The Challenges of Multilingual Search

Paul Clough

The Information School
University of Sheffield

ISKO UK conference 8-9 July 2013

Outline

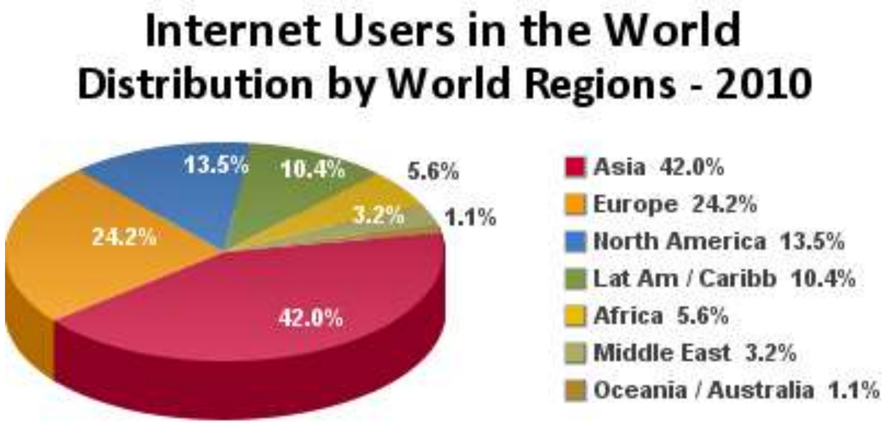
- What is multilingual search?
- MLIR and CLIR techniques
- Four challenges
    - Thinking beyond search
    - Translation and language resources
    - Providing effective user support
    - Going from research to practice

The “Grand [search] Challenge”

*“Given a query in any medium and **any language**, select relevant items from a **multilingual** multimedia collection which can be in any medium and **any language**, and present them in the style or order most likely to be useful to the querier, with identical or near identical objects in **different** media or **languages** appropriately identified.”*

Douglas W. Oard and David Hull, AAAI Symposium on Cross-Language IR, Spring 1997, Stanford, USA

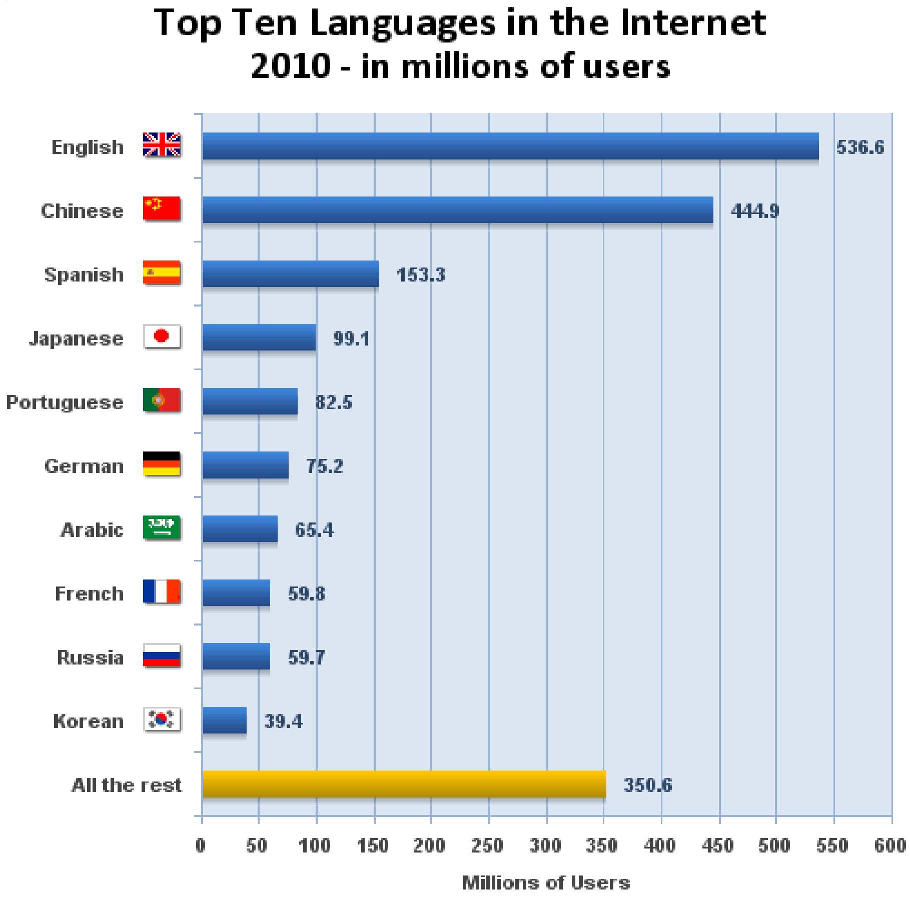
The need for multilingual search



Internet Users in the World
Distribution by World Regions - 2010

Source: Internet World Stats - www.internetworldstats.com/stats.htm
Basis: 1,966,514,816 Internet users on June 30, 2010
Copyright © 2010, Miniwatts Marketing Group

http://www.internetworldstats.com/stats.htm



Top Ten Languages in the Internet
2010 - in millions of users

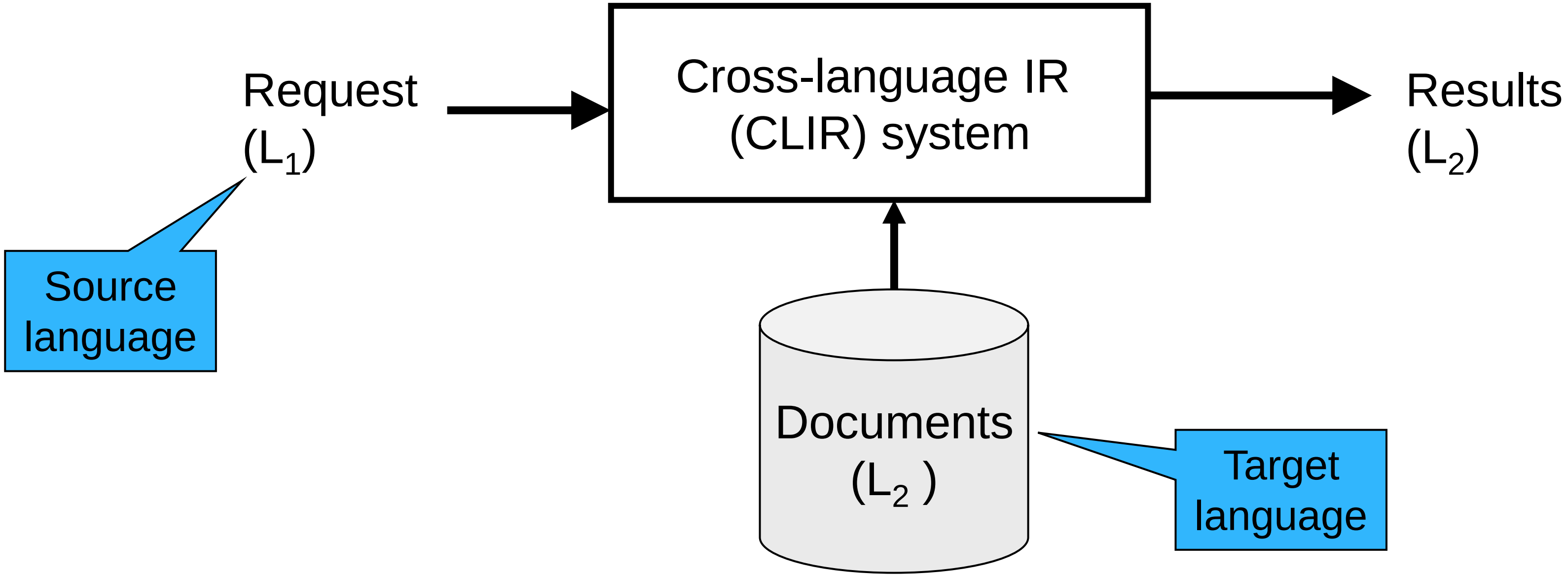
Source: Internet World Stats - www.internetworldstats.com/stats7.htm
Estimated Internet users are 1,966,514,816 on June 30, 2010
Copyright © 2000 - 2010, Miniwatts Marketing Group

What is multilingual search?

