LTI SUMMIT 2019 ◦ BRUSSELS ◦ 24 JUNE 2019

Lexicala API

Ilan Kernerman

Transforming Dictionary Products into Cross-lingual Lexical Data Services

KDICTIONARIES

lexicala

HIGHLIGHTS

- Transformation & Innovation
- Sources & Resources
- Infrastructure & Functionalities
- RDF & Linked Data
- Market & Users

TRENDS

Globalization

systemize | standardize | uniformize → ***localize***

+

Digitization

automize | customize | personalize

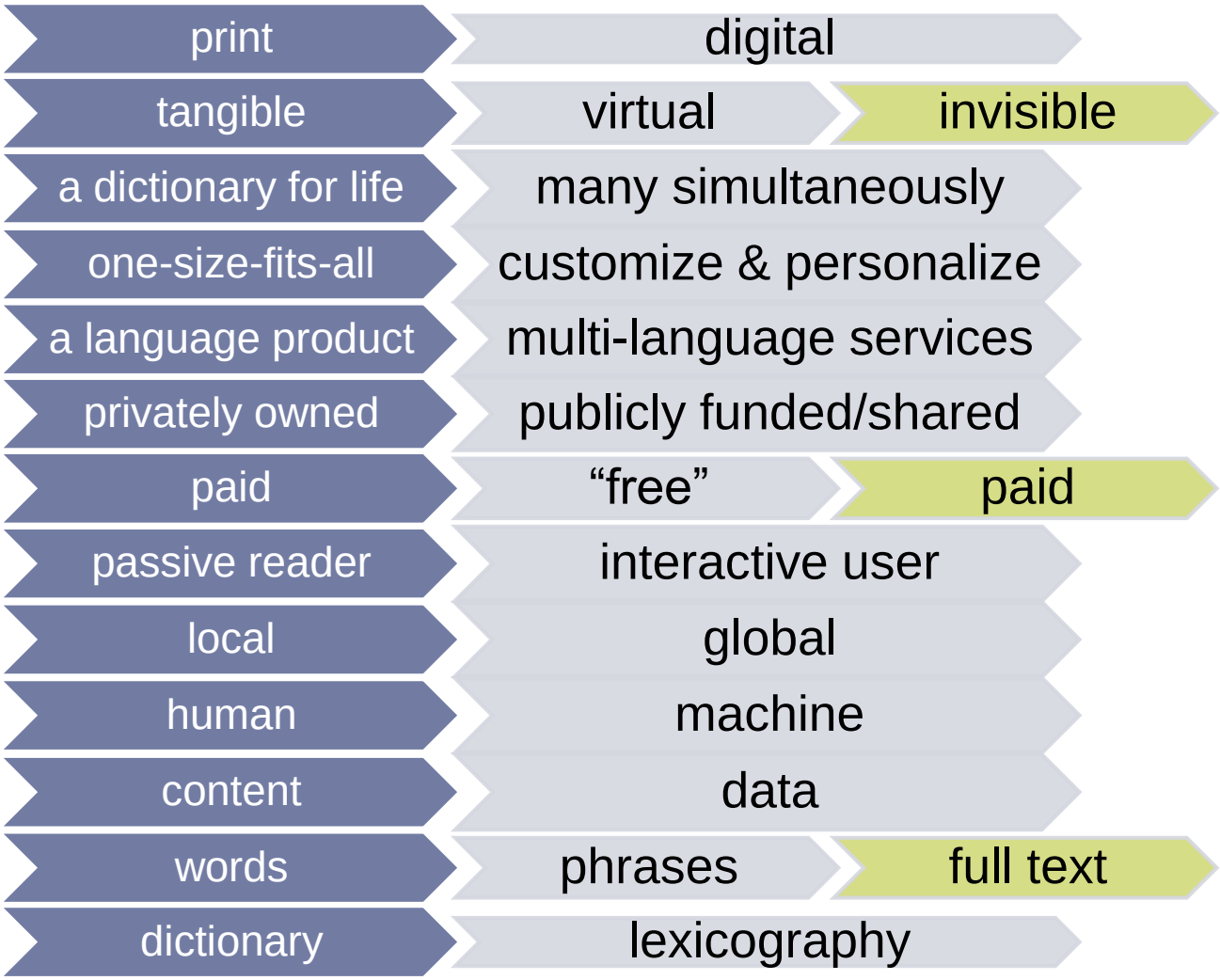
human→machine | content→data

=

Datacization

the process of transforming information resources previously accessed directly by humans into resources primarily accessed by software (Erin McKean, 2017)

TRANSFORMATION



K DICTIONARIES

- Global dictionary leader, founded 1993, Tel Aviv
- Collaborate worldwide
  - industry
  - academia
  - professional associations
- Create multilayer crosslingual lexical resources
  - auto-generated
  - human-curated
  - machine-oriented

