



Adventures in Semantic Interoperability

Casper Grathwohl

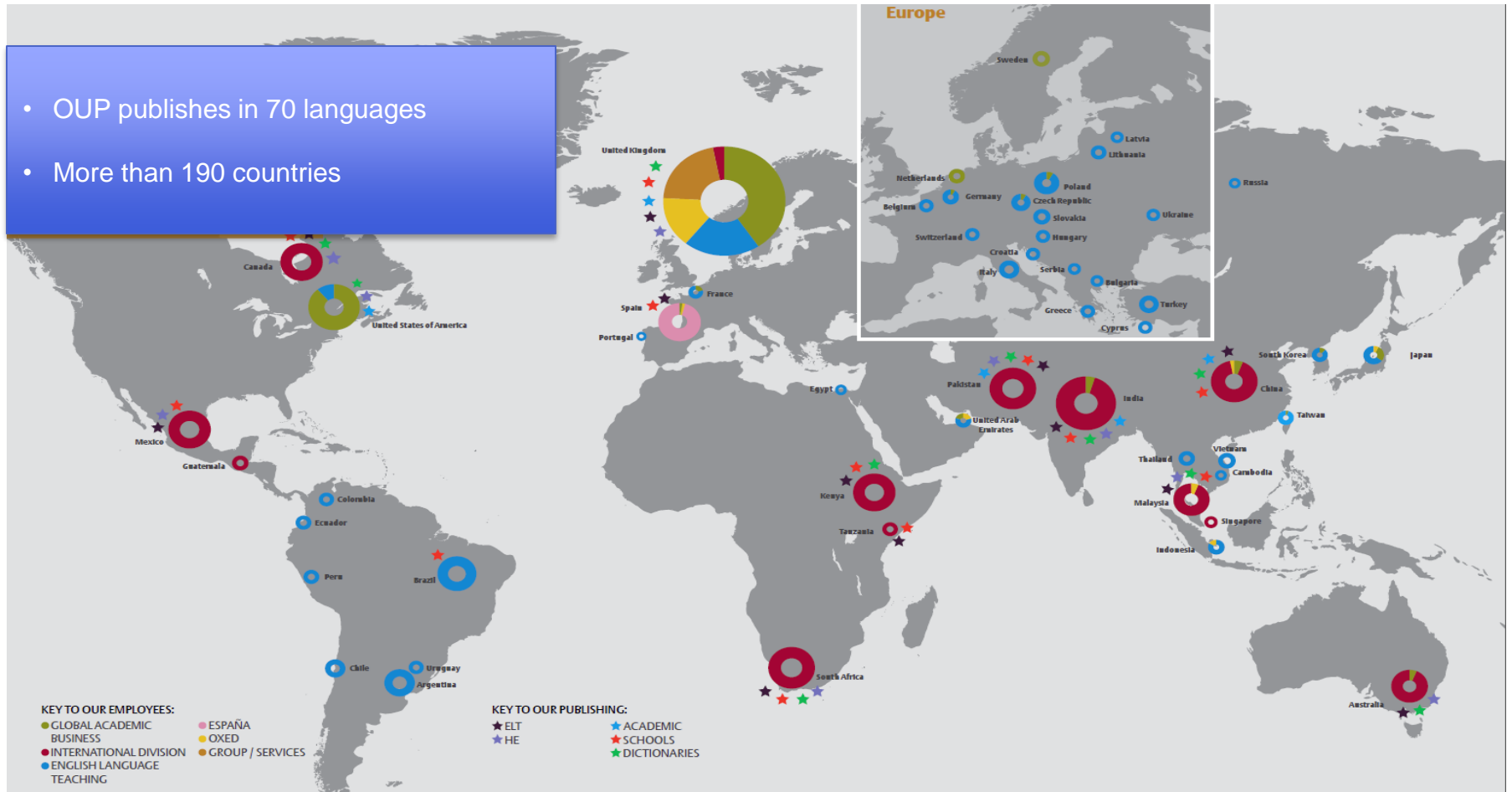
President, Dictionaries and Language Services

LT Innovate Industry Summit
25 July 2019



Oxford's global reach:

