Kernerman
DICTIONARY News

Number 27 • July 2019
kdictionaries.com/kdn

Lexicala API:
A new era in dictionary data

Ilan Kernerman and Dorielle Lonke

The Lexicala API is a REST API providing access to cross-lingual lexical data of K Dictionaries (KD) across 50 languages, including monolingual linguistic resources, over 150 language pairs, and numerous multilingual combinations. It enables flexible search options and returns JSON responses, as well as JSON-LD encoding RDF representation of KD data for Linked Data purposes and full integration with Semantic Web technologies. The formal release is on July 11, 2019.

1 Lexicala API: A new era in dictionary data | **Ilan Kernerman and Dorielle Lonke**
4 **K Dictionaries & Lexicala Workshops**
5 **Abstracts from the Globalex Workshop on Lexicography and Neologism**
Linguistics terminology and neologisms in Swahili: Rules vs. practice | **Gilles-Maurice de Schryver and Jutta De Nul**
Beyond frequency: On the dictionarisation of new words in Spanish | **Judit Freixa and Sergi Torner**
New words for the *Duden* | **Kathrin Kunkel-Razum**
New Estonian words and senses: Detection and description | **Margit Langemets, Jelena Kallas, Kaisa Norak and Indrek Hein**
A system for evaluating multiple data inputs to prioritize neologisms for inclusion in dictionaries | **Katherine Connor Martin**
Using the Hypothes.is web annotation tool for neologism collection | **Erin McKean**
The Korean Neologism Investigation Project: Current status and key issues | **Kilim Nam, Sujin Lee and Hae-Yun Jung**
New words in Japanese and the design of *UniDic* electronic dictionary | **Teruaki Oka**
Adding neologisms to the Hebrew online dictionary *Rav-Milim* | **Noga Porath**
The formation of neologisms in a lesser used language: The case of Frisian | **Hindrik Sijens and Hans Van de Velde**
Anglicisms and language-internal neologisms: Dealing with new words and expressions in *The Danish Dictionary*| **Lars Trap-Jensen**
Exploring criteria for the inclusion of trademarks in general language dictionaries of Modern Greek | **Anna Vacalopoulou**
Neologisms in a Dutch online portal | **Vivien Waszink**
14 Lexicography in higher education institutions: European Master in Lexicography with an Erasmus Mundi joint degree | **Stefan J. Shierholz**
18 Dictionaries of the future – the future of dictionaries: Challenges for lexicography in a digital society | **Stefan J. Shierholz**
19 Jacek Fisiak (1936-2019) | **Arleta Adamska-Sałaciak**
20 Deny Arnos Kwary. In-Memoriam | **Dora Amalia. Sandro Nielsen**
22 AsiaLex 2020: Lexicography and Language Documentation | **Dora Amilia and Luh Anik Mayani**
22 **META-Forum 2019**
23 Adam Kilgarriff Prize 2019 | **Michael Rundell**
24 **K Dictionaries & Lexicala News**

Editor | **Ilan Kernerman**

K DICTIONARIES LTD
8 Nahum Hanavi Street
Tel Aviv 6350310 Israel
+972-3-5468102
kdl@kdictionaries.com
https://lexicala.com

ISSN 1565-4745

The Lexicala API Team

Maayan Orner, Manager

Vova Dzhuranyuk, Developer

Roi Sadika, Developer

Dorielle Lonke, Coordinator

1. Resources

The Lexicala API offers data from three different KD resources: *Global*, *Password*, and *Random House*.

The Global series. A network of multi-layered and inter-linked lexicographic datasets for 25 European and Asian languages. Each language has at the basic layer its own monolingual core, featuring detailed and varied semantic and syntactic information, including alternative spellings and scripts; phonetic transcription; grammatical categorization, gender and number; sense disambiguation and attributes such as synonyms, antonyms, subject domain, register, etc.; examples of usage and different types of multiword units. Most of these cores (22 languages) have translation equivalents for each sense, example and expression to at least one other language (e.g., Korean to Japanese) and up to 18 languages (in the case of French). When several bilingual versions are available, they are juxtaposed and form a multilingual set, i.e. each item includes translations in several languages.

The Password series. A semi-bilingual English learner's dictionary with translations in 45 languages. The English entries include a definition and example(s) of usage for each sense, as well as a brief translation equivalent of that sense of the headword, including for each multiword unit and sub-entry. Most of the language versions are complemented by a bilingual index to the specific sense(s) of the English entry. The index is then expanded into a multilingual glossary by automatically adding the other language translations of *Password*, thus generating translations indirectly from the core language to any of the other languages via the English intermediary.

Random House Webster's College Dictionary (RHWCD). A comprehensive monolingual dictionary of American English. The last edition of this legacy dictionary was published by Random House in 2005, and in 2009 KD acquired full rights for its use. It includes over 133,000 entries with 191,000 senses, and offers a deep and extensive description of contemporary English language, including etymological, geographical and biographical information. KD has continued to reformat the data and update the contents in making RHWCD its flagship English dictionary, serving both native speakers of English and upper-level non-native users.

These resources combine various forms of human-curated and machine-generated content, and are supplemented by rich morphological lists of inflected word forms.

2. Infrastructure and Functionalities

The Lexicala API uses ElasticSearch as a back end, utilizing language specific functionalities for a flexible, fast search, that deals with natural language challenges for multiple languages. It is hosted on Amazon Web Services (AWS), prioritizing service reliability and scalability.