- **Multilingual Information Access (MLIA)**
  - Accessing, querying and retrieving information from collections in any language
  - Includes search and browse functionalities
- **Cross-Language Information Retrieval (CLIR)**
  - Querying multilingual collections in one language in order to retrieve documents in other languages
  - Bilingual information retrieval
- **Multilingual Information Retrieval (MLIR)**
  - Process information (queries, documents, both) in multiple languages
  - Includes cross-language retrieval

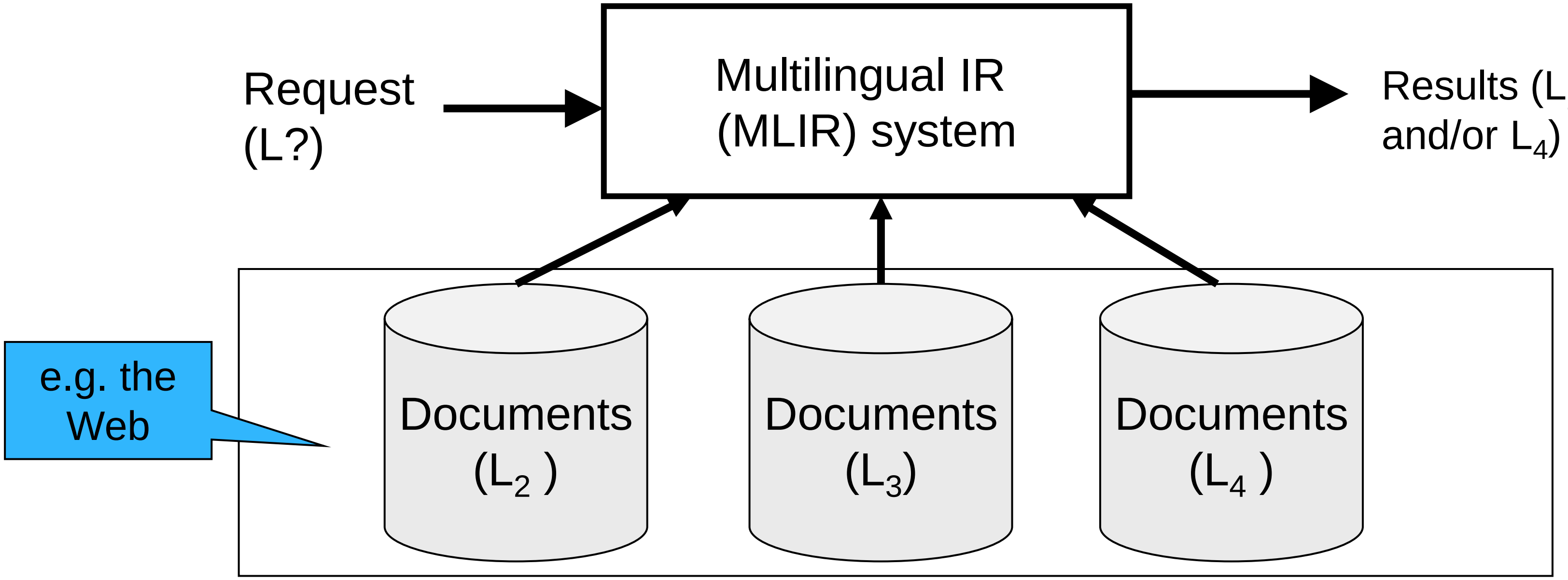
Cross-language search

- Documents and user requests are in different languages (**bilingual retrieval**)

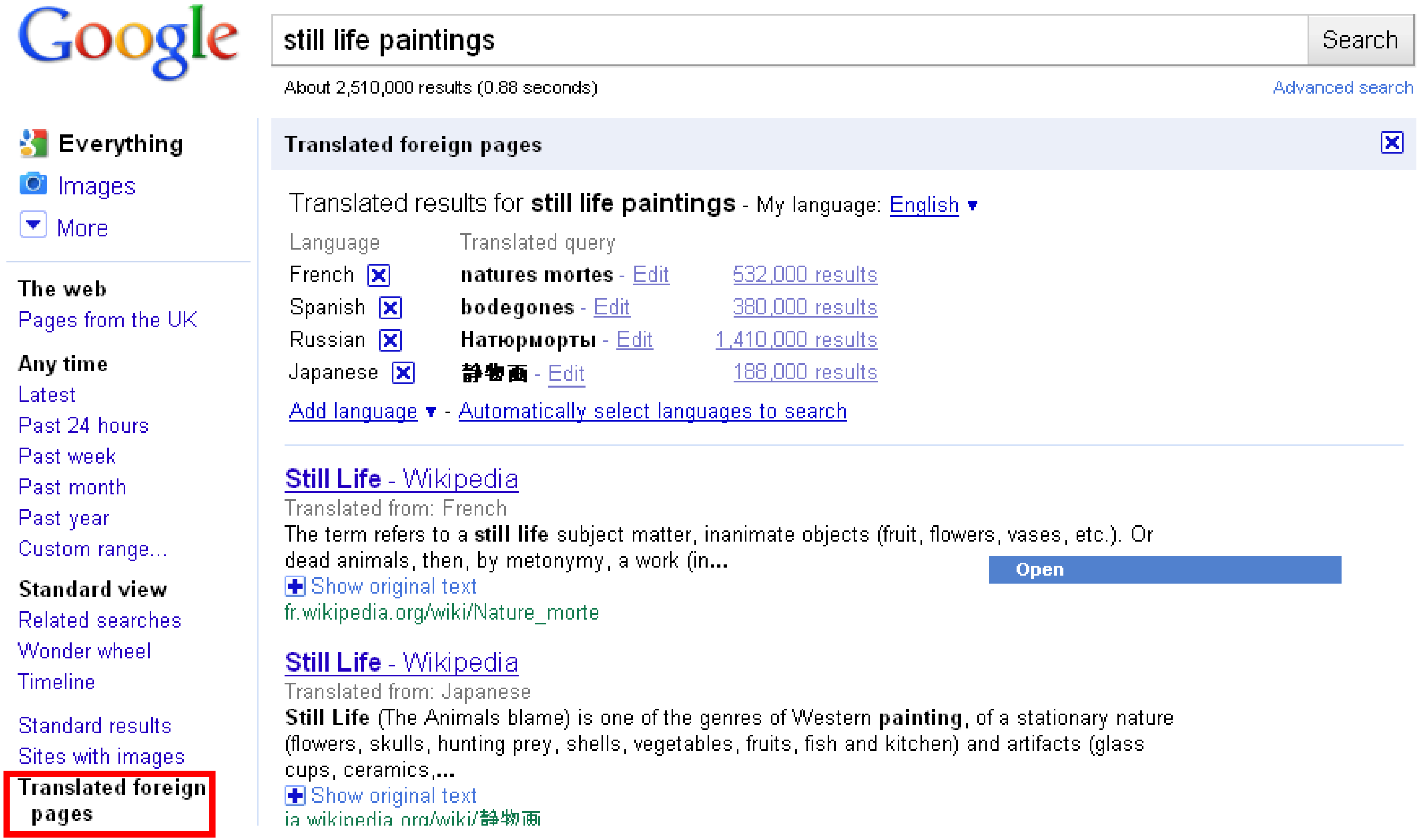


Multilingual search

- Documents in collection in different languages, search requests in any language



Google web search



still life paintings

Search

About 2,510,000 results (0.88 seconds)

Advanced search

- Everything
- Images
- More

The web

Pages from the UK

Any time

- Latest
- Past 24 hours
- Past week
- Past month
- Past year
- Custom range...

Standard view

- Related searches
- Wonder wheel
- Timeline

- Standard results
- Sites with images
- **Translated foreign pages**

Translated foreign pages ☒

Translated results for **still life paintings** - My language: English ▾

Language	Translated query	
French ☒	**natures mortes** - Edit	532,000 results
Spanish ☒	**bodegones** - Edit	380,000 results
Russian ☒	**Натюрморты** - Edit	1,410,000 results
Japanese ☒	**静物画** - Edit	188,000 results

Add language ▾ - Automatically select languages to search

Still Life - Wikipedia

Translated from: French

The term refers to a **still life** subject matter, inanimate objects (fruit, flowers, vases, etc.). Or dead animals, then, by metonymy, a work (in**...**

Open

+ Show original text

fr.wikipedia.org/wiki/Nature_morte

Still Life - Wikipedia

Translated from: Japanese

Still Life (The Animals blame) is one of the genres of Western **painting**, of a stationary nature (flowers, skulls, hunting prey, shells, vegetables, fruits, fish and kitchen) and artifacts (glass cups, ceramics,**...**

+ Show original text

ja.wikipedia.org/wiki/静物画

Obvious question …

… "Why do users want to retrieve documents they presumably can't read?"