- OUP publishes in 70 languages
- More than 190 countries



Oxford Academic Dictionaries Vision

OXFORD
UNIVERSITY PRESS



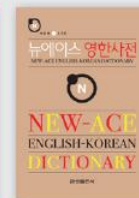
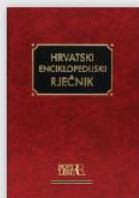
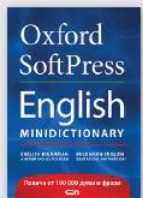
*To enable rich, diverse language
communication in the digital age*

Oxford Global Language Solutions

Available monolingual datasets

Language	No. of headwords	No. of definitions
Arabic	7,059	56,000
Bulgarian	20,000	36,000
Catalan	55,000	117,000
Chinese (simplified)	73,000	96,000
Chinese (traditional)	50,000	67,000
Croatian	69,000	169,000
Czech	110,000	85,000
Danish	100,000	61,000
Dutch	69,000	87,000
English (UK)	90,000	350,000
English (US)	90,000	350,000
Finnish	100,000	100,000

Language	No. of headwords	No. of definitions
Hungarian	75,000	150,000
Indonesian	36,000	59,000
Italian	96,000	145,000
Japanese	65,000	89,000
Korean	301,000	368,000
Latvian	30,000	27,000
Malay	20,000	20,000
Norwegian	81,500	86,000
Polish	100,000	130,000
Portuguese	228,500	376,500
Romanian	67,000	100,000
Russian	41,000	100,000





T-Mobile 2:56 PM

General Dictionary

British English
Oxford Dictionary of English

American English
New Oxford American Dictionary

Spanish
Diccionario General de la Lengua Española

Spanish-English
Gran Diccionario Oxford - Español-Inglés

Apple Dictionary

Simplified Chinese
现代汉语规范词典

Simplified Chinese-English
牛津英汉汉英词典

Traditional Chinese
五南國語活用辭典

Danish
Politikens Nudansk Ordbog

Dutch
Prisma woordenboek Nederlands

7:37 PM

Dictionary

Japanese
スーパー大辞林

Japanese-English
ウイズダム英和辞典 / ウイズダム和英辞典

Korean
뉴에이스 국어사전

Korean-English
뉴에이스 영한사전 / 뉴에이스 한영사전

Portuguese
Dicionário de Português licenciado para Oxf...

Russian
Толковый словарь русского языка

Simplified Chinese
现代汉语规范词典

Simplified Chinese-English
牛津英汉汉英词典

Spanish
Diccionario General de la Lengua Española Vox

Spanish-English
Gran Diccionario Oxford - Español-Inglés • In...


Thai
พจนานุกรมไทย ฉบับทันสมัยและสมบูรณ์

Turkish



About 22,100,000 results (0.84 seconds)

spec·tac·u·lar

/spek'takjələr/ 

adjective

adjective: **spectacular**

1. beautiful in a dramatic and eye-catching way.

"spectacular mountain scenery"

synonyms: [striking](#), [picturesque](#), [eye-catching](#), [breathtaking](#), [arresting](#), [glorious](#); *informal* out of this world

"a spectacular view"

antonyms: [unimpressive](#), [dull](#)

- strikingly large or obvious.

"the party suffered a spectacular loss in the election"

synonyms: [impressive](#), [magnificent](#), [splendid](#), [dazzling](#), [sensational](#), [dramatic](#), [remarkable](#), [outstanding](#), [memorable](#), [unforgettable](#)

"a spectacular victory"

antonyms: [unimpressive](#)

noun

noun: **spectacular**; plural noun: **spectaculars**

1. an event such as a pageant or musical, produced on a large scale and with striking effects.

Origin

ENGLISH

spectacle

ENGLISH

oracular

→ spectacular

late 17th century

late 17th century: from [spectacle](#), on the pattern of words such as *oracular*.

新 牛津词典

柯林斯词典

英英释义

双语例句

继续查词

牛津词典

verb

1 ~ (sth) (from sth) (into sth) | ~ sth (as sth)

to express the meaning of speech or writing in a different language 翻译；译

[VN] He translated the letter into English. 

他把这封信译成了英文。

Her books have been translated into 24 languages. 

她的书被译成了24种语言。

'Suisse' had been wrongly translated as 'Sweden'. 

Suisse被错译成Sweden (瑞典)。

Can you help me translate this legal jargon into plain English? 

