Validating the OntoLex-*lemon* lexicography module with K Dictionaries’ multilingual data

JULIA BOSQUE-GIL\(^1,2\), DORIELLE LONKE\(^3\), JORGE GRACIA\(^2\), ILAN KERNERMAN\(^3\)

[1] Ontology Engineering Group
Universidad Politécnica de Madrid, Spain

{jbosque,jogracia}@unizar.es

{dorielle,ilan}@kdictionaries.com

eLex’19: Smart Lexicography. October 1-3, Sintra, Portugal
1. Introduction and motivation
2. K Dictionaries multilingual Global Series
3. Problems of past RDF representations
4. The *lexicog* module
5. Methodology
6. Applying *lexicog*
7. Conclusions
The set of best practices for exposing, sharing and connecting data on the Web referred to as Linked Data (LD; Bizer et al., 2009) are progressively being adopted in lexicography.

LD enhances the tendency to standardise the ways of representation, query and enrichment of lexical content.
Facilitating interoperability with external resources, enhancing depth, precision and cross-linguality

Improved features are an advantage in the emerging multilingual digital single market, in Europe and eventually worldwide

Making KD data a golden standard for LD-compliant lexicographic data increases appeal and uniqueness in the private sector
OntoLex-lemon (McCrae et al., 2017) and its predecessor, lemon, have been the preferred choice by developers to convert lexicographic resources into LD
But there are situations in which **no perfect match** is available between the elements of the model and those found in lexicographic entries.

**OntoLex Module for Lexicography:**

*lexicog*
Goal

- Validate *lexicog* with an actual use case as well as to introduce some recommendations for future applications.

- Examine how the *limitations* of the OntoLex-*lemon* model already reported in the literature with respect to KD are successfully addressed by the module.
The data
Detailed and multi-layered lexicographic datasets
- Compiled with advanced corpus-based analysis tools
- Developed within a single, systematic framework
- Elaborate and robust XML Schema (DTD)
- Underlying monolingual layer can either be used on its own, or serve as a base for producing multilingual versions
- Complemented by idiomatic translation equivalents, individual sets can be cross-linked to other sets
- 25 language cores and nearly 100 language pairs
Junge $^1$ ['jʊŋə] 

*nm (gensg -n, Nordd. umg. nompl Jungs, [jʊŋs]) 1 Nordd. #Mädchen; männliches Kind*

{ar} - عَلَامَةٌ [wɔˈlaːmun] m sg, وَلْدَ، [ˈwalaːdun] m sg

◊ als ich noch ein kleiner Junge war

{ar} - حينما كنت لا أزال غلاما صغيراً.

2 coll=Kerl, Bursche; (junger) Mann

{ar} - شَاپُ [ˈʃaːbbun] m sg

◊ Der neue Lehrling ist ein tüchtiger Junge.

{ar} - المتدرب الجديد شاب كفء.

♦ Junge, Junge! drückt Staunen oder Bewunderung aus

{ar} - يَا لِلْعَجْبِ [jaː laːˈʕaːdʒabi] -

Junge $^2$ ['jʊŋə] 

*mt (gensg -n, nompl -n) zool sehr junges Tier*

{ar} - صَنِيْغَرُ الْحَيَّاَتُانَ [sˤa'ɾiːru laːˈħaːwaːni] m sg

◊ Unsere Katze hat Junge bekommen.

{ar} - وَلَدتْ قطتنا قططَا صغيرًا.
Problems of past RDF representations of KD’s data

- Previous conversions of KD Global series **monolingual** data with the **lemon** model (Klimek & Brümmer, 2015):
  - **Lexical relations** (e.g. compositional phrases and their relation to the elements embedding them)
  - Lack of **ontological references** for `lemon:LexicalSense` instances
  - Gaps in LexInfo, high amount of **ad-hoc** classes

- **OntoLex-lemon** (2016) providing **new classes and relations**
Problems of past RDF representations of KD’s data

- Previous conversions of KD Global series multilingual data with the OntoLex-\textit{lemon} model (Bosque-Gil et al. 2016a)
  - Part of \textit{LDL4HELTA}
  - \textit{Round-tripping} condition
Problems of past RDF representations of KD’s data

1. Loss of **structural information** reflecting lexical distinctions
2. Lack of elements for the **annotations** in the microstructure
3. Lack of **guidelines** for the application of OntoLex-*lemon* for lexicographic content
4. **Mismatches** between LexInfo elements and KD’s DTD tags and values
The model
Modelling existing lexicographic resources as LD by...

overcoming the limitations of OntoLex when modelling lexicographic information as LD in a way that is agnostic to the underlying lexicographic view and minimises information loss
Modelling existing lexicographic resources as LD by...

providing a model of linguistic objects in lexicography to capture the underlying original structure and annotations of the lexicographic entry in a way that keeps the purely lexical content separate from the lexicographic one
The methodology
An incremental approach was taken, starting with the basics of a single entry and gradually adding more complex elements.

