

Multilingual Lexicon Generation

Ahmad R. Shahid

Dimitar Kazakov

University of York, UK

Wikipedia

- Started in 2001
 - Attracting 684 million visitors yearly by 2008
- 262 languages
- 11,637,885 articles
 - 2,612,422 articles in English (~22%)
- 22 languages with over 100,000 articles
- 57 languages with over 10,000 articles

Wikipedia

- Dynamic in nature
 - Uses wiki for online editing, and hence the name Wikipedia
- Covers articles on every conceivable topic: politics, religion, economics, science, engineering etc.
- More than 75,000 active contributors

Art - Wikipedia, the free encyclopedia - Microsoft Internet Explorer

Address <http://en.wikipedia.org/wiki/Art> Go Links

Make a donation to Wikipedia and give the gift of knowledge! [Log in / create account](#)

[article](#) [discussion](#) [view source](#) [history](#)

Art

From Wikipedia, the free encyclopedia

navigation

- Main page
- Contents
- Featured content
- Current events
- Random article

search

Go Search

interaction

- About Wikipedia
- Community portal
- Recent changes
- Contact Wikipedia
- Donate to Wikipedia
- Help

tools

- What links here
- Related changes
- Upload file
- Special pages
- Permanent link
- Cite this page

languages

- العربية
- Aragonés
- Armãneashce
- Asturianu
- Azərbaycan
- Беларуская (тарашкевіца)
- Boarisch
- Bosanski
- Brezhoneg
- Български
- Català
- Česky

languages

- العربية
- Aragonés
- Armãneashce
- Asturianu
- Azərbaycan
- Беларуская (тарашкевіца)
- Boarisch
- Bosanski
- Brezhoneg
- Български
- Català
- Česky

Today

СРЕДА 15.05.2013 11:58



Barack Obama, President of the United States, speaking at a podium during a press conference in Washington, D.C., on May 15, 2013.

President Obama

His speech gives a sense of optimism for the Balkan region's progress.



President Obama's visit to the Balkans is a significant step towards strengthening ties with the region.

Balkan region's progress

Obama's visit is a sign of progress in the Balkan region.



The meeting was held in a formal setting, discussing regional issues.

Obama's visit

Obama's visit to the Balkans is a significant event.



The meeting was held in a formal setting, discussing regional issues.

Obama's visit

Obama's visit to the Balkans is a significant event.



The meeting was held in a formal setting, discussing regional issues.

Obama's visit

Obama's visit to the Balkans is a significant event.

Obama's visit

Obama's visit to the Balkans is a significant event.

Obama's visit

Obama's visit to the Balkans is a significant event.

Obama's visit

Obama's visit to the Balkans is a significant event.



The meeting was held in a formal setting, discussing regional issues.

Obama's visit

Obama's visit to the Balkans is a significant event.

Obama's visit

Obama's visit to the Balkans is a significant event.

Obama's visit

Obama's visit to the Balkans is a significant event.

Obama's visit

Obama's visit to the Balkans is a significant event.

Obama's visit

Obama's visit to the Balkans is a significant event.

- 1. Obama's visit to the Balkans is a significant event.
- 2. Obama's visit to the Balkans is a significant event.
- 3. Obama's visit to the Balkans is a significant event.
- 4. Obama's visit to the Balkans is a significant event.
- 5. Obama's visit to the Balkans is a significant event.
- 6. Obama's visit to the Balkans is a significant event.
- 7. Obama's visit to the Balkans is a significant event.

- 1. Obama's visit to the Balkans is a significant event.
- 2. Obama's visit to the Balkans is a significant event.
- 3. Obama's visit to the Balkans is a significant event.
- 4. Obama's visit to the Balkans is a significant event.
- 5. Obama's visit to the Balkans is a significant event.
- 6. Obama's visit to the Balkans is a significant event.
- 7. Obama's visit to the Balkans is a significant event.