你能帮助我用浅显易懂的英语来说明这一法律术语吗？

[V] I don't speak Greek so Dina offered to translate for me. 

我不懂希腊语，于是迪娜主动给我翻译。

My work involves translating from German. 

我的工作包括德语翻译。

2 [V] ~ (as sth)

to be changed from one language to another 被翻译；被译成



Search the Dictionary

FAMED



famed [feɪmd]

adjective

1. known about by many people; renowned
2. widely reported or rumoured.

ORIGIN

Middle English: past participle of archaic fame (verb), from Old French famer, from Latin fama

WORD POPULARITY (30 DAYS)



Played often

Played rarely

Oxford Dictionaries:

To enable rich, diverse language communication in the digital age



www.oed.com



www.oxforddictionaries.com



[Language Solutions](#)



[API & Developer Program](#)



<http://en.bab.la/dictionary>



[Oxford Global Languages](#)

Asset Creation

Off-the-shelf Data

- Monolingual lexical data and wordlists
- Bilingual lexical data
- Example sentence banks at sense level
- Synonym content (Thesauri)
- Inflected forms linked to monolingual or bilingual dictionary data
- Corpora-derived n-grams and frequency
- Human audio pronunciation files
- Hyphenation information
- Morphology

One consistent XML structure with intelligent metadata

All linked at a sense level within a language

Asset Creation

New Words and Prioritization Engine

OXFORD
UNIVERSITY PRESS

- Our “Asset Creation” team is always acquiring new language content and data
- This includes new words and new language data sets (i.e. via corpus)
- Much of this is automated
- This is backed up by Oxford’s word renowned lexicographers and editorial team.



Oxford Language Data

OXFORD
UNIVERSITY PRESS

Google

YAHOO!

facebook

amazon



Microsoft

SONY

CASIO.

Baidu 百度



Tencent
腾讯

SAMSUNG

zynga.



