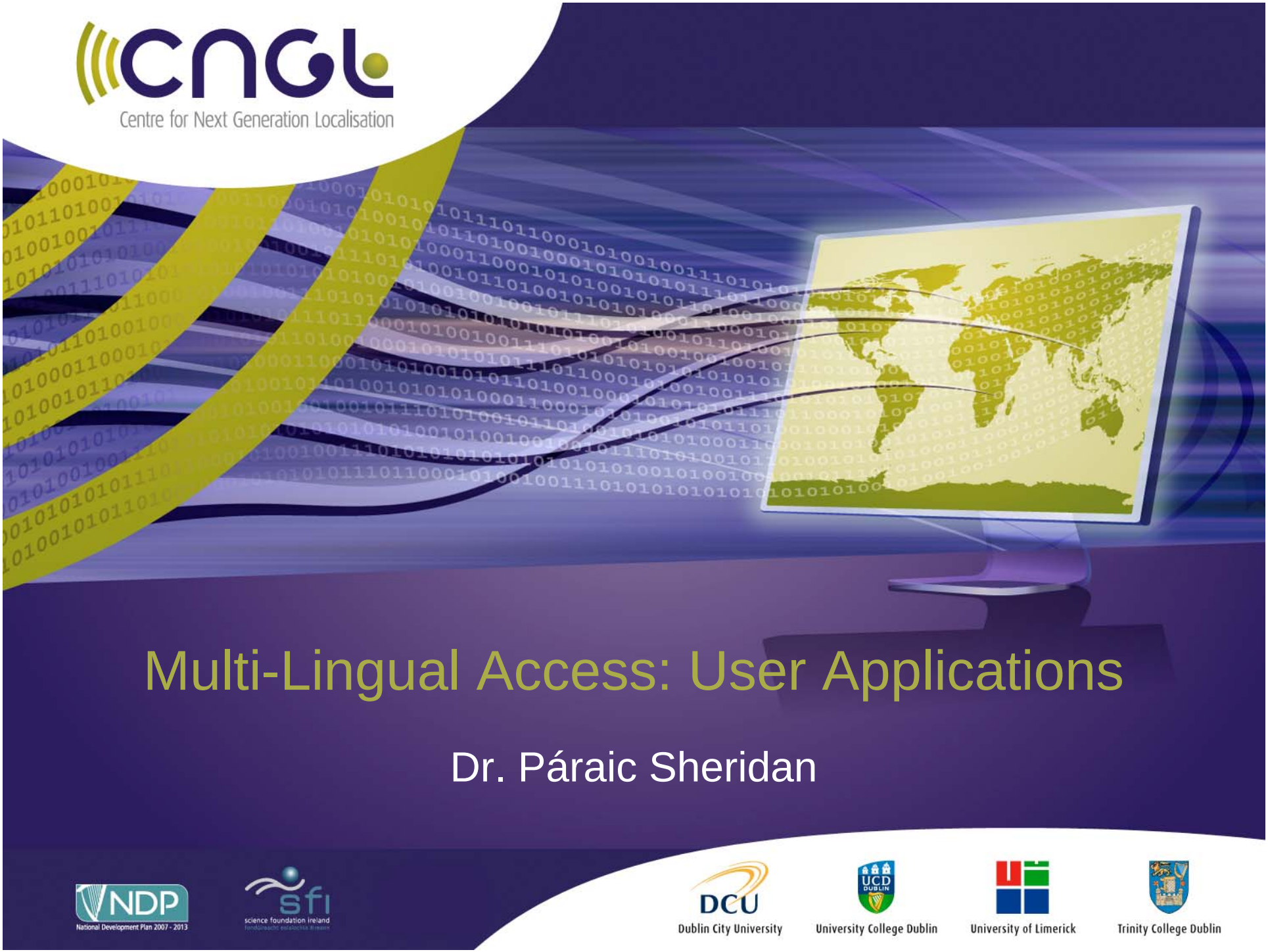
Commercial Cross-Language Search

Páraic Sheridan, Centre for Next Generation Localisation, Dublin City University, Ireland

TrebleCLEF Summer School on
Multilingual Information Access
Hotel Santa Croce in Fossabanda, Pisa, Italy
15-19 June 2009


CNGL
Centre for Next Generation Localisation
Multi-Lingual Access: User Applications
Dr. Páraic Sheridan
NDP
National Development Plan 2007 - 2013
sfi
science foundation ireland
DCU
Dublin City University
UCD DUBLIN
University College Dublin
University of Limerick
Trinity College Dublin

Outline

- Introduction
- Economic Imperative for Multilingual Information Access
- The World Wide Web
- Enterprise Search
- Areas of Opportunity

Introduction

By way of Introduction:

- PhD thesis on Cross-Language search using ‘Similarity Thesauri’ – ETH Zurich 1998.
- 10 years at TextWise LLC in Syracuse, NY.
  - CINDOR cross-language search in Government & Enterprise market (US focus)
  - More recently semantic technologies & online advertising.
- Now at Centre for Next Generation Localisation (CNGL) at Dublin City University (DCU).

Centre for Next Generation Localisation

Our work is guided by the vision of enabling people to interact with content, products and services in their own language, according to their own culture, and according to their own personal needs.

http://www.cngl.ie/

CNGL – 13 Partners…

The Centre for Next Generation Localisation

- Industry-Academia Partnership
- Funded jointly by Science Foundation Ireland (SFI) and Industry Partners
- 2007-2012

- DCU, TCD, UCD, UL
- IBM, Microsoft, Symantec, Dai Nippon Printing, SDL, VistaTEC, Alchemy, Traslán, Speech Storm