- Some users are **multilingual**
  - Can formulate searches and judge relevance in many languages
  - Want the convenience of a single query
- Some users are **monolingual**
  - Want to query in their native language
  - Can judge relevance even if results not translated
  - Have access to document translation
  - Objects retrieved are language-independent (e.g. images)

Matching queries and documents

- MLIR and CLIR involves matching queries in one language with documents in another
  - Translate the query (**query translation**)
  - Translate the documents (**document translation**)
  - Translate queries and documents (e.g. into an intermediate language)
  - No translation
- MLIR may also involve a **merging** step

Translation resources

- The use of machine-readable, human-crafted **bilingual dictionaries** (and thesauri, word-lists and other similar resources)
- The use of translation resources generated by statistical approaches from suitable training data
  - Require parallel or comparable corpora
- The use of a ‘full’ **machine translation** system
  - Commercial or open source

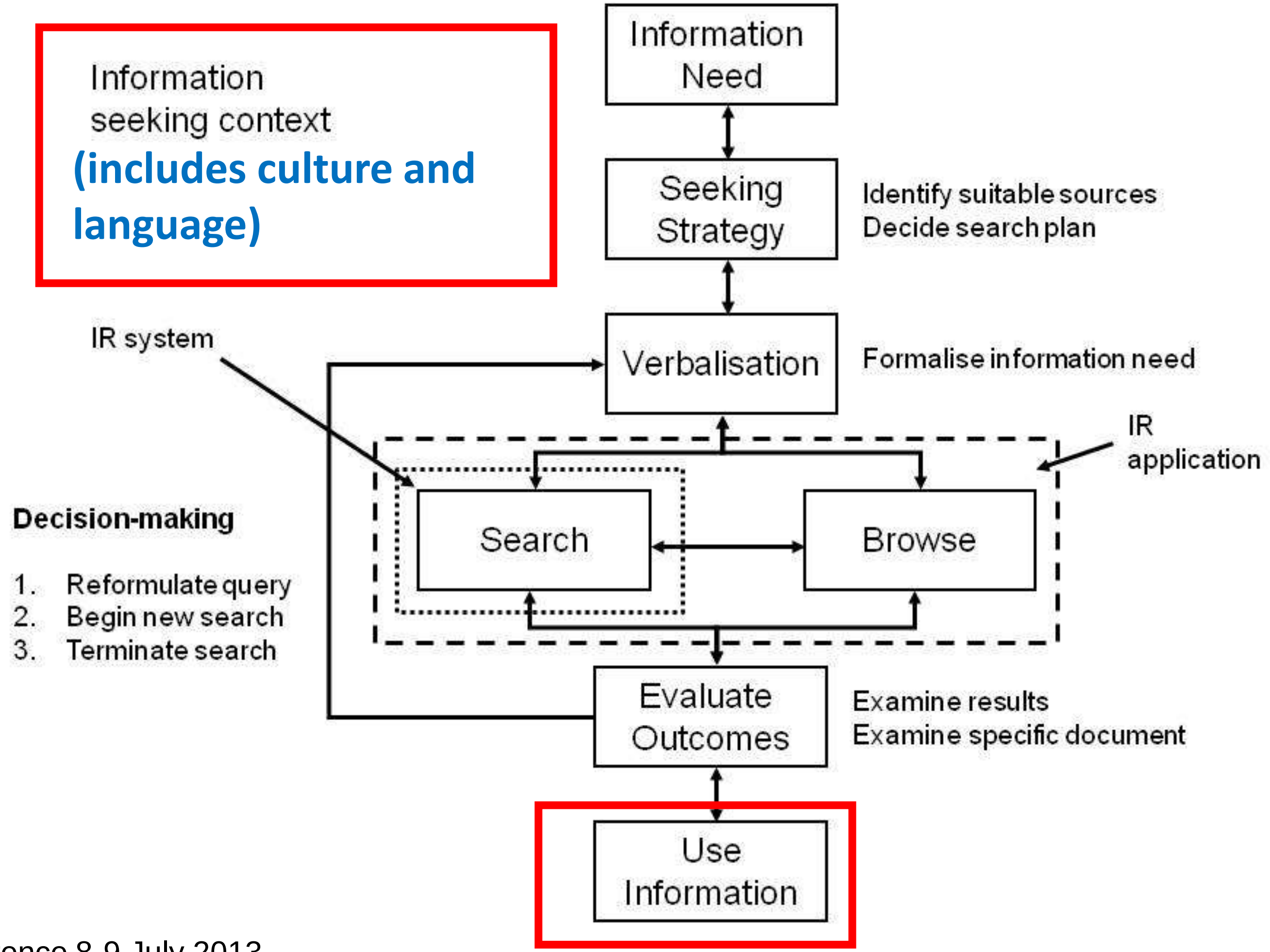
Does CLIR/MLIR work?

- In the lab CLIR/MLIR using various approaches and translation resources is effective
- Results from the Cross-Language Evaluation Forum (CLEF) have shown steady increases
  - **1997** 50-60% of monolingual baseline for various tasks
  - **2003** 80% onwards for most commonly used language pairs
  - **2009** up to 99% of monolingual baseline
- Main findings
  - Well-tuned MT systems are highly effective
  - Combinations of approaches work well

Challenges for multilingual search

1. Thinking beyond search

Search in the broader context



Developing global applications

"Globalization is the process of making applications work seamlessly **utilizing the user's preferred language and culture**. It involves not just programming and deployment skills but also cultural, translation and language expertise as well."

http://www-01.ibm.com/software/globalization/topics/index.html

Multilingual search will typically be part of a wider globalization process and concerned with developing multilingual applications (or websites) that include search functions (i.e. search is not the end goal)

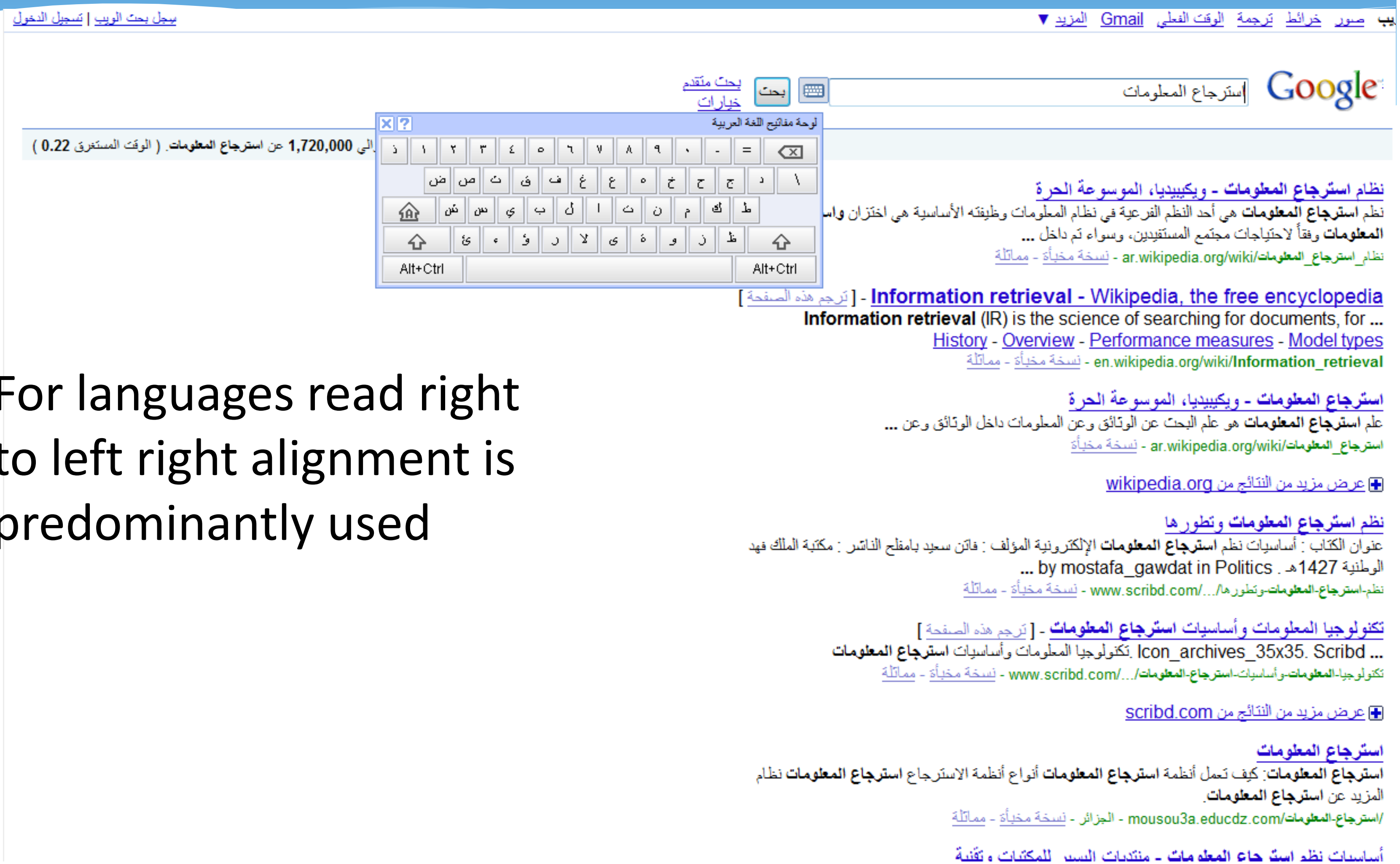
The impact of culture

- Culture is the behaviour typical of a group of people
  - Members of the same culture are likely to have the same knowledge of certain things and would think and act similarly in certain situations
- Aspects to consider (for design) include
  - Religion, customs, colours, metaphors, icons and flags, and **language**
- Crossing the cultural boundary includes adapting to a given market’s cultural conventions (**localisation**)