LEX → NLP

- machine translation & localization
- e-learning & deep learning
- multilinguality
- word processing, spellcheck, auto-complete
- text-to-speech, speech-to-text, OCR
- search engine optimization
- knowledge graphs
- information retrieval
- *word sense disambiguation & induction*

INNOVATION

- Evolve dictionary into lexical data
- Enhance interoperability by Linked Data methods
- Offer smart multilingual and cross-lingual features
- Combine manual/automatic content generation
- Address both human and machine use
- Cooperate with NLP and knowledge stakeholders
- Interact with end users and public bodies
- Adjust market strategy and dissemination models

Natural Lexicography Processing

RESOURCES

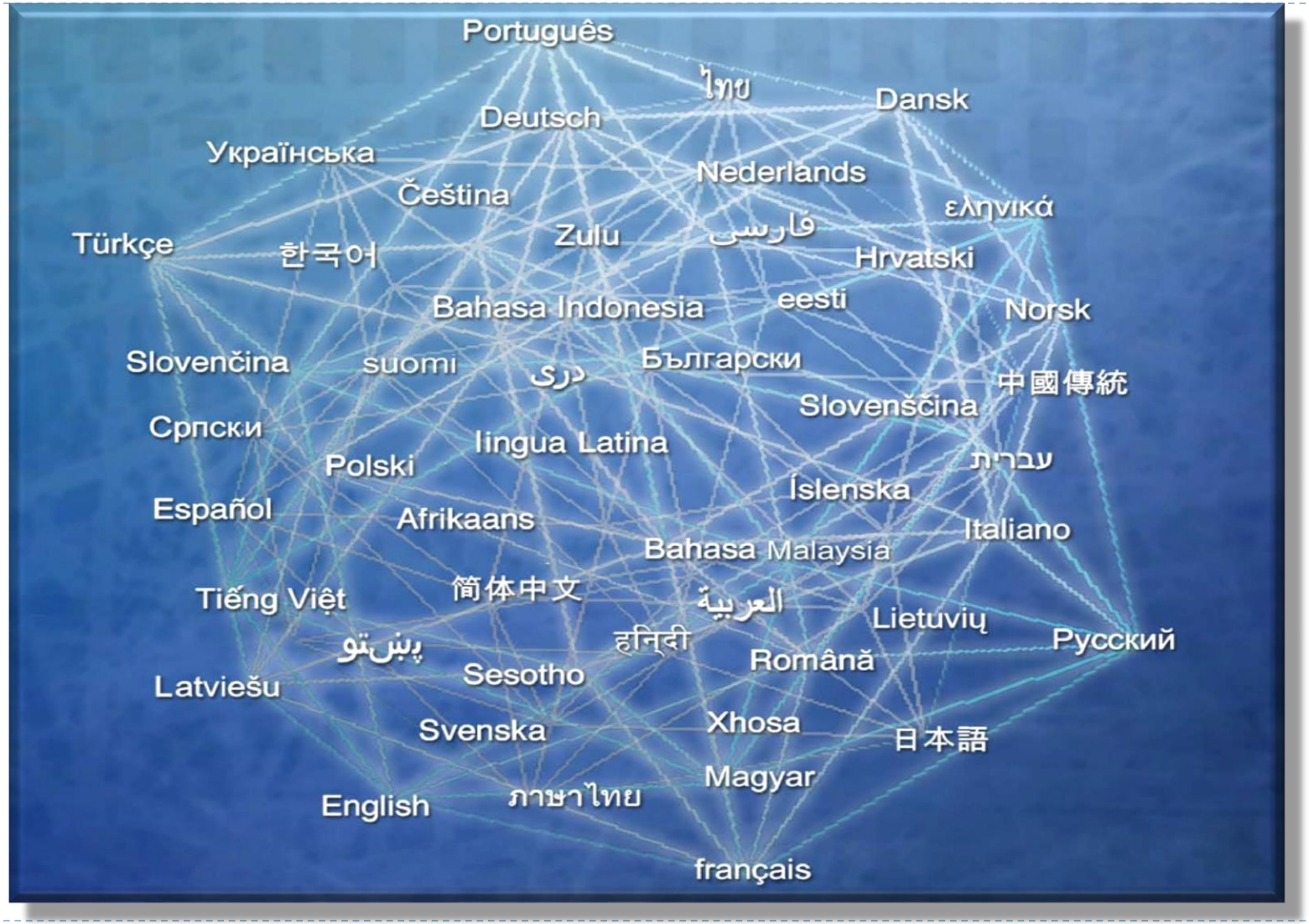
- 50 languages
- mono- / bi- / multilingual lexicographic sets
- XML / RDF / JSON formats
- morphology / pronunciation / supplements
- editorial & data processing tools
- online & mobile applications
- REST API