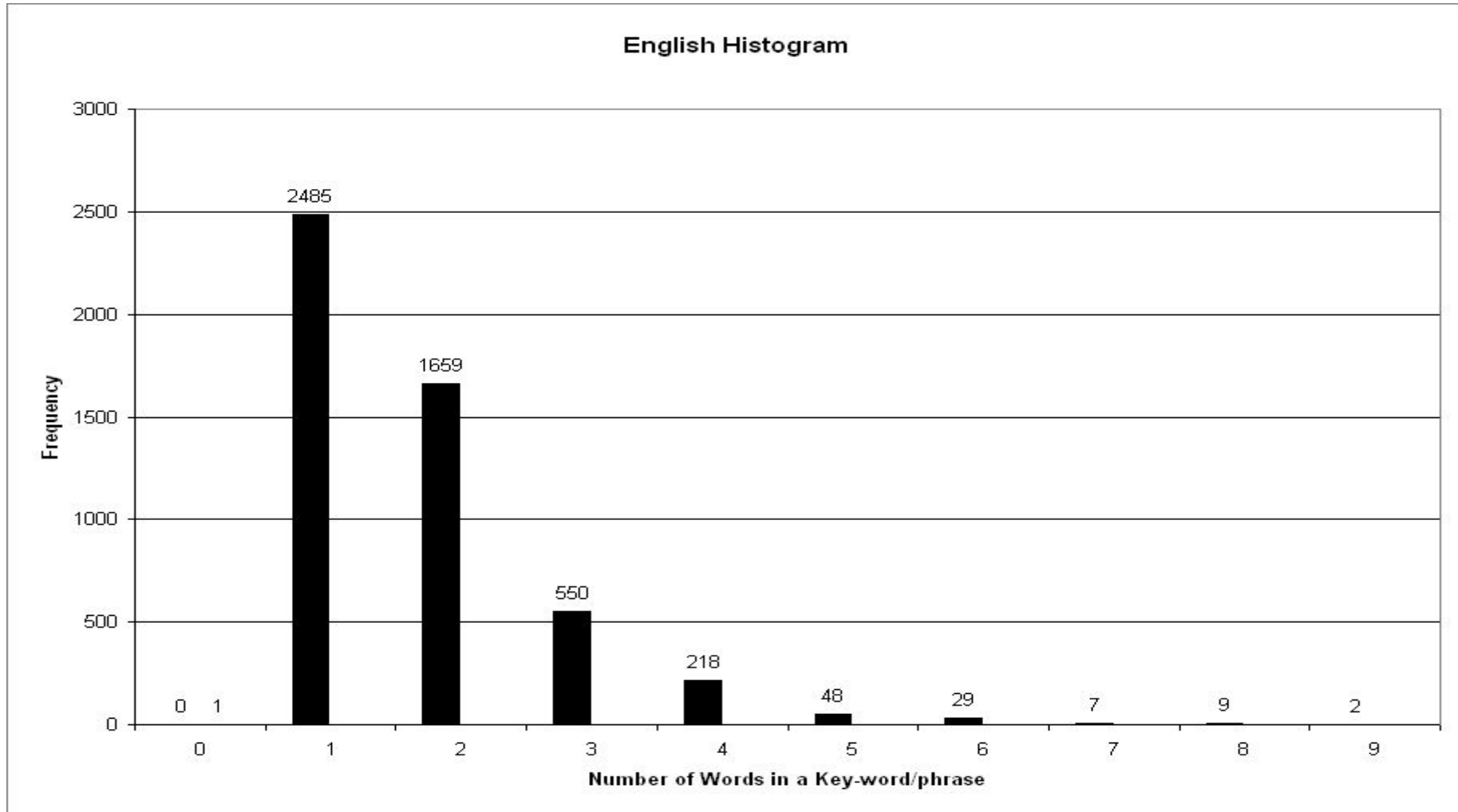
Creating the Lexicon

- User gives the starting point, a Wikipedia page on any topic of the user's liking
- The crawler crawls on that page, collecting interesting links
- Once reasonable number of links have been collected
 - Goes to the links one by one
 - Extracts the **title** of the page
 - Goes to articles in the target languages
 - Extracts **titles** in them as well
 - Puts them in the form of tuples, creating **entries** in the lexicon