Aspects to consider for localisation

- **Translation** of the product' s interface and documentation
- **Colours, images, graphics and icons** adapting to cultural and legal requirements
- **Rendering** displaying text correctly (e.g. does the new text fit inside the allocated space?)
- **Fonts** making sure the correct fonts and characters for the target language
- **Bi-directional text** needed in Arabic and other languages
- **Locale data** how to display dates, time, number, currency and other regional data

Formatting and page layout



ويب صور خرائط ترجمة الوقت الفعلي Gmail المزيد ▼
سجل بحث الويب | تسجيل الدخول
Google
بحث
بحث متقدم
خيارات
لوحة مفاتيح اللغة العربية
ذ ١ ٢ ٣ ٤ ٥ ٦ ٧ ٨ ٩ ٠ - = ⌫
ض ص ث ق ف غ ع ه خ ح ج د \
ش س ي ب ل ا ت ن م ك ط
ئ ء ؤ ر لا ى ة و ز ظ
Alt+Ctrl
Alt+Ctrl

استرجاع المعلومات

إلى 1,720,000 عن **استرجاع المعلومات**. (الوقت المستغرق 0.22)

نظام استرجاع المعلومات - ويكيبيديا، الموسوعة الحرة
نظم **استرجاع المعلومات** هي أحد النظم الفرعية في نظام المعلومات وظيفته الأساسية هي اختزان **واس**
المعلومات وفقاً لاحتياجات مجتمع المستفيدين، وسواء تم داخل **...**
ar.wikipedia.org/wiki/نظام_**استرجاع_المعلومات** - نسخة مخبأة - مماثلة

Information retrieval - Wikipedia, the free encyclopedia - [ترجم هذه الصفحة]
Information retrieval (IR) is the science of searching for documents, for **...**
History - Overview - Performance measures - Model types
en.wikipedia.org/wiki/**Information_retrieval** - نسخة مخبأة - مماثلة

استرجاع المعلومات - ويكيبيديا، الموسوعة الحرة
علم **استرجاع المعلومات** هو علم البحث عن الوثائق وعن المعلومات داخل الوثائق وعن **...**
ar.wikipedia.org/wiki/**استرجاع_المعلومات** - نسخة مخبأة

عرض مزيد من النتائج من wikipedia.org

نظم استرجاع المعلومات وتطورها
عنوان الكتاب : أساسيات نظم **استرجاع المعلومات** الإلكترونية المؤلف : فاتن سعيد بامفلح الناشر : مكتبة الملك فهد
الوطنية 1427هـ . by mostafa_gawdat in Politics **...**
www.scribd.com/.../نظم-**استرجاع-المعلومات**-وتطورها - نسخة مخبأة - مماثلة

تكنولوجيا المعلومات وأساسيات استرجاع المعلومات - [ترجم هذه الصفحة]
... Icon_archives_35x35. Scribd .تكنولوجيا المعلومات وأساسيات **استرجاع المعلومات**
www.scribd.com/.../تكنولوجيا-**المعلومات**-وأساسيات-**استرجاع-المعلومات** - نسخة مخبأة - مماثلة

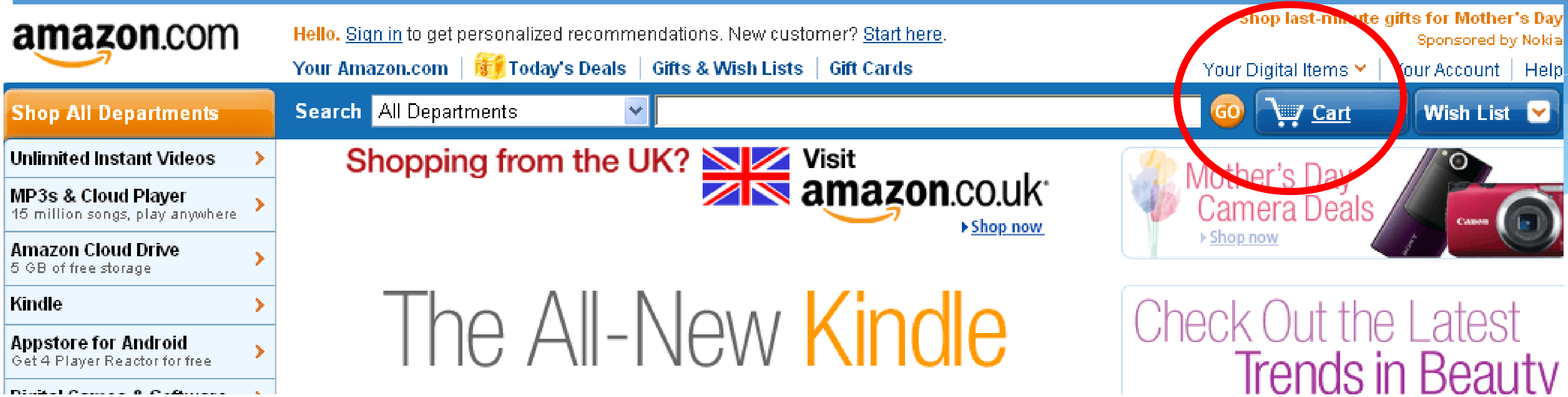
عرض مزيد من النتائج من scribd.com

استرجاع المعلومات
استرجاع المعلومات: كيف تعمل أنظمة **استرجاع المعلومات** أنواع أنظمة الاسترجاع **استرجاع المعلومات** نظام
المزيد عن **استرجاع المعلومات**.
mousou3a.educdz.com/**استرجاع-المعلومات**/ - الجزائر - نسخة مخبأة - مماثلة

أساسيات نظم استرجاع المعلومات - منتديات اليسير للمكتبات و تقنية

For languages read right to left right alignment is predominantly used

Differences in terminology



amazon.com

Hello. Sign in to get personalized recommendations. New customer? Start here.

Your Amazon.com | Today's Deals | Gifts & Wish Lists | Gift Cards

Shop last-minute gifts for Mother's Day
Sponsored by Nokia

Your Digital Items | Your Account | Help

Shop All Departments | Search | All Departments | GO | Cart | Wish List

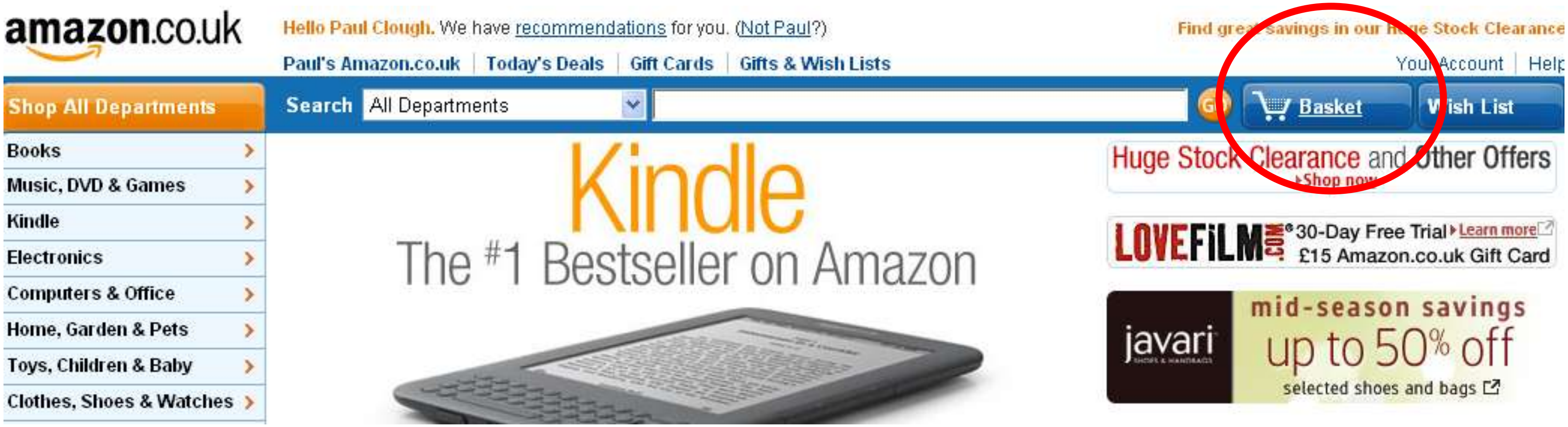
- Unlimited Instant Videos
- MP3s & Cloud Player
  15 million songs, play anywhere
- Amazon Cloud Drive
  5 GB of free storage
- Kindle
- Appstore for Android
  Get 4 Player Reactor for free
- Digital Games & Software

Shopping from the UK? Visit amazon.co.uk
Shop now

Mother's Day Camera Deals
Shop now

The All-New Kindle

Check Out the Latest Trends in Beauty



amazon.co.uk

Hello Paul Clough. We have recommendations for you. (Not Paul?)

