Bulgaria. Association for Computational Linguistics.

Benko, V. (2016). Two years of aranea: Increasing counts and tuning the pipeline. In *LREC*.

Kirov, C., Sylak-Glassman, J., Que, R., and Yarowsky, D. (2016). Very-large scale parsing and normalization of wiktionary morphological paradigms. In *LREC*.

Liebeck, M. and Conrad, S. (2015). 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.

MediaWiki (2018). Api:client code — mediawiki, the free wiki engine. [Online; accessed 1-October-2018].

Nivre, J., Abrams, M., Agić, Ž., Ahrenberg, L., Antonsen, L., Aplonova, K., Aranzabe, M. J., Arutie, G., Asahara, M., Ateyah, L., Attia, M., Atutxa, A., Augustinus, L., Badmaeva, E., Ballesteros, M., Banerjee, E., Bank, S., Barbu Mititelu, V., Basmov, V., Bauer, J., Bellato, S., Bengoetxea, K., Berzak, Y., Bhat, I. A., Bhat, R. A., Biagetti, E., Bick, E., Blokland, R., Bobicev, V., Börstell, C., Bosco, C., Bouma, G., Bowman, S., Boyd, A., Burchardt, A., Candito, M., Caron, B., Caron, G., Cebiroğlu Eryiğit, G., Cecchini, F. M., Celano, G. G. A., Čéplö, S., Cetin, S., Chalub, F., Choi, J., Cho, Y., Chun, J., Cinková, S., Collomb, A., Çöltekin, Ç., Connor, M., Courtin, M., Davidson, E., de Marneffe, M.-C., de Paiva, V., Diaz de Ilarraza, A., Dickerson, C., Dirix, P., Dobrovoljc, K., Dozat, T., Droganova, K., Dwivedi, P., Eli, M., Elkahky, A., Ephrem, B., Erjavec, T., Etienne, A., Farkas, R., Fernandez Alcalde, H., Foster, J., Freitas, C., Gajdošová, K., Galbraith, D., Garcia, M., Gärdenfors, M., Garza, S., Gerdes, K., Ginter, F., Goenaga, I., Gojenola, K., Gökırmak, M., Goldberg, Y., Gómez Guinovart, X., Gonzáles Saavedra, B., Grioni, M., Grūzītis, N., Guillaume, B., Guillot-Barbance, C., Habash, N., Hajič, J., Hajič jr., J., Hà Mỹ, L., Han, N.-R., Harris, K., Haug, D., Hladká, B., Hlaváčová, J., Hociung, F., Hohle, P., Hwang, J., Ion, R., Irimia, E., Ishola, Ọ., Jelínek, T., Johannsen, A., Jørgensen, F., Kaşıkara, H., Kahane, S., Kanayama, H., Kanerva, J., Katz, B., Kayadelen, T., Kenney, J., Kettnerová, V., Kirchner, J., Kopacewicz, K., Kotsyba, N., Krek, S., Kwak, S., Laippala, V., Lambertino, L., Lam, L., Lando, T., Larasati, S. D., Lavrentiev, A., Lee, J., Lê Hồng, P., Lenci, A., Lertpradit, S., Leung, H., Li, C. Y., Li, J., Li, K., Lim, K., Ljubešić, N., Loginova, O., Lyashevskaya, O., Lynn, T., Macketanz, V., Makazhanov, A., Mandl, M., Manning, C., Manurung, R., Mărănduc, C., Mareček, D., Marheinecke, K., Martínez Alonso, H., Martins, A., Mašek, J., Matsumoto, Y., McDonald, R., Mendonça, G., Miekka, N., Misirpashayeva, M., Missilä, A., Mititelu, C., Miyao, Y., Montemagni, S., More, A., Moreno Romero, L., Mori, K. S., Mori, S., Mortensen, B., Moskalevskyi, B., Muischnek, K., Murawaki, Y., Müürisep, K., Nainwani, P., Navarro Horñiacek, J. I., Nedoluzhko, A., Nešpore-Bērzkalne, G., Nguyễn Thị, L., Nguyễn Thị Minh, H., Nikolaev, V., Nitisaroj, R., Nurmi, H., Ojala, S., Olúòkun, A., Omura, M., Osenova, P., Östling, R., Øvrelid, L., Partanen, N., Pascual, E., Passarotti, M., Patejuk, A., Paulino-Passos, G., Peng, S., Perez, C.-A., Perrier, G., Petrov, S., Piitulainen, J., Pitler, E., Plank, B., Poibeau, T., Popel, M., Pretkalniņa, L., Prévost, S., Prokopidis, P., Przepiórkowski, A., Puolakainen, T., Pyysalo, S., Rääbis, A., Rademaker, A., Ramasamy, L., Rama, T., Ramisch, C., Ravishankar, V., Real, L., Reddy, S., Rehm, G., Rießler, M., Rinaldi, L., Rituma, L., Rocha, L., Romanenko, M., Rosa, R., Rovati, D., Roșca, V., Rudina, O., Rueter, J., Sadde, S., Sagot, B., Saleh, S., Samardžić, T., Samson, S., Sanguinetti, M., Saulīte, B., Sawanakunanon, Y., Schneider, N., Schuster, S., Seddah, D., Seeker, W., Seraji, M., Shen, M., Shimada, A., Shohibussirri, M., Sichinava, D., Silveira, N., Simi, M., Simionescu, R., Simkó, K., Šimková, M., Simov, K., Smith, A., Soares-Bastos, I., Spadine, C., Stella, A., Straka, M., Strnadová, J., Suhr, A., Sulubacak, U., Szántó, Z., Taji, D., Takahashi, Y., Tanaka, T., Tellier, I., Trosterud, T., Trukhina, A., Tsarfaty, R., Tyers, F., Uematsu, S., Urešová, Z., Uria, L., Uszkoreit, H., Vajjala, S., van Niekerk, D., van Noord, G., Varga, V., Villemonte de la Clergerie, E., Vincze, V., Wallin, L., Wang, J. X., Washington, J. N., Williams, S., Wirén, M., Woldemariam, T., Wong, T.-s., Yan, C., Yavrumyan, M. M., Yu, Z., Žabokrtský, Z., Zeldes, A., Zeman, D., Zhang, M., and Zhu, H. (2018). Universal dependencies 2.3. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Ricco, J. (2017). Using python to scrape html tables with merged cells. [Online; accessed 7-October-2018].

Roland, O. (2011). Dictionary builder. https://github.com/newca12/dictionary-builder. [Online; accessed 1-October-2018].

Wikipedia contributors (2017). Wiktionary:parsing. [Online; accessed 1-October-2018].

Zesch, T., Müller, C., and Gurevych, I. (2008). Extracting lexical semantic knowledge from wikipedia and wiktionary. In *LREC*, volume 8, pages 1646–1652.