Creating the General Dictionary

English	German	French	Polish	Bulgarian	Greek	Chinese
Wikipedia	Wikipedia	Wikipédia	Wikipedia	Уикипедия	Βικιπαίδεια	維基百科
Encyclopedia	Enzyklopädie	Encyclopédie	Encyklopedia	Енциклопедия	Εγκυκλοπαίδεια	百科全书
English language	Englische Sprache	Anglais	Język angielski	Английски език	Αγγλική γλώσσα	英语
Venice	Venedig	Venise	Wenecja	Венеция	Βενετία	威尼斯
Film director	Regisseur	Réalisateur	Reżyser	Режисьор	Σκηνοθέτης	電影導演
Uniform Resource Locator	Uniform Resource Locator	Uniform Resource Locator	Uniform Resource Locator	Унифициран локатор на ресурси	Uniform Resource Locator	統一資源定位符
Web search engine	Suchmaschine	Moteur de recherche	Wyszukiwarka internetowa	Търсачка	Μηχανή αναζήτησης	搜索引擎
University	Hochschule	Université	Uniwersytet	Университет	Πανεπιστήμιο	大學
Monopoly	Monopol	Monopole	Monopol	Монопол	Μονοπώλιο	壟斷
Computer	Computer	Ordinateur	Komputer	Компютър	Ηλεκτρονικός υπολογιστής	計算機
University of Oxford	University of Oxford	Université d'Oxford	Uniwersytet Oksfordzki	Оксфордски университет	Πανεπιστήμιο της Οξφόρδης	牛津大学
Population density	Bevölkerungsdichte	Densité de population	Gęstość zaludnienia	Гъстота на населението	Πυκνότητα πληθυσμού	人口密度
Presidential system	Präsidentielles Regierungssystem	Régime présidentiel	System prezydencki	Президентска република	Προεδρική Δημοκρατία	總統制
Dictatorship	Diktatur	Dictature	Dyktatura	Диктатура	Δικτατορία	專政
European Community	Europäische Gemeinschaft	Communauté européenne	Wspólnota Europejska	Европейска общност	Ευρωπαϊκή Κοινότητα	歐洲共同體
Benazir Bhutto	Benazir Bhutto	Benazir Bhutto	Benazir Bhutto	Беназир Бхуто	Μπιναζίρ Μπούτο	贝娜齐尔·布托
Thomas Edison	Thomas Alva Edison	Thomas Edison	Thomas Alva Edison	Томас Едисън	Τόμας Έντισον	托马斯·爱迪生
Art	Kunst	Art	Sztuka	Изкуство	Τέχνη	艺术
California	Kalifornien	Californie	Kalifornia	Калифорния	Καλιφόρνια	加利福尼亚州
Buddhism	Buddhismus	Bouddhisme	Buddyzm	Будизъм	Βουδισμός	佛教

The Histogram



Results

- Unigrams make the bulk of entries (~50%)
 - Followed by bigrams (~33%)
- More than 5,000 entries in the lexicon
 - Visited 726,715 English articles
- Only a little more than 1% of the total had corresponding pages in all the other six languages

Categories and Domain Specific Dictionaries

- Every article on Wikipedia belongs to one or more categories
 - For instance, the one on **Algorithm** belongs to categories:
 - Algorithms, Arabic words and phrases, Discrete mathematics, Mathematical logic, Theoretical computer science, Articles with example pseudocode

Categories in Wikipedia

- Wikipedia defines a hierarchy of categories
 - Each category may have many subcategories
 - Which in turn may have subcategories
- To create domain specific dictionaries categories could be used
 - Only those articles would be visited which belong to a set of short-listed categories
 - Starting point must be more specific