Paul's Amazon.co.uk | Today's Deals | Gift Cards | Gifts & Wish Lists

Find great savings in our Huge Stock Clearance

Your Account | Help

Shop All Departments | Search | All Departments | GO | Basket | Wish List

- Books
- Music, DVD & Games
- Kindle
- Electronics
- Computers & Office
- Home, Garden & Pets
- Toys, Children & Baby
- Clothes, Shoes & Watches

Kindle
The #1 Bestseller on Amazon

Huge Stock Clearance and Other Offers
Shop now

LOVEFiLM.com 30-Day Free Trial Learn more
£15 Amazon.co.uk Gift Card

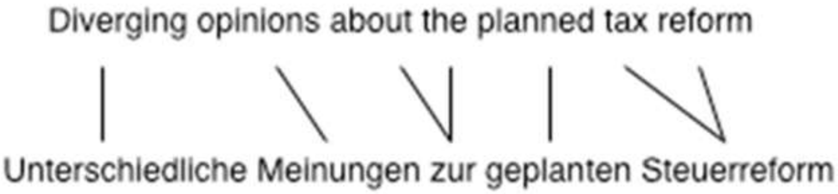
javari
mid-season savings
up to 50% off
selected shoes and bags

Challenges for multilingual search

2. Languages and translation

What is translation?

- Isn’t translation just replacing words in one language (source) with words in another language (target)?



Diverging opinions about the planned tax reform

Unterschiedliche Meinungen zur geplanten Steuerreform

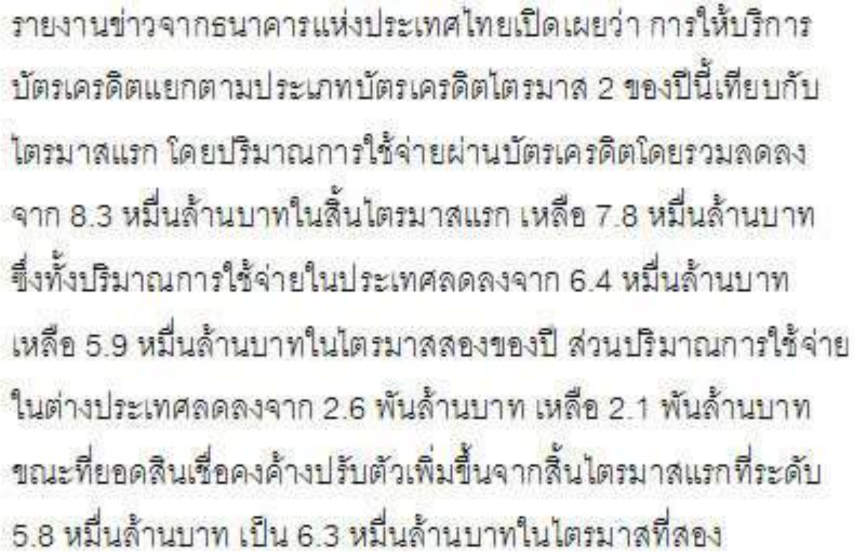
- Not quite so simple …
  - Often includes changes in syntax
  - Often includes translation of semantics and use of culturally-specific terms or concepts

Further translation challenges

- Out-Of-Vocabulary (OOV) terms
- Ambiguity (source and target languages)
  - Lexical ambiguity (e.g. bat: cricket or animal?)
  - Syntactic ambiguity (e.g. "I saw the boy with the telescope")
- Translation of phrases (e.g. "George <u>Bush</u>")
- Translation of proper names
- Word inflections (e.g. in Swedish)
- Compounds (e.g. in German and Dutch)
- Word segmentation (e.g. in Arabic and Chinese)
- …

Handling different languages

- Character sets
- Punctuation and marks
- Word separation
- Digits
- Writing direction
- Formatting styles
- Character shapes
- Sort order and case
- Date and time formatting



รายงานข่าวจากธนาคารแห่งประเทศไทยเปิดเผยว่า การให้บริการ
บัตรเครดิตแยกตามประเภทบัตรเครดิตไตรมาส 2 ของปีนี้เทียบกับ
ไตรมาสแรก โดยปริมาณการใช้จ่ายผ่านบัตรเครดิตโดยรวมลดลง
จาก 8.3 หมื่นล้านบาทในสิ้นไตรมาสแรก เหลือ 7.8 หมื่นล้านบาท
ซึ่งทั้งปริมาณการใช้จ่ายในประเทศลดลงจาก 6.4 หมื่นล้านบาท
เหลือ 5.9 หมื่นล้านบาทในไตรมาสสองของปี ส่วนปริมาณการใช้จ่าย
ในต่างประเทศลดลงจาก 2.6 พันล้านบาท เหลือ 2.1 พันล้านบาท
ขณะที่ยอดสินเชื่อคงค้างปรับตัวเพิ่มขึ้นจากสิ้นไตรมาสแรกที่ระดับ
5.8 หมื่นล้านบาท เป็น 6.3 หมื่นล้านบาทในไตรมาสที่สอง

Lexical resources for translation

- The use of machine-readable, human-crafted **bilingual dictionaries** (and thesauri, word-lists and other similar resources)
- The use of translation resources generated by statistical approaches from suitable training data
  - Require parallel or comparable corpora
- The use of a ‘full’ **machine translation** system
  - Commercial or open source

The "resource bottleneck"

- There are approximately 6,800 known languages in the world and just over 2,000 have a writing system
  - **Around 300** have some kind of language processing tools
  - CLIR/MLIR performance depends on the availability of high-quality translation resources and language processing tools
- **"Resource bottleneck"**
  - Finding ways to acquire, maintain and update language tools and resources in economic manner
  - Building and sustaining resources is costly

Example research: ACCURAT

- ACCURAT project investigated automatically gathering materials from various online sources to train MT systems
  - Focused on generating resources for “rare” language pairs and specific domains (e.g. automotive industry)
  - Used **comparable corpora** to derive translations
- Used various online resources
  - Wikipedia, news sites, parallel versions of websites, social media and crawls of general web pages
- Download Toolkit: http://www.accurat-project.eu/