A basic API search is performed by looking up a headword, which returns all corresponding entries. This type of query returns a JSON document containing partial lexical information on entries that match the search criteria, including their unique entry ID. It is also possible to search for inflected forms, as well as by grammatical gender, number, part of speech and subcategorization, to obtain more specific results.

The inflected forms are provided either by morphological word form lists or by an automated *stemmer* functionality, allowing for greater flexibility when searching for a specific lemma. The purpose of the stemmer is to create a stem form from the analyzed word. The stem does not have to be a valid word, for example, a stemming algorithm can reduce *fishing*, *fished* or *fishes* to the stem *fish*, or the words *argue*, *argued* or *arguing* to the stem *argu*.

It is also possible to query the entire collection of dictionary entries (or senses) by a unique entry (or sense) ID. This type of response consists of a full dictionary entry, including translations, syntactic and semantic information, compositional phrases and usage examples. This type of result contains elaborate information about each headword, and offers all available translations at once, emphasizing the uniqueness of KD resources in multi- and cross-linguality and linking between languages and datasets.

3. RDF and Linked Data

Besides relying on unique multilingual resources, a prominent feature setting the Lexicala API apart from other dictionary APIs is the option of obtaining JSON-LD formatted RDF representation of lexical data, designed for Linked Data (LD) interoperability.

The RDF data is modelled according to the state-of-the-art *Lexicog* module of the OntoLex-*lemon* model – which was designed with linguistic and lexicographic data in mind and constitutes the *de facto* standard of lexicographic data representation in RDF – and is serialized in JSON-LD, a popular format in the Semantic Web (cf. W3C Ontology-Lexica Community Group at https://www.w3.org/community/ontolex/ and *Lexicog* module specification at http://www.w3.org/ns/lemon/lexicog#).

LD methods are at the forefront of the current generation of powerful language

technology solutions, and are at the heart of human-machine interaction. Providing quality cross-lingual lexical data, with the LD-driven option of linking to other sources, substantially widens the offering of data resources to Lexicala API users and for various integration, research and development purposes. The possibility of linking KD data to other enriched or annotated resources can be of great value for NLP and machine-learning tasks, which places the Lexicala API in leverage to other Dictionary APIs, providing added value in computational aspects to traditional lexicography and language related content.

4. Market and Users

The Lexicala API targets a broad range of users, starting with individual developers of a wide variety of applications who are looking for quality lexical data with rich multilingual extensions, through NLP researchers and computer scientists in need of large lexical corpora for processing, parsing or training a machine, and on to all types and sizes of online and offline translation, localization, learning and other language services.

The possibilities for using KD data are diverse, as the varying focal points of each resource and the wide selection of languages and information offer solutions to many different issues, and the flexible search and accessibility allow easy processing and seamless integration with other applications.

The new Lexicala API has an important role in two ongoing projects funded by the European Union's Horizon 2020 research and innovation programme, in which KD is participating: *Lynx* – Building the Legal Knowledge Graph for Smart Compliance Services in Multilingual Europe (No. 780602, http://lynx-project.eu/); and, *Elexis* – European Lexicographic Infrastructure (No. 731015, https://elex.is/).

Sample API extracts:
1. excerpt from *Abbau* in Global German
2. *chair* (noun) in the Password multilingual set
3. *smile* in RHWCD
4. results for *azul* in Global Spanish, when searching by parameters

1.


```
        },
        "tr": {
          "text": "atlıkarıncanın sökülmesi"
        }
      }
    }
  ]
},
{
  "id": "DE_SE00000042",
  "semantic_subcategory": "von Bodenschätzen",
  "translations": {
    "ar": [
      {
        "text": "إستِخراج",
        "pronunciation": {
          "value": "ʔistixˈraːʤun"
        },
        "gender": "masculine",
        "number": "singular"
      },
      {
        "text": "إستِدلال",
        "pronunciation": {
          "value": "ʔistiyˈlaːlun"
```


2.


```
  },
  "ko": {
    "text":
  },
  "lt": {
    "text":
  },
  "lv": {
    "text":
  },
  "ml": {
    "text":
  },
  "nl": {
    "text":
  },
  "no": {
    "text":
  },
  "pl": {
    "text": "krzesło"
  },
  "prs": {
    "text": "صندلی"
  },
  "ps": {
    "text": "چوکی: کرسی: رئیس (دغونډی"
  },
  "pt": {
    "text": "cadeira"
  },
  "ro": {
    "text": "scaun"
  },
  "ru": {
    "text": "стул"
  },
  "sk": {
    "text": "stolička"
  },
  "sl": {
    "text": "stol"
  },
  "sr": {
    "text": "stolica"
  },
  "sv": {
    "text": "stol"
  },
  "th": {
    "text": "เก้าอี้"
```


3.


```
{
  "id": "RDE00064769_0",
  "source": "random",
  "language": "en",
  "headword": {
    "text": "smile",
    "pronunciation": {
      "value": "smaɪl"
    },
    "pos": [
      "verb",
      "noun"
    ]
  },
  "senses": [
    {
```


4.


```
{
  "n_results": 2,
  "page_number": 1,
  "results_per_page": 10,
  "n_pages": 1,
  "available_n_pages": 1,
  "results": [
    {
      "id": "ES_DE00006683",
      "language": "es",
      "headword": {
        "text": "azul",
        "pos": "noun"
      },
      "senses": [
        {
          "id": "ES_SE00009139",
          "definition": "color parecido al cielo"
        }
      ]
```