Computer Science Specific Dictionary

English	German	Japanese	Chinese	Arabic	Korean	Bulgarian	Thai	Greek	Urdu
Computer science	Informatik	計算機科学	计算机科学	معلوماتية	전산학	Информатика	วิทยาการคอมพิวเตอร์	Επιστήμη υπολογιστών	null
Computer	Computer	コンピュータ	计算机	حاسوب	컴퓨터	Компютър	คอมพิวเตอร์	Ηλεκτρονικός υπολογιστής	سمارتده
Programming language	Programmiersprache	プログラミング言語	编程语言	لغة برمجة	프로그래밍 언어	Език за програмиране	ภาษาโปรแกรม	Γλώσσα προγραμματισμού	null
Internet	Internet	インターネット	互联网	إنترنت	인터넷	Интернет	อินเทอร์เน็ต	Διαδίκτυο	جالين
Alan Turing	Alan Turing	アラン・チューリング	艾伦·图灵	ألان تورنج	앨런 튜링	Альн Тюринг	แอลัน ทัวริง	Άλαν Τούρινγκ	null
Operating system	Betriebssystem	オペレーティングシステム	操作系统	نظام تشغيل	운영 체제	Операциона система	ระบบปฏิบัติการ	Λειτουργικό σύστημα	عملیاتی نظام
Binary numeral system	Dualsystem	二進法	二进制	نظام عد ثنائي	이진법	Двоична бройна система	เลขฐานสอง	Διαδικό σύστημα	null
Fortran	Fortran	FORTRAN	Fortran	فورتران	포트란	FORTRAN	ภาษาฟอร์แทรน	Fortran	null
Web browser	Webbrowser	ウェブブラウザ	网页浏览器	متصفح وب	웹 브라우저	Уеббраузер	เว็บเบราว์เซอร์	Web browser	متصفح جال
Computer software	Software	ソフトウェア	軟件	برمجيات	컴퓨터 소프트웨어	Програмно осигуряване	ซอฟต์แวร์	Λογισμικό	null