Xiaomi

SHARP

bing



PlayStation

Sogou 搜狗

Academic Dictionaries Developer and API Programme

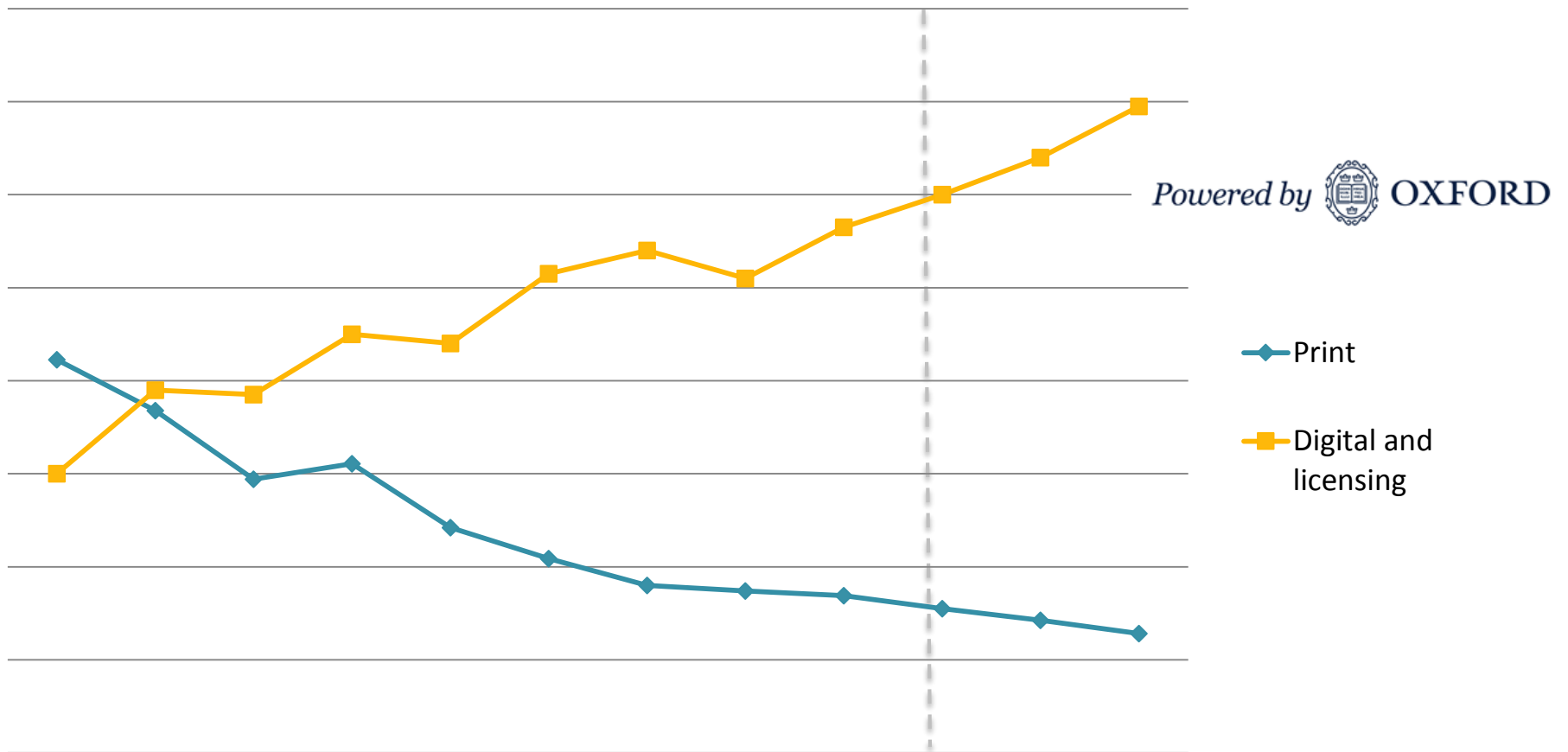
OXFORD
UNIVERSITY PRESS



The image shows a screenshot of the Oxford Dictionaries API developer page. At the top left is the Oxford Dictionaries logo. To the right are navigation links: DOCUMENTATION, SUPPORT (with a dropdown arrow), USE CASES, SIGN IN, and a REGISTER button. The main content area features a large grid of diverse human faces. Overlaid on this grid is the text "Oxford Dictionaries API" in a large blue font. Below this, in a smaller grey font, is the text "Enhance your app with our world-renowned dictionary data." At the bottom of this text block is a blue button with the text "GET STARTED" in white.

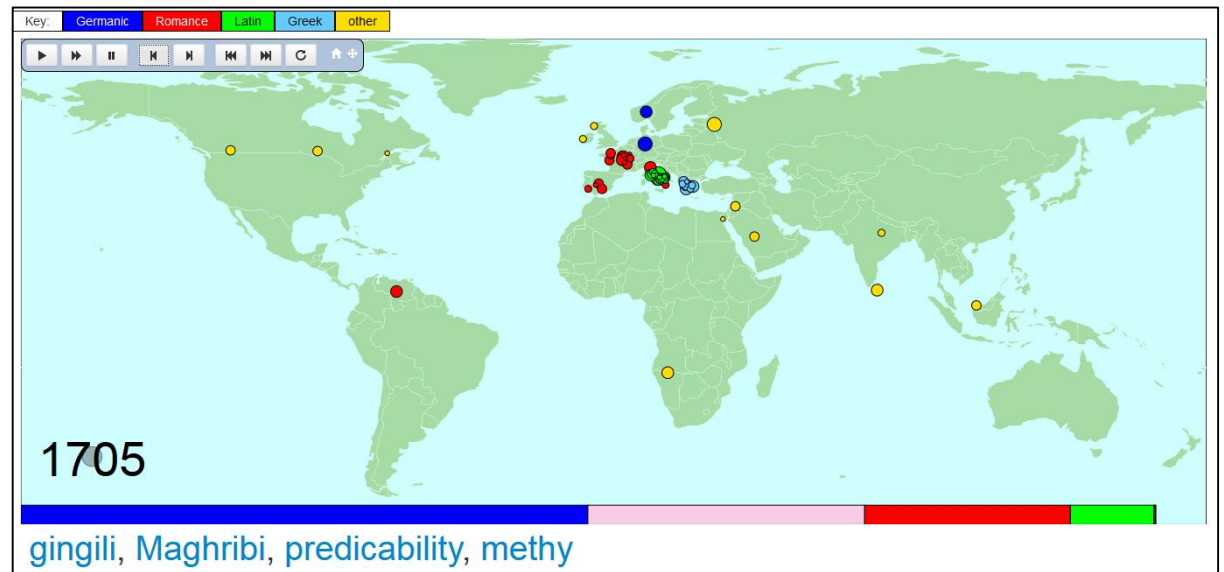
<https://developer.oxforddictionaries.com/>