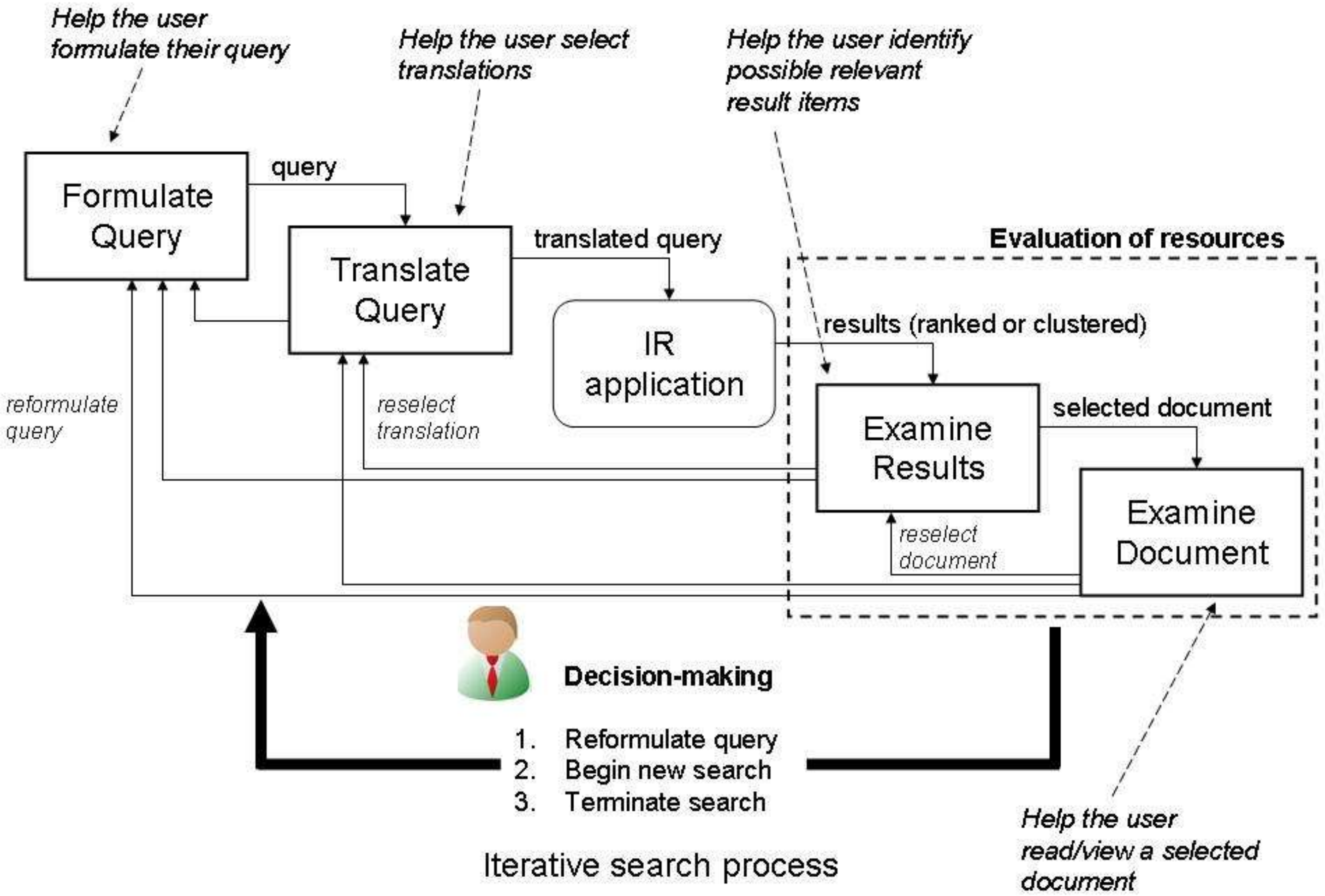
Challenges for multilingual search

3. Providing effective user interaction

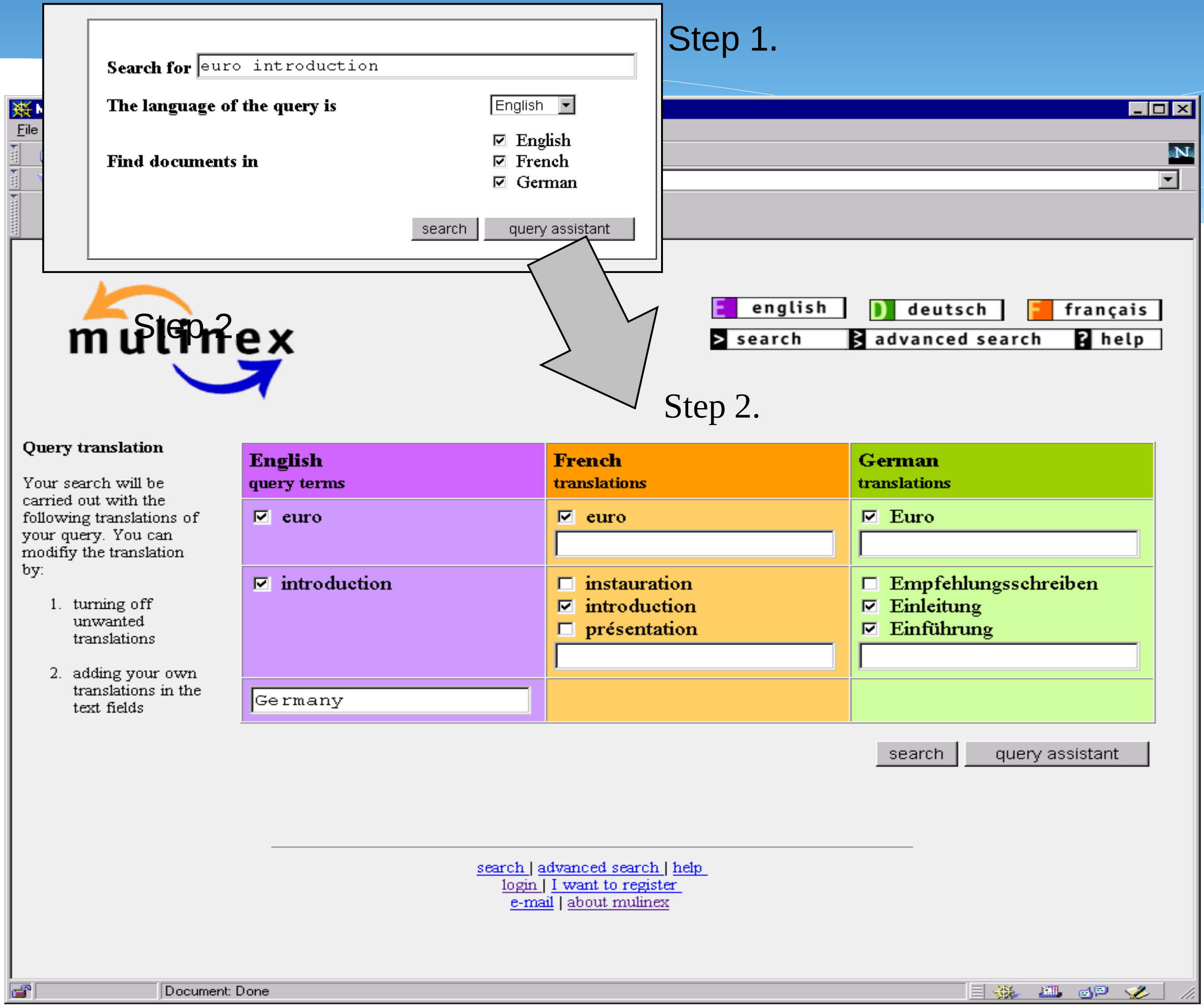
Adapting to language differences

- Individuals can have a **range of foreign language** abilities and knowledge, e.g. passive vs. active skills
  - Can use translation to deal with language differences
- Many people can use **English**
  - Often an interlingua (or default language) for many interfaces
- User’s language skills will affect the design of interfaces, for example for search
  - e.g. monolingual users may need help formulating queries

Supporting CLIR/MLIR



Query translation (Mulinex)



Step 1.

Search for euro introduction

The language of the query is English

Find documents in

☑ English
☑ French
☑ German

Step 2.