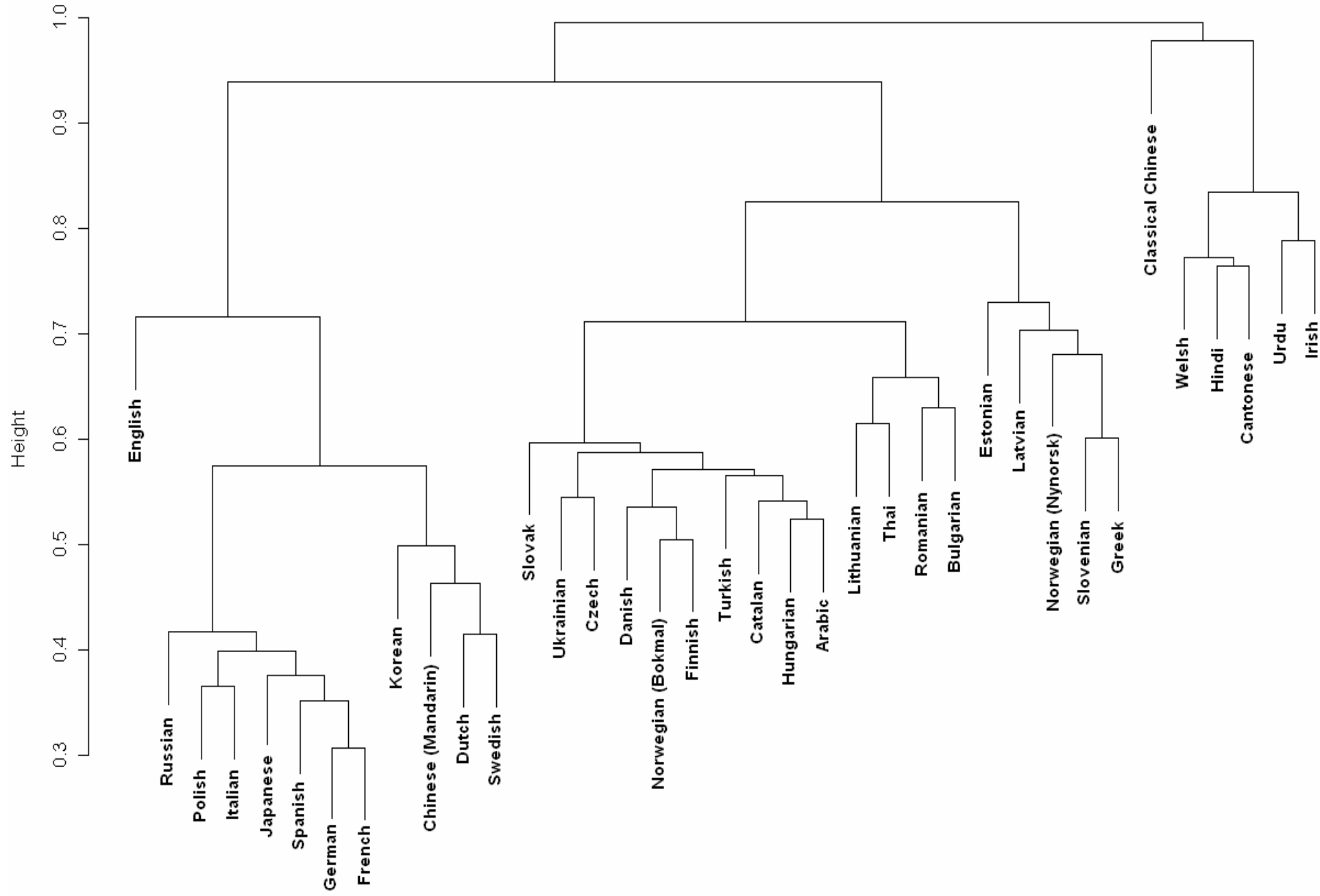
CS specific dictionary

- The result is a 37-language, 2,500 entry dictionary
- The list of categories

Similarities and relationships between languages

- Similarity between any two languages is defined in terms of the ratio of number of common entries in both the languages to the total number of entries in both
 - Similarity defines if any two languages are related
 - Linguistically related languages and the ones sharing the cultural background are more likely to have common entries in the lexicon

Cluster Dendrogram



Language relationships

- There may be some relationship between geographical distance
 - French and German
- and overlap, as well as linguistic relatedness
 - Slovak-Ukrainian-Czech, Turkish-Finnish-Hungarian.
- But it is far from clear.

Category Translations

- Very useful as a teaching resource (specially for undergraduate students)
 - Students know the concept in their own language, but may not know its English translation
 - Even when the English term is known, it may be easier to memorise it if it can be linked to items studied in one's mother tongue.

Category Translations

- It would be useful for students who return home after obtaining their degrees in a language different from their mother tongue and have to communicate in their own language once at home.
- The aim is to provide translations of some key concepts.

Categories Computer Science

A

- [\[+\] Algorithms](#) (47)
- [\[+\] Artificial intelligence](#) (29)
- [\[+\] Computer science awards](#) (4)

C

- [\[+\] Cellular automata](#) (5)
- [\[+\] Computer science competitions](#) (2)
- [\[+\] Computational science](#) (15)
- [\[+\] Computer architecture](#) (16)
- [\[+\] Computer programming](#) (23)
- [\[+\] Concurrency](#) (6)

D

- [\[+\] Data structures](#) (12)
- [\[+\] Databases](#) (17)

E

- [\[+\] Computer science education](#) (1)
- [\[+\] Events \(computing\)](#) (0)

G

- [\[+\] Computer graphics](#) (28)

H

- [\[+\] Human-computer interaction](#) (15)

L

- [\[+\] Computer science lists](#) (0)
- [\[+\] Computer science literature](#) (3)

M

- [\[+\] Mathematical optimization](#) (5)

O

- [\[+\] Operating systems](#) (37)

O cont.

- [\[+\] Computer science organizations](#) (8)

P

- [\[+\] Programming languages](#) (42)

S

- [\[+\] Computer scientists](#) (25)
- [\[+\] Computer security](#) (17)
- [\[+\] Software engineering](#) (28)

T

- [\[+\] Theoretical computer science](#) (16)

W

- [\[+\] Computer science websites](#) (0)

μ

- [\[+\] Computer science stubs](#) (6)

Pages in category "Computer science"

The following 19 pages are in this category, out of 19 total. This list may sometimes be slightly out of date ([learn more](#))

- [Computer science](#)
- [List of computer science fields](#)
- *
 - [Topic outline of computer science](#)
 - [Portal:Computer science](#)

A

- [ACM Computing Classification System](#)
- [Adaptive educational hypermedia](#)

A cont.