lexicala

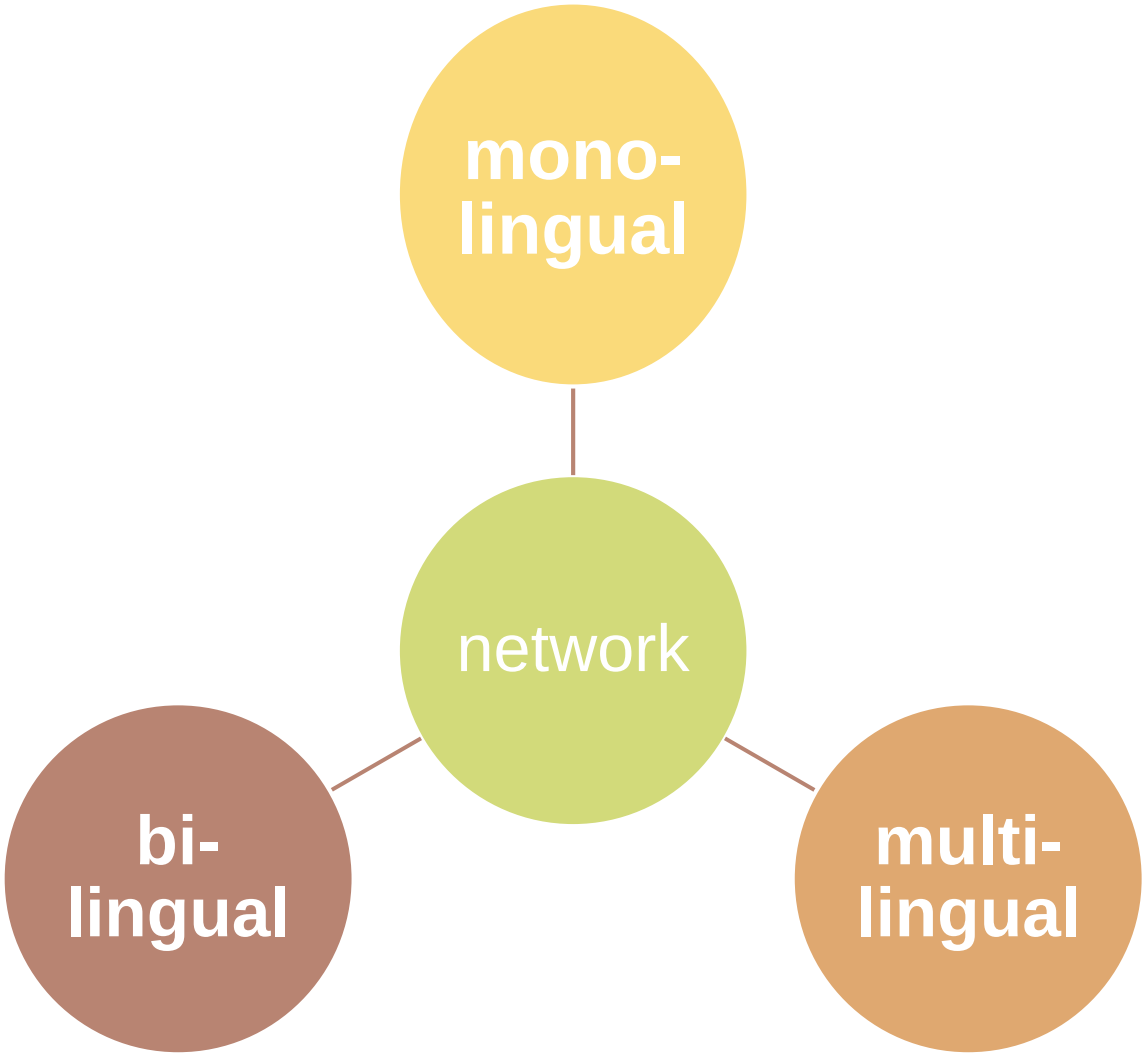
LANGUAGES

Afrikaans | Arabic | Azerbaijani | Bulgarian | Catalan | Chinese Simplified / Traditional | Croatian | Czech | Danish | Dari | Dutch | English | Estonian | Farsi | Finnish | French | Frisian | German | Greek | Hebrew | Hindi | Hungarian | Icelandic | Indonesian | Italian | Japanese | Kazakh | Korean | Latin | Latvian | Lithuanian | Malay | Norwegian | Pashto | Polish | Portuguese Brazil / Portugal | Romanian | Russian | Serbian | Slovak | Slovene | Spanish | Swedish | Thai | Turkish | Ukrainian | Urdu | Vietnamese | Welsh

CROSS-LINGUAL



MULTI-LAYER



MAPPING

- AlternativeScripting
- AlternativeSpelling
- Antonym
- CompositionalPhrase
- CrossReference
- Definition
- Example
- GeographicalUsage
- GrammaticalGender
- GrammaticalNumber
- HomographNumber
- Lemma
- PartOfSpeech
- Pronunciation
- RangeOfApplication
- Register
- SenseIndicator
- SenseQualifier