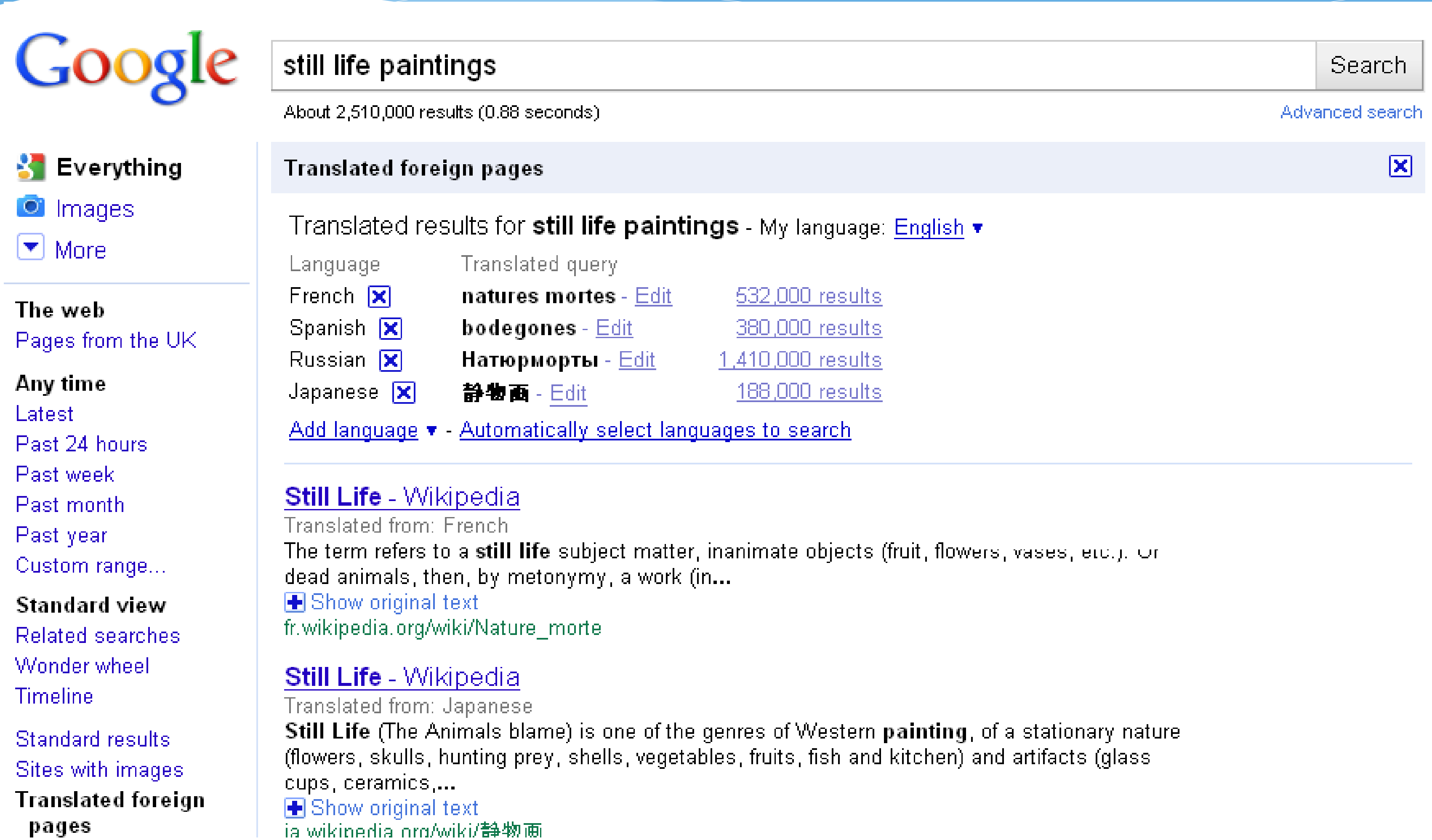
Query translation

Your search will be carried out with the following translations of your query. You can modifiy the translation by:

1. turning off unwanted translations
2. adding your own translations in the text fields

English query terms	French translations	German translations
☑ euro	☑ euro	☑ Euro
☑ introduction	☐ instauration ☑ introduction ☐ présentation	☐ Empfehlungsschreiben ☑ Einleitung ☑ Einführung
Germany		

Translation of search results



still life paintings

Search

About 2,510,000 results (0.88 seconds)

Advanced search

The web

Pages from the UK

Any time

Latest

Past 24 hours

Past week

Past month

Past year

Custom range...

Standard view

Related searches

Wonder wheel

Timeline

Standard results

Sites with images

Translated foreign pages

Translated foreign pages ☒

Translated results for **still life paintings** - My language: English ▾

Language	Translated query	
French ☒	**natures mortes** - Edit	532,000 results
Spanish ☒	**bodegones** - Edit	380,000 results
Russian ☒	**Натюрмортьı** - Edit	1,410,000 results
Japanese ☒	**静物画** - Edit	188,000 results

Add language ▾ - Automatically select languages to search

Still Life - Wikipedia

Translated from: French

The term refers to a **still life** subject matter, inanimate objects (fruit, flowers, vases, etc.). Or dead animals, then, by metonymy, a work (in**...**

⊞ Show original text

fr.wikipedia.org/wiki/Nature_morte

Still Life - Wikipedia

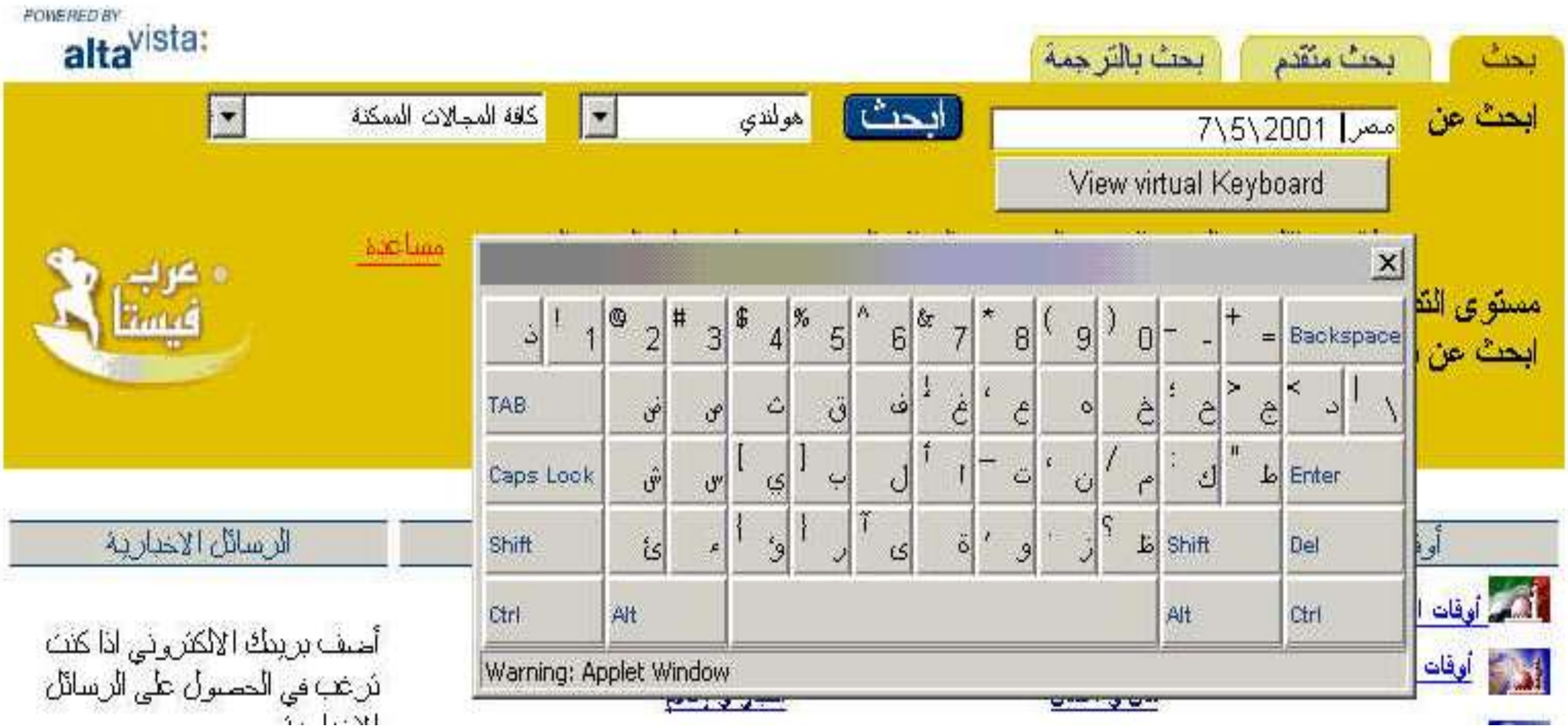
Translated from: Japanese

Still Life (The Animals blame) is one of the genres of Western **painting**, of a stationary nature (flowers, skulls, hunting prey, shells, vegetables, fruits, fish and kitchen) and artifacts (glass cups, ceramics,**...**