- [Adaptive hypermedia](#)
- [Authoring of adaptive hypermedia](#)

C

- [History of computer science](#)
- [Computer scientist](#)

E

- [Empirical modelling](#)

H

- [Charles Leonard Hamblin](#)

I

- [Informatics](#)

I cont.

- [Information and Computer Science](#)

M

- [MALINTENT](#)

O

- [Overlapping subproblem](#)

P

- [Program \(mathematical object\)](#)

S

- [Klaus Samelson](#)

U

- [User talk:Spychiehalla](#)

Categories Artificial Intelligence

A

- [\[+\] Artificial intelligence applications](#) (4)
- [\[+\] Artificial immune systems](#) (0)
- [\[+\] Artificial intelligence associations](#) (0)
- [\[+\] Automated planning and scheduling](#) (0)

C

- [\[+\] Cognitive architecture](#) (0)
- [\[+\] Computer vision](#) (12)
- [\[+\] Artificial intelligence conferences](#) (0)
- [\[+\] Constraint satisfaction](#) (0)

E

- [\[+\] Expert systems](#) (1)

F

- [\[+\] Artificial intelligence in fiction](#) (4)

G

- [\[+\] Game artificial intelligence](#) (6)

H

- [\[+\] History of artificial intelligence](#) (0)

K

- [\[+\] Knowledge engineering](#) (0)
- [\[+\] Knowledge representation](#) (16)

L

- [\[+\] Artificial intelligence laboratories](#) (0)
- [\[+\] Logic programming](#) (3)

M

- [\[+\] Machine learning](#) (6)
- [\[+\] Multi-agent systems](#) (1)

N

- [\[+\] Natural language processing](#) (6)

O

- [\[+\] Ontology \(computer science\)](#) (2)
- [\[+\] Optimization algorithms](#) (3)

P

- [\[+\] Philosophy of artificial intelligence](#) (1)
- [\[+\] Artificial intelligence publications](#) (1)

R

- [\[+\] Artificial intelligence researchers](#) (2)
- [\[+\] Robotics](#) (16)
- [\[+\] Rule engines](#) (1)

S

- [\[+\] Search algorithms](#) (2)

T

- [\[+\] Turing tests](#) (0)

U

- [\[+\] Artificial intelligence stubs](#) (0)

Pages in category "Artificial intelligence"

The following 139 pages are in this category, out of 139 total. This list may sometimes be slightly out of date ([learn more](#))

- [Portal:Artificial intelligence](#)
- [Topic outline of artificial intelligence](#)

2

- [20Q](#)

A

- [AI-complete](#)
- [AIML](#)
- [ASR-complete](#)
- [Action selection](#)
- [Admissible heuristic](#)
- [Affective computing](#)
- [Agent Systems Reference Model](#)
- [AgentSheets](#)
- [Anticipation \(artificial intelligence\)](#)
- [Anytime algorithm](#)

C cont.

- [Computational intelligence](#)
- [Computer Audition](#)
- [Computer vision](#)
- [Computer-assisted proof](#)
- [Connectionist expert system](#)
- [Constructionist design methodology](#)

D

- [Darwin Among the Machines](#)
- [Darwin machine](#)
- [Data pack](#)
- [Decision lists](#)
- [Decision-tree pruning](#)
- [Diagnosis \(artificial intelligence\)](#)
- [Discovery system](#)
- [Dynamic time warping](#)

M cont.

- [Mark Stephen Meadows](#)
- [Means-ends analysis](#)
- [MindRACES](#)
- [Mindpixel](#)
- [Model-based reasoning](#)
- [Moravec's paradox](#)
- [Morphological computation](#)

N

- [Neats vs. scruffies](#)
- [Neural modeling fields](#)
- [Neuro-fuzzy](#)
- [Nouvelle AI](#)

O

- [Ontology learning](#)

Categories Translation

- CS is much more general than AI
 - Subcategories go many levels deep
- 2,000 CS related categories were extracted
- 450 AI related categories were extracted
 - Leaf nodes were also considered on the first page