Prior to conversion, each XML path was mapped to a corresponding LD element.

After the mapping stage, a URI naming strategy was established.
Checks that the predicates are in place and that the correct relations occur.

Checks that all necessary information is present and that nothing was left out during conversion.

Checks that only relevant information is present by limiting what could appear in the document.

Checks that the URIs are well-defined using Regular Expressions.
"title": "lexicog:entry instance",
"description": "a single dictionary entry belonging to a lexicographic resource",
"bsonType": "object",
"properties": {
  "@id": {
    "description": "lexicog:Entry URI",
    "bsonType": "string",
    "pattern": "^[kd-base:DE[0-9]{8}$"
  },
  "@type": {
    "bsonType": "string",
    "enum": [
      "lexicog:Entry"
    ]
  },
  "lexicographicEntryIn": {
    "description": "reversed predicate - lexicog:entry",
    "bsonType": "object",
    "properties": {
      "@id": {
        "description": "lexicog:LexicographicResource URI",
        "bsonType": "string",
        "pattern": "^[kd-base:mlds-[A-Z0-9]+""}
    },
    "@type": {
      "bsonType": "string",
      "enum": [
        "lexicog:LexicographicResource"
      ]
    }
  }
}
The instantiation
arte [ˈarte] nn/f 1 = inspiración; manifestación humana con intención estética

La música, la danza y la pintura son formas de arte.

Muziek, dans en schilderen zijn vormen van kunst.
Musikk, dans og maling er kunsttyper.
Musik, dansen og billedkunsten er kunststarter.
Music, dans och måleri är konstarter.

Music, dance and painting are art forms.

Distintas formas de arte han acompañado siempre a la humanidad.

Artes plásticas artes que utilizan el dibujo o el volumen: la pintura, la escultura y la arquitectura.
Dictionary Entry: *arte* (art)

Embedded compositional phrase: *artes plásticas* (fine arts)

Embedded synonym: *inspiración* (inspiration)
lexicog:Entry vs ontolex:LexicalEntry

:mlds-ES3 a lexicog:LexicographicResource;
dc:language "es";
lexicog:entry :ES DE00005536.

:ES DE00005536 a lexicog:Entry;
lexicog:describes :lexiconES/arte-n.

:lexiconES a lime:Lexicon;
Nested entries

Dictionary Entry: *besuchen* (visit. v), *Besuch* (visit. n), *Besucher* (visitor).

An element `<NestEntry>` grouping them together
Nested entries

:lexiconDE/besuchen-v a ontolex:LexicalEntry .
:lexiconDE/Besuch-n a ontolex:LexicalEntry . :lexiconDE/Besucher-n a ontolex:LexicalEntry .


Usage examples and their translations

- A sense with usage examples
- Each example in the source language has in turn a translation into the target language
Usage examples and their translations

:lexiconES/arte-n-SE00007455-sense a ontolex:LexicalSense ;
lexicog:usageExample
:lexiconES/arte-n-SE00007455-sense-TC00017355-ex .

:lexiconES/arte-n-SE00007455-sense-TC00017355-ex a
lexicog:UsageExample ;
rdf:value "La música, la danza y la pintura son formas de arte."@es ;
rdf:value "Muziek, dans en schilderen zijn vormen van kunst."@nl .
Conclusion
Conclusion

- *lexicog* addresses the loss of **structural and lexical information in the original resource**, and provides elements to capture data frequently found in lexicographic records.

- An **incremental approach + a JSON schema** → a solid first output to manually validate in subsequent steps towards a flawless conversion and linking to other sources.

- Future work: cross-lingual **linking** between different KD cores and to external resources.
Validating the OntoLex-\textit{lemon} lexicography module with K Dictionaries’ multilingual data

JULIA BOSQUE-GIL, DORIELLE LONKE, JORGE GRACIA, ILAN KERNERMAN

Acknowledgements:
Supported by the European Union’s Horizon 2020 research and innovation programme through the projects Prêt-à-LLOD, Elexis and Lynx.