⊞ Show original text

ja.wikipedia.org/wiki/静物画

Inputting non-ASCII characters



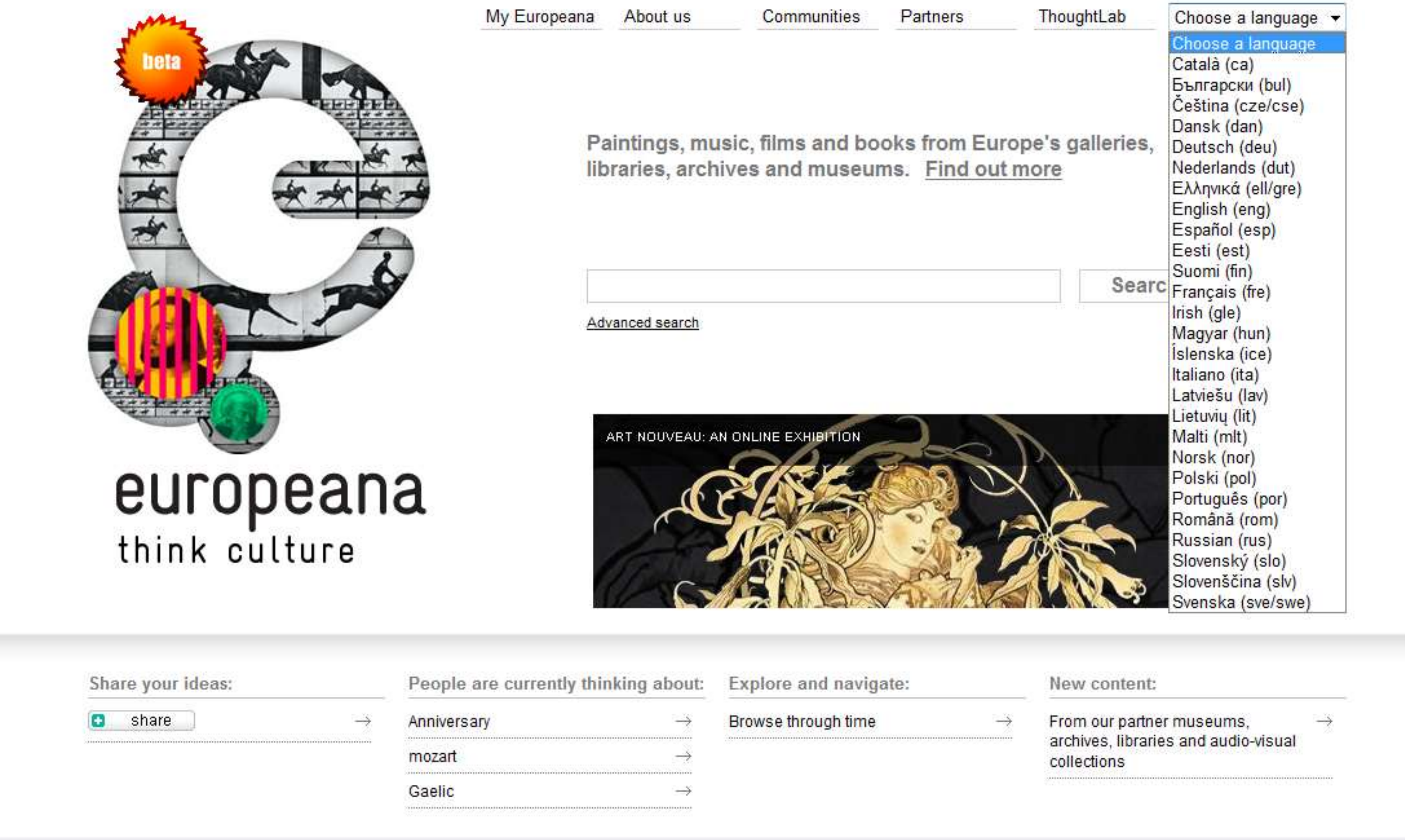
مصر 7\5\2001

Warning: Applet Window

أضف بريدك الالكتروني اذا كنت ترغب في الحصول على الرسائل

Example of non-ASCII keyboard character input (Arabvista.com).

Giving users a choice of interface language



Paintings, music, films and books from Europe's galleries, libraries, archives and museums. Find out more

Share your ideas:

People are currently thinking about:

- Anniversary →
- mozart →
- Gaelic →

Explore and navigate:

- Browse through time →

New content:

- From our partner museums, archives, libraries and audio-visual collections →

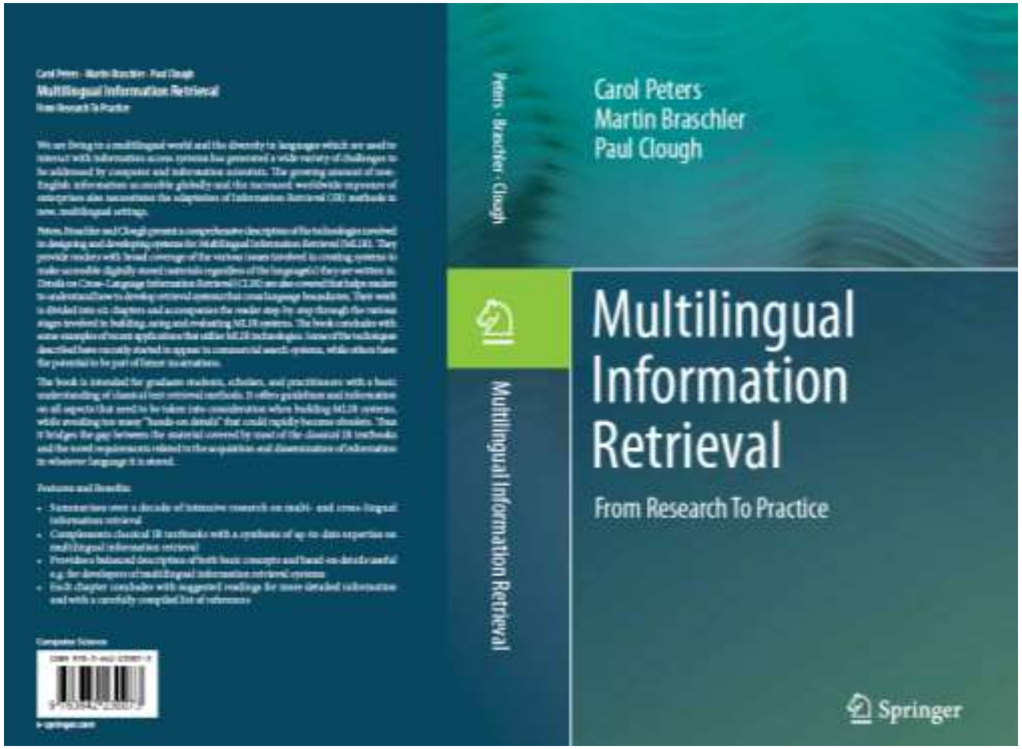
Challenges for multilingual search

4. Going from research to practice

Are we stuck in the lab?

- The technologies for CLIR/MLIR are proven in the lab, but where are the applications in practice?
  - For example, consider the lack of CLIR/MLIR in large-scale information services, such as Amazon
- We need to go from research to practice
  - Guidelines for developing CLIR/MLIR applications
  - Collaborations between academics and business
  - Effective methods for knowledge transfer (in both directions)

Research to practice



Carol Peters · Martin Braschler · Paul Clough

Multilingual Information Retrieval

From Research To Practice

We are living in a multilingual world and the diversity in languages which are used to interact with information access systems has generated a wide variety of challenges to be addressed by computer and information scientists. The growing amount of non-English information accessible globally and the increased worldwide exposure of enterprises also necessitates the adaptation of Information Retrieval (IR) methods to new, multilingual settings.

Peters, Braschler and Clough present a comprehensive description of the technologies involved in designing and developing systems for Multilingual Information Retrieval (MLIR). They provide readers with broad coverage of the various issues involved in creating systems to make accessible digitally stored materials regardless of the language(s) they are written in. Details on Cross-Language Information Retrieval (CLIR) are also covered that helps readers to understand how to develop retrieval systems that cross language boundaries. Their work is divided into six chapters and accompanies the reader step-by-step through the various stages involved in building, using and evaluating MLIR systems. The book concludes with some examples of recent applications that utilise MLIR technologies. Some of the techniques described have recently started to appear in commercial search systems, while others have the potential to be part of future incarnations.

The book is intended for graduate students, scholars, and practitioners with a basic understanding of classical text retrieval methods. It offers guidelines and information on all aspects that need to be taken into consideration when building MLIR systems, while avoiding too many "hands-on details" that could rapidly become obsolete. Thus it bridges the gap between the material covered by most of the classical IR textbooks and the novel requirements related to the acquisition and dissemination of information in whatever language it is stored.

Features and Benefits:

- Summarizes over a decade of intensive research on multi- and cross-lingual information retrieval
- Complements classical IR textbooks with a synthesis of up-to-date expertise on multilingual information retrieval
- Provides a balanced description of both basic concepts and hand-on details useful e.g. for developers of multilingual information retrieval systems
- Each chapter concludes with suggested readings for more detailed information and with a carefully compiled list of references

Computer Science

ISBN 978-3-642-23007-3

► springer.com

Peters · Braschler · Clough

Multilingual Information Retrieval

Carol Peters
Martin Braschler
Paul Clough

Multilingual Information Retrieval

From Research To Practice

Springer

Summary

- Search systems are used interactively so you must consider and design for **end users**
  - **Implication**: multilingual search is NOT just an engineering problem; you will also need to understand users' cultural background and language abilities
- Searching is part of wider information seeking activities supported by search **applications**
  - **Implication**: you may have to consider localisation of the whole service and not just querying support
- Research has shown we can do cross-language search in multiple languages well (**in the lab**)
  - **Implication**: the challenge is going from research to practice

Questions?

Paul Clough

Senior Lecturer in Information Retrieval

p.d.clough@sheffield.ac.uk

http://ir.shef.ac.uk/cloughie/