Using Lexemes in Abstract Wikipedia: How can we improve the data?

Ariel Gutman (Google.org fellow for Abstract Wikipedia)
Wikidata Quality Days, July 10th, 2022
Abstract Wikipedia's ultimate goals

- Represent articles in a language-agnostic way ("abstract content")
- Render them in the different language editions of Wikipedia using Natural Language Generation techniques.
Abstract Wikipedia’s architecture

Source: Multilingual Wikipedia architecture on Commons, created by Denny
Lexicographical data is key!
Example generation

Generic claim: use plural

"Cacti use CAM photosynthesis."

Which plural?

Forms

Cactaceae (Q14560)

uses

CAM photosynthesis

1 reference

Thank you to VIGNERON for bringing this lexeme to my attention!
Main issue: inconsistency

- Inconsistency within lexemes.
- Inconsistency across lexemes.
- (Unwarranted) inconsistency across languages.
Inconsistency within lexemes

- Inconsistency between lexeme-level statement and forms' grammatical features.
- Redundant repetition of lexeme-level statements (here the lexical category).
Inconsistency within lexemes

- Inconsistency use of language codes.
- Why is en-x-Q7976 used instead of en-us? (Answer: it used to be a technical limitation)
Inconsistency within a single form

- The form contains contradictory grammatical features.
- There is no machine-readable indication of a disjunction.
- The lexeme-level statement is enough.
- For NLG purposes, it should be augmented with a preferred gender (which I've done).
Inconsistency across lexemes

- Should **third-person singular** be represented as one or two features?
- And how should the (English) present tense be represented?

L1885-F2 | has en
--- | ---
Grammatical features: third person, singular, simple present

L1883-F4 | is en
--- | ---
Grammatical features: third-person singular, present indicative

L3006-F2 | goes en
--- | ---
Grammatical features: simple present, third-person singular
The solution?
A linguistic model of lexicographic data
But this has already been done?

- **Documentation pages** for languages
- **Lexeme forms** enforce consistency when creating lexemes
- **Lexical masks** serve to validate data
- **The problem**: these are not always consistent among each other...
- ... and may represent different conceptions of such models.
Basic principles

Disclaimer: the following slides use an assertive tone, but discussion is welcome!
Use lexeme forms for grammatical* inflection

“one for each relevant combination of grammatical features”

*Grammatical = morpho-phonological
And other variants?

- **Regional or dialectal variation**: insofar the pronunciation of grammatical features differ - create distinct lexemes (with appropriate language code).
- **Orthographic** or “light” dialectal **variation** - use spelling variants.
- **Historical variation**: use qualified statements, ranks (single preferred rank)
- **Note**: abbreviation is not a grammatical feature!
  - Frequently occurring abbreviations may be treated as spelling variants.
  - Domain-specific variations could be handled in statements.
Use lexeme statements for recurring features
Prefer “atomic” features

Third-person singular → third person & singular
Present indicative → present tense & indicative mood
Define a concise set of features

1. Per language
2. Per part-of-speech
3. Universally
Inventory of features

- Each part-of-speech requires specific grammatical categories (feature types).
- Lexeme-level categories need a corresponding property.
- Each feature can take specific values.
- The feature values should be *instances of* the feature type.
Examples

Swedish verbs
- Tense: past, present
- Voice: active, passive
- Mood: infinitive, imperative, participle (supine), indicative (unmarked)

Swedish nouns
- Gender: common, neuter
- Number: singular, plural
- Definiteness: definite, indefinite
- Case: unmarked, genitive

Swedish pronouns
- Gender: common, neuter, masculine, feminine
- Number: singular, plural
- Definiteness: definite, indefinite
- Case: nominative, genitive, oblique

Swedish adjectives
- Gender: common, neuter, masculine, feminine
- Number: singular, plural
- Definiteness: definite, indefinite, predicative (?)
- Comparison degree: positive, comparative, superlative
Hierarchy of features

- Grammatical features can be organized hierarchically.
  - To reflect this we can use the *subclass of* property.
    - Alternatively: create a new property such as *linguistic subtype of*
    - This may be qualified to apply only in certain languages
- The features are both instances and subclasses of the grammatical category

Gender hierarchy in Swedish.
Source: Gutman, Ivanov & Kirchner (2019)
Hierarchy of features: current state

Subclasses of gender (single level)

Sub-instances of gender (single level)

Source: Wikidata Graph Builder
Usage of: Unmarked features

- For a more sparse representation of lexemes we can use **unmarked features**.
- An unmarked form may represent either:
  - A form valid for all values of the unmarked category.
  - A default form which is overridden by a more specific one.
  - A stem from which regular forms can be derived.

In Wikidata:

<table>
<thead>
<tr>
<th></th>
<th>Sg.</th>
<th>Pl.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nom.</td>
<td>Vater</td>
<td>Väter</td>
</tr>
<tr>
<td>Gen.</td>
<td>Vaters</td>
<td>Väter</td>
</tr>
<tr>
<td>Dat.</td>
<td>Vater</td>
<td>Vätern</td>
</tr>
<tr>
<td>Acc.</td>
<td>Vater</td>
<td>Väter</td>
</tr>
</tbody>
</table>

In Wiktionary:

- *Vater* (strong, genitive *Vaters*, plural *Väter*, diminutive *Väterchen* m)
Workshop ideas

- Pick a documentation page on a specific language and improve it
  - What features, parts-of-speech are relevant for that language
- Improve/clean-up the type-hierarchy of one or more grammatical features
- **Pick a specific part of speech in a specific language and model it**
  - Improve language-specific documentation
  - Create/improve Entity Schemas for that part-of-speech
  - Create a script to edit Lexemes in a bulk in accordance with model
- Create scripts to clean inconsistencies in lexemes
- Create property proposals for missing properties
  - Missing lexeme-level properties (e.g. [Grammatical Person property](#))
  - Linguistic subtype of property
Thank you!

The floor is yours for discussion.