Revenue by category 2010-2020



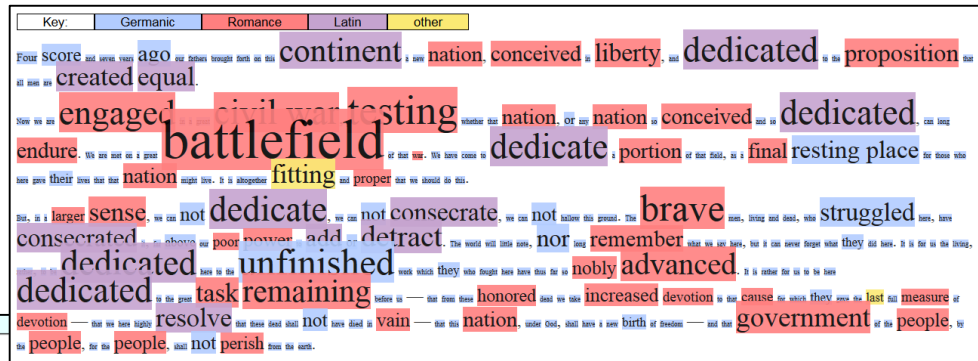
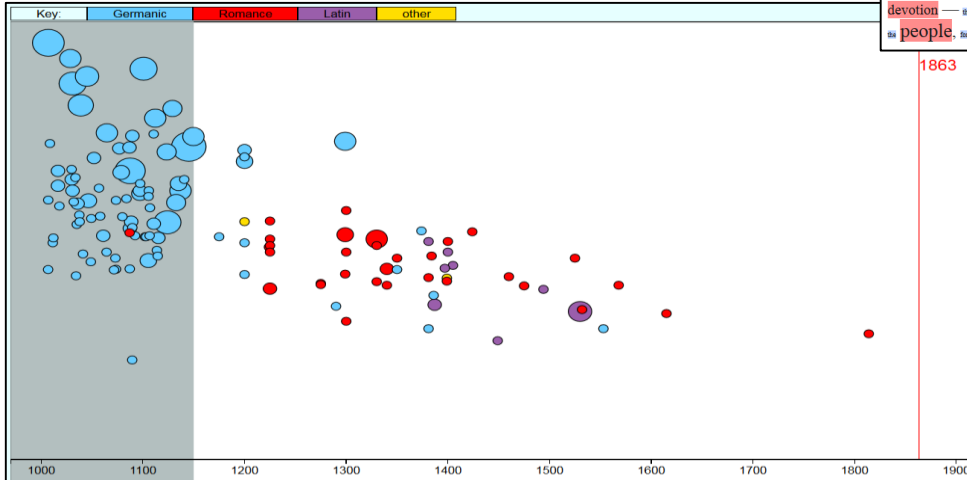
The OED in two minutes

An interactive animation exploring when and how words entered the English language, and how the lexicon as a whole developed as a result. This exploits date, etymology, and frequency information across the OED dataset.



Example applications: Text metrics

This application takes a piece of text – historical or modern – and maps each word to OED data in order to explore and visualize the author’s style, choice of vocabulary, use of neologisms, etc.



It can even rewrite pieces of text using alternative vocabulary from different historical periods. Try turning ‘Call me Maybe’ into a medieval lyric:

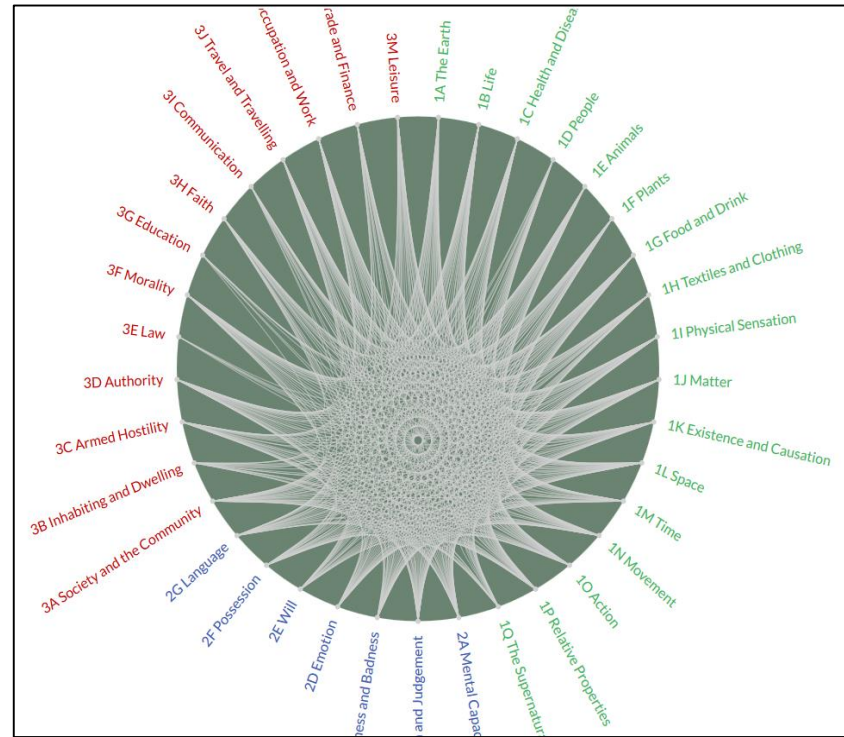
O, I just met you,
And this is **unrocked**,
But here’s my **number**,
So **clepe** me, **mayhap**!
It’s hard to look right
At you **pap-hawk**,
But here’s my **number**,
So **glew** me, **happen**!
And all the other **theows**,
Try to chase me,
But here’s my **sum-total**,
So **grede** me, **peradventure**!

<http://wordrobot.com/apps/textmetrics/home>

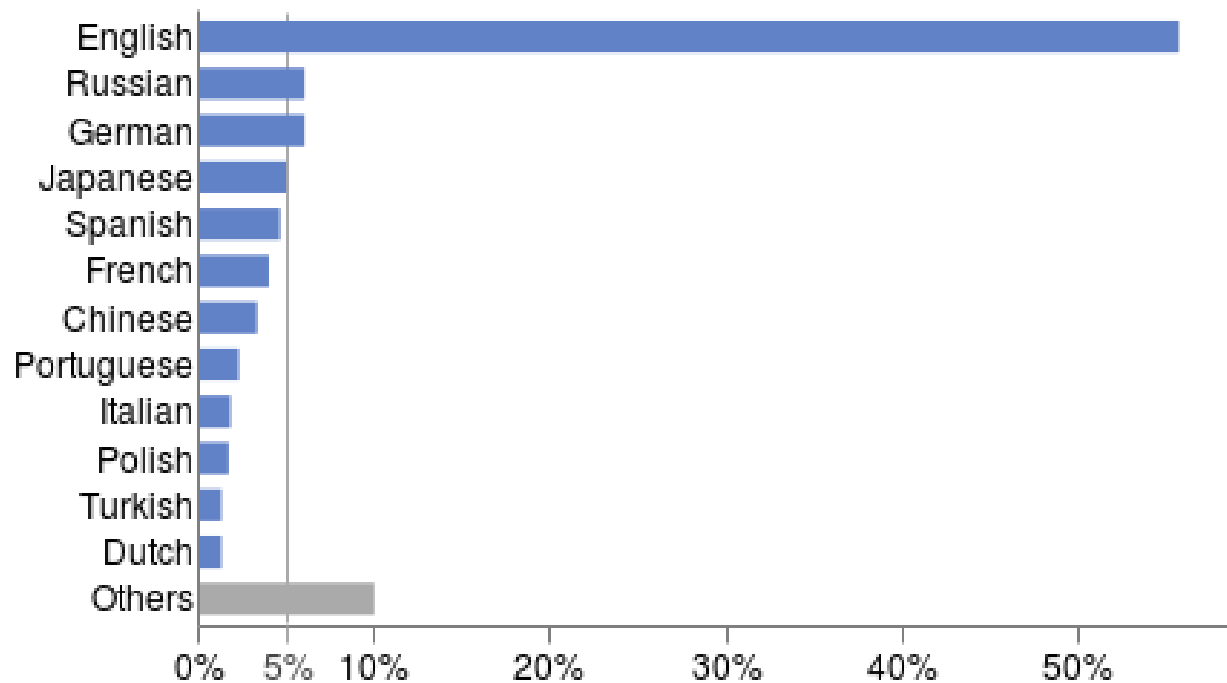
Username: demo / Password: XCZt9fLr

Example applications: mapping metaphor

A University of Glasgow and Arts and Humanities Research Council project that uses systematic patterns in the history of English to track the emergence of concepts and metaphorical associations at a cognitive and cultural level.



