Lexicala's multilingual lexical data solutions for language service providers

Lexicala by K Dictionaries is launching a new range of services for the language technology industry and the academia. The focus is on multilingual lexical data solutions that merge human created and curated language resources with automatically generated content. The solutions feature expert parallel corpora, domain classification, morphology, text annotation, and other premium resources for enhancing natural language processing (NLP) tasks and training machine learning (ML) models.

Lexicala is the trade name used by K Dictionaries (KD) to offer cross-lingual resources to language service providers (LSPs) and for collaboration in research and innovation projects. It relies on and expands datasets and methodologies developed by KD over the last three decades, which eventually enable infinite ways of extracting refined monolingual, bilingual and multilingual components and applying them for translation, spellchecking, semantic technologies, speech recognition, knowledge management, language learning, as well as for online dictionaries.

KD has been an international dictionary leader since the 1990s. It began with the breakthrough semi-bilingual dictionary for learners of English, published by prominent local publishing partners in dozens of language editions for millions of students worldwide, and reaching many more users in recent years in digital formats, mainly through collaboration with Cambridge University Press & Assessment on Cambridge Dictionary online, the world's no. 1 website for English learner's dictionaries. At the turn of the century, KD became involved in innovative multilingual dictionaries, starting in cooperation with Finland's leading electronic dictionary publisher, Kielikone, and currently made available on Naver's Open Dictionary PRO platform. In the early 2000s, it pioneered a new approach to creating multi-layer lexicographic resources as part of the Global series, built on an extensive core for each language. This monolingual core serves as a base for developing bilingual pairs and multilingual networks, which are all inter-connected within an overall framework sharing a common technical infrastructure and interoperable with exterior resources. In the last decade, KD was proactive in pioneering the incorporation of lexicographic data and linguistic linked (open) data technologies, primarily in connection with the W3C OntoLex community group, as well as in EU projects, particularly with Semantic Web Company.

lexicala
by K DICTIONARIES

LITHME WG1 Workshop, University of Luxembourg, Esch-sur-Lazette
Human Language Data for Machine Learning. Ilan Kernerman
LITHME – Language In The Human-Machine Era – COST Action CA19102
Working Group 1: Computational Linguistics
5-6 September 2022

SEMANTiCS & LTI 2022, Vienna, Austria
Semi-automated Generation of Multilingual Domain Taxonomies. Ilan Kernerman and Martin Kaltenböck
15 September 2022

LLODream, Vilnius, Lithuania
Linking lexicographic resources to language proficiency-level applications. Kris Heylen, Ilan Kernerman and Carole Tiberius
21 September 2022

TAUS 2M 2022, San Jose, CA, USA
Massively Multilingual Conference & Expo
11-13 October 2022

COLING 2022, Gyeongju, Korea
The 29th International Conference on Computational Linguistics
12-17 October 2022

The first major step in the new Lexicala direction has been accomplished this year with the delivery of expert parallel corpora to one of Asia's top information technology conglomerates and completing the integration of the first language pair into its Neural Machine Translation systems. As part of this venture, Lexicala provided over a quarter-of-a-million bilingual segments to train ML models and improve the translation engines' performance. The segments consisted of usage examples composed of full sentences and short phrases that were extracted from dictionary entries between the required Asian language and European ones. The corpora were developed by converging manually created content with smart data generation methods, then followed by thorough curation of each sentence pair by local language experts. Contracting with such a leading global player demonstrates the ability to offer highest quality cross-lingual lexical data supported by automated processes and perfected by professional linguists and translators.

Lexicala's expert parallel corpora resources contain millions of bilingual and multilingual segments in more than 20 languages, including low-resourced combinations. Moreover, many of these segments can apply to languages for specific purposes, as they stem from selected senses of dictionary entries that include a subject field label.

Besides the existing collection of KD resources that encompass 50 languages, the new Lexicala services aim to cover all human languages. They include bilingual and multilingual translation, domain classification, morphological forms, part-of-speech tagging, sense alignment, name entity annotation, IPA and audio pronunciation, and other cross-lingual NLP applications. This underlines Lexicala's expansion to the world of AI and illustrates its potential to offer outstanding solutions to the language technology community.

Ilan Kernerman
CEO, Lexicala by K Dictionaries

lexicala Solutions