- SubCategorization
- SubjectField
- Synonym
- [Translation]

GLOBAL

Gruppe [ˈgrʊpə] *w* (*Gen. Sg.* **Gruppe**, *Pl.* **Gruppen**)
1 <mit gemeinsamem Merkmal> gewisse Anzahl von Personen oder Dingen an einem Ort oder mit einem gemeinsamen Merkmal
AR جَمَاعَة [ʒaˈmaːʕatun] *f sg*, جَوْف [fawʒun] *m sg*, مَجْمُوعَة [maʒˈmuːʕatun] *f sg*
BR **grupo** *m*, **agrupamento** *m*
DK **gruppe** *common*
EN **group**
JA グループ
NL **groep** *m/f*
NO **gruppe** *m*
SV **grupp** *common*
TR **grup, küme**
◊ *eine Gruppe Schulkinder an der Bushaltestelle*
AR مَجْمُوعَة مِنَ التِّلْمِيذِ الَّذِينَ يَنْتَظِرُونَ الْحَافِلَة.
BR *um grupo de estudantes na parada do ônibus*
DK *en gruppe skolebørn ved busstoppestedet*
EN *a group of schoolchildren at the bus stop*
NL *een groep schoolkinderen bij de bushalte*
NO *en gruppe med skolebarn på bussholdeplassen*
SV *en grupp skolbarn vid busshållplatsen*
TR *otobüs durağındaki bir öğrenci grubu*