Most common content languages for websites



["Usage of content languages for websites"](#). W3Techs.com.

Retrieved 24 March 2015.

Oxford Global Languages

Aims and objectives

- Internet connectivity continues to grow
- But the internet is dominated by a few, large languages
- This creates a “Digital Divide”
- This is a global issue



The screenshot shows the Oxford Dictionaries website interface for isiZulu. At the top, there is a navigation bar with links for 'ABOUT', 'SIGN IN', and 'SITE LANGUAGE'. The main header features the Oxford Dictionaries logo and the text 'Be part of the living dictionary'. Below this is a search bar with the text 'Search the dictionary' and a search button. A blue button labeled 'ADD A WORD' is positioned below the search bar. The main content area is divided into three columns. The first column shows a plate of food with the text 'What is your favourite dish? Can you translate any of the ingredients?'. The second column shows a basket of fruit with the text 'This is a tricky fruit to translate. Do you agree? If you know the translation, submit it here.'. The third column shows a group of people looking at a tablet with the text 'What is the isiZulu Living Dictionary?'. At the bottom, there are three sections: 'Word of the Day' with the word 'umama', a Twitter feed with the text 'Ngabe yikuphi ukudla okuyintandokazi yakho? Ungabwazi ukubumusha noma...', and a red button labeled 'Do you need help getting started?'.

Google India Blog

News and Notes from Google India

Hindi Dictionary in Search

When you search for the meaning of a word in English, for instance “meaning of nostalgic”, you’ll get a dictionary straight in Google Search. Today, in collaboration with the [Oxford University Press](#), we’re bringing the Rajpal & Sons Hindi dictionary online. This new experience supports transliteration so you don’t even need to switch to a Hindi keyboard. So the next time when you’d like to know more about a word, say Nirdeshak, you can just type in Nirdeshak ka matlab in Search, and you’ll instantly get to see word meanings and dictionary definitions on the search results page, including English translations.

<https://india.googleblog.com/2017/04/bringing-down-language-barriers-making.html>



Nirdeshak ka matlab



सभी

समाचार

वीडियो

चित्र

मानचित्र

पुस्तकें

उड़ान

निर्देशक

लिप्यंतरित वर्शन: **nirdeshak**.

विशेषण

निर्देश देनेवाला।

nirdeshak का अनुवाद इस भाषा में करें:

अंग्रेज़ी



1. Director

फ़ीडबैक



अधिक परिभाषाएं और शब्द उद्गम

Meaning of निर्देशक (Nirdeshak) in English | निर्देशक का ...

shabdkosh.raftaar.in > Meaning-of-निर...



Google Research Blog

The latest news from Research at Google

A Large Corpus for Supervised Word-Sense Disambiguation

Wednesday, January 18, 2017

Posted by Colin Evans and Dayu Yuan, Software Engineers

Understanding the various meanings of a particular word in text is key to understanding language. For example, in the sentence *"he will receive stock in the reorganized company"*, we know that "stock" refers to *"the capital raised by a business or corporation through the issue and subscription of shares"* as defined in the [New Oxford American Dictionary](#) (NOAD), based on the context. However, there are more than 10 other definitions for "stock" in NOAD, ranging from "goods in a store" to "a medieval device for punishment". For a computer algorithm, distinguishing between these meanings is so difficult that it has been described as "AI-complete" in the past ([Navigli, 2009](#); [Ide and Veronis 1998](#); [Mallery 1988](#)).

In order to help further progress on this challenge, we're happy to announce the [release of word-sense annotations](#) on the popular [MASC](#) and [SemCor](#) datasets, manually annotated with senses from the NOAD. We're also releasing mappings from the NOAD senses to [English Wordnet](#), which is more commonly used by the research community. This is one of the largest releases of fully sense-annotated English corpora.