CS Categories

English	German	Japanese	Chinese	Arabic	Korean	Bulgarian	Thai	Greek
Computer science	Informatik	計算機科学	计算机科学	حوسبة	전산학	Информатика	วิทยาการคอมพิวเตอร์	Επιστήμη υπολογιστών
Computer architecture	Rechnerarchitektur	コンピュータアーキテクチャ	電腦架構	null	컴퓨터 구조	null	null	null
Semantics	Semantik	意味論	null	null	의미론	Семантика	null	null
Algorithms	Algorithmus	アルゴリズム	算法	خوارزميات	알고리즘	Алгоритми	ขั้นตอนวิธี	Αλγόριθμοι
Artificial intelligence	Künstliche Intelligenz	人工知能	人工智能	ذكاء اصطناعي	인공지능	null	ปัญญาประดิษฐ์	Τεχνητή νοημοσύνη
Computer programming	Programmierung	プログラミング	程序设计	برمجة	컴퓨터 프로그래밍	null	การเขียนโปรแกรม	null
Operating systems	Betriebssystem	オペレーティングシステム	操作系统	نظم تشغيل	운영 체제	null	ระบบปฏิบัติการ	null
Programming languages	Programmiersprache	プログラミング言語	程序设计语言	لغات برمجة	프로그래밍 언어	Езици за програмиране	ภาษาโปรแกรม	Γλώσσες προγραμματισμού
Linux	Linux	Linux	Linux	لينكس	리눅스	ГНУ/Линукс	null	null
Cryptography	Kryptologie	暗号技術	密码学	تعمية	암호학	Криптография	null	Κρυπτογραφία

AI Categories

English	German	French	Japanese	Chinese	Arabic	Korean	Bulgarian	Thai	Greek
Artificial intelligence	Künstliche Intelligenz	Intelligence artificielle	人工知能	人工智能	ذكاء اصطناعي	인공지능	Изкуствен интелект	ปัญญาประดิษฐ์	Τεχνητή νοημοσύνη
Chess	Schach	Échecs	チェス	国际象棋	شطرنج	체스	Шахмат	null	Σκάκι
Game theory	Spieltheorie	Théorie des jeux	ゲーム理論	博弈论	نظرية الألعاب	게임 이론	Теория на игрите	ทฤษฎีเกม	Θεωρία παιγνίων
Search algorithms	null	Algorithme de recherche	検索アルゴリズム	搜尋演算法	null	검색 알고리즘	null	null	null
Machine learning	Maschinelles Lernen	null	機械学習	机器学习	تعلم آلي	기계 학습	null	การเรียนรู้ของเครื่อง	null
Robotics	Robotik	Robotique	ロボット工学	机器人学	null	로봇공학	Роботика	null	null
Computer vision	Maschinelles Sehen	Vision par ordinateur	コンピュータビジョン	计算机视觉	رؤية حاسوبية	null	null	คอมพิวเตอร์วิทัศน์	null
Fuzzy logic	Fuzzy-Logik	Logique floue	フuzzy論理	模糊逻辑	منطق ضبابي	퍼지	null	ตรรกศาสตร์คลุมเครือ	null
Taxonomy	Taxonomie	Taxinomie	分類学	生物分類學	null	분류학	Таксономия	อนุกรมวิธาน	Ταξινόμια

Bibliography

- Alexander E.R. and Patrick S., Mining Wiki Resources for Multilingual Named Entity Recognition, *Proceedings of ACL-08: HLT*. (2008).
- Christian B., Mathieu M. and Gilles S., The PAPHILLON Project: Cooperatively Building a Multilingual Lexical Database to Derive Open Source Dictionaries & Lexicons, *Proceedings of the 2nd Workshop NLPXML 2002, Post COLING 2002 Workshop*. (2002).
- James B., JMdict: a Japanese-Multilingual Dictionary, *Coling 2004 Workshop on Multilingual Linguistic Resources*. (2004).
- Mathieu L., Multilingual Dictionary Construction and Services Case Study with the Fe* Projects, *Proceedings of PACLING'97*. (1997).
- Shahid, A. and Kazakov, D. (2009). Automatic Multilingual Lexicon Generation using Wikipedia as a Resource, *International Conference on Agents and Artificial Intelligence*, Portugal (to appear).