PASSWORD

jar[2] [dʒɑː] (*past tense*, *past participle* **jarred**) *verb*
1 (with **on**) to have a harsh and startling effect (on):
Her sharp voice jarred on my ears.

af knars, tril
ar يُحْدِثُ صَوْتا مُزْعِجا
az cingildəmək
bg дразня
br vibrar
ca desafinar, grinyolar
cs skřípat
de wehtun
dk skurre
el πειράζω, ερεθίζω
es chirriar, discordar
et (kõrva) riivama
fa اثر ناخوشایند داشتن
fi ärsyttää
fr écorcher
fy sear dwaan
he לִצרוֹם
hi खटकना
hr vrijeđati kome uho
hu sért
id menghantam
is nísta
it urtare
ja 耳ざわりである
ko 귀에 거슬리는 소리를 내다
lt rėžti
lv griezīgi skanēt
ml menyakitkan telinga
nl knarsen
no skrape, skurre
pl drażnić
prs اثر ناخوشایند داشتن
ps ناوړه زغ ایستل
pt vibrar
ro a irita, a zgâria
ru раздражать
sk škrípať, vŕzgať
sl praskati
sr parati
sv gnissla, skära
th รบกวน
tr bozuk ve çatlak ses çıkartmak
tw 發出刺耳聲且令人不舒服
uk дратувати, справляти неприємне враження
ur جسم میں لرزش پیدا کرنا
vi gây cảm giác khó chịu
zh 发出刺耳声

RANDOM HOUSE WEBSTER'S

fruit /frut/ *n*., *pl*. **fruits**, or (*esp*. collectively) **fruit**, *n*.
1. the edible part of a plant developed from a flower and containing one or more seeds with any accessory tissues, as the peach, mulberry, or banana.
2. the developed ovary of a seed plant with its contents and accessory parts, as the pea pod, nut, tomato, or pineapple.
3. any product of plant growth useful to humans or animals.
4. the spores and accessory organs of ferns, mosses, fungi, algae, or lichen.
5. anything produced or accruing; product, result, or effect; return or profit.
6. *Slang: Disparaging and Offensive*. a contemptuous term used to refer to a male homosexual.
--- *v.i*., *v.t*.
7. to bear or cause to bear fruit.

[1125–75; ME < OF < L frūctus enjoyment, profit, fruit, derivative of **fruī** to enjoy the produce of]

usage Definition 6 is a slur and should be avoided. It is used with disparaging intent and is perceived as insulting.

INFRASTRUCTURE

- Making use of Elasticsearch as a backend
  - utilizing language specific functionalities for a flexible, fast search
  - dealing with natural language challenges for multiple languages
- Hosted on Amazon Web Services (AWS)
  - prioritizing service reliability and scalability

SEARCH

- Basic search performed by looking up a headword
  - returns all corresponding entries
  - returns JSON document
  - including the entry's unique entry (or sense) ID
- Also possible to search for
  - inflected forms
  - grammatical gender and number
  - part of speech and subcategorization
  - syntactic and semantic information
  - compositional phrases
  - usage examples
  - translations

INFLECTED FORMS

- Provided either by
  – morphological word form lists, or
  – automated *stemmer* functionality
- Stemmer creates a stem form of the analyzed word
- Stem does not have to be a valid word, e.g.
  – reduce *fishing*, *fished* or *fishes* to the stem ***fish***
  – reduce *argue*, *argued* or *arguing* to the stem ***argu***

RDF

- Option of obtaining JSON-LD formatted RDF representation of lexical data, designed for Linked Data (LD) interoperability
- Modelled according to state-of-the-art *Lexicog* module of the OntoLex-*lemon* model

LD ADVANTAGES

- Linked Data methods are at the forefront of the current generation of powerful LT solutions, and at the heart of human-machine interaction
- LD-driven option of linking to enriched or annotated other sources, widens offering of data resources
- Added value for NLP and machine-learning tasks, expanding computational aspects of traditional lexicography and language related content

USERS

- **individual developers** of a multitude of applications and word games looking for reliable lexical data with rich multilingual extensions
- **NLP researchers and computer scientists** in need of large lexical corpora for processing, parsing or training a machine
- **Providers** of online and offline translation, localization, learning, and other language services

H2020 PROJECTS

- LYNX

  Building the Legal Knowledge Graph for Smart Compliance Services in Multilingual Europe

  Dec. 2017 – Nov. 2020

  €3M. Innovation Action

  10 partners

- ELEXIS

  European Lexicographic Infrastructure

  Feb. 2018 – Jan. 2022

  €5M. Research & Innovation

  17 partners

DEMO (1) basic search parameters

- source (*Global*, *Password, Random*)
- language (de, en, es, fr, ja, ko, ru, ...)
- text (*car*, *chair*, *working*, …)

This query returns all entries in the Spanish core (language=es) of the Global series (source=global) with the headword “azul” (text=azul). There are two entries – a noun and an adjective.

DEMO (2) specific semantic criteria

- pos (= noun, verb, ...)
- number (= singular, plural, …)
- gender (= masculine, feminine, …)
- subcategorization (= transitive, countable, …)
- monosemous (= true/false)
- polysemous (= true/false)

This query returns all entries in the Polish dictionary (language=pl) of the Global series (source=global) that are plural nouns (pos=noun, number=plural). There are 217 such entries.

DEMO (3.1) inflected forms & word stems

- morph (= true/false) – searches for the text in both headwords *and* inflections, including in supplementary morphological lists. This is based on existing human-curated data and semi-automated word form lists.

  Searching “houses” will return the entry “house” (noun) even though the word “houses” is not an entry in the English dictionary (it is a plural inflection of “house”).

DEMO (3.2) inflected forms & word stems

- analyzed (= true/false) – a *stemmer* algorithm that "strips" words to their stem, disregarding diacritics and case (uppercase/lowercase). This function is completely automatic.

  This query returns the entries "working" (adj.), "work" (verb), "work" (noun), "hard-working" (adjective), "working class" (noun), "work on" (verb) and any other entry with the stem "work" in its headword.

DEMO (4.1) searching by entry or sense ID

- For example, the entry ID of the entry “azul” (noun) in the Spanish core is **ES_DE00006683**.

  This query returns the complete entry “azul” (noun), a monosemous word (one sense) including an example (“*El azul es un color primario*”), compositional phrases (“azul marino”), a synonym (“añil”) and translations to multiple target languages.

DEMO (4.2) searching by entry or sense ID

- An example for a polysemous word, the word “Abbau” in German, which has three senses.

DEMO (4.3) searching by entry or sense ID

- To find the entry “chair” in the *Password* series, first obtain the entry ID for the entry “chair” – PW00003877 – then the query returns this entry with translations to 45 languages.

DEMO (4.4) searching by entry or sense ID

- Likewise, in *Random House Webster's College Dictionary*, obtain the entry "smile". ☺

THANK YOU

[θæŋk juː]

interjection

a common elliptical expression
used to express gratitude or appreciation

Thank you for your attention!

Thank you for your cooperation.

KDICTIONARIES lexicala

https://lexicala.com/