<https://research.googleblog.com/2017/01/a-large-corpus-for-supervised-word.html>

Prêt-à-LLOD

A European Union Initiative

- Prêt-à-LLOD aims to increase the uptake of language technologies by creating and exploiting ready-to-use multilingual Linguistic Linked Open Data (LLOD).
- The focus of the proposal creating a single digital market in Europe for multilingual NLP data.
 - Language technology specialists spend over 80% of their time on cleaning, organizing and collecting datasets
 - Few specialists take advantage of linked data technologies
- OUP will work on 2 pilot projects
 - **Pilot 1** covers sense linking for dictionaries, building on work done already
 - **Pilot 2** covers linking dictionaries to corpora

Prêt-à-LLOD

Who is involved?

National University of Ireland Galway	University of Zaragoza
Universidad Politécnica de Madrid	Universität Bielefeld
Goethe-Universität Frankfurt	DFKI GmbH
Semalytix	Oxford University Press
Semantic Web Company	Derilinx

High quality curated corpora

OXFORD
UNIVERSITY PRESS

Parallel Corpora

Major world language pairs, English to low-resource languages, and more available on request

Monolingual World Languages

Eg., English, French, Arabic, Portuguese, Hindi, etc

Low Resourced Languages

Eg., Indonesian, Bangla, Georgian, Hausa, Tamil, etc.

Domain Specific

Eg., Legal, Medical, Financial, etc.

Deep Domain

*Eg.,
Neuropharmacology;
Oncology; etc.*

OUP Academic Corpus Program

- Over 14m scholarly and general-interest books in English, all domain-classified in a 4-6 level taxonomy
 - » Eg., practitioner → medicine → brain science → neuropharmacology
- Full rights to exploit commercially
- Translation licensing program -- aligning parallel corpora
 - All professional human translations
 - Thousands of translated books into major world languages (+100m tokens, ontologically and taxonomically classified)
 - +200 titles (+15m tokens) translated into medium and low resource languages (Turkish; Indonesian; Georgian; etc.)

Dictionary Sense-Linking Project

- Creating a system for linking 2 dictionaries at the sense level.
That is, for generating links between the senses in 2 dictionaries whenever they have the same meaning.
- Dictionaries involved:
 - On one side, a **monolingual dictionary**
 - On the other, a **bilingual dictionary among the following**:
 - English – German, Spanish, Russian, Chinese, French, Italian
 - Linking through EN, the common language in each dictionary pair

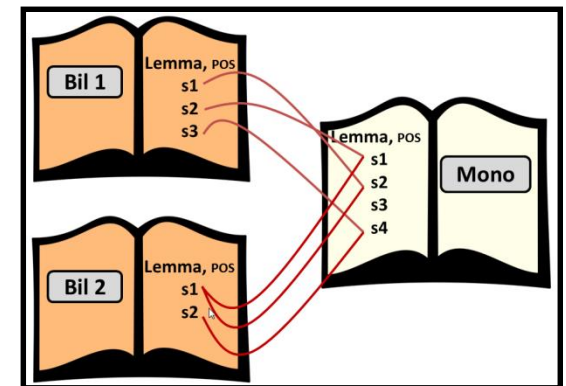
Strategy: a double-layer approach

Layer 1. Applying a standard Machine Learning binary classifier

Layer 2. Applying a meta-classifier on top

Results

- Precision: 91% (fraction of sense pairs correctly classified as sense links by the system)
- Recall: 80% (fraction of sense links the system was able to identify)



Current Development/Interest Areas

High quality
monolingual and
parallel corpora

Domain
mapping and
ontologies

Automated
multilingual
word-sense
linking

Training data
from OUP's
publishing
corpus

Tools derived
from OUP's rich
lexical data

Indian
languages

Collaboration Opportunities

Oxford brings...

- Deep lexical data experience; untapped content
- Collaborative relationships with all major Big Tech
- Power brand
- Global footprint

Oxford's looking for...

- Data and language tech software experience
- Great ideas for next-gen tools/services
- Open to partnering opportunities



Thank you

Casper Grathwohl
President, Dictionaries and Language Services, OUP



Oxford Dictionaries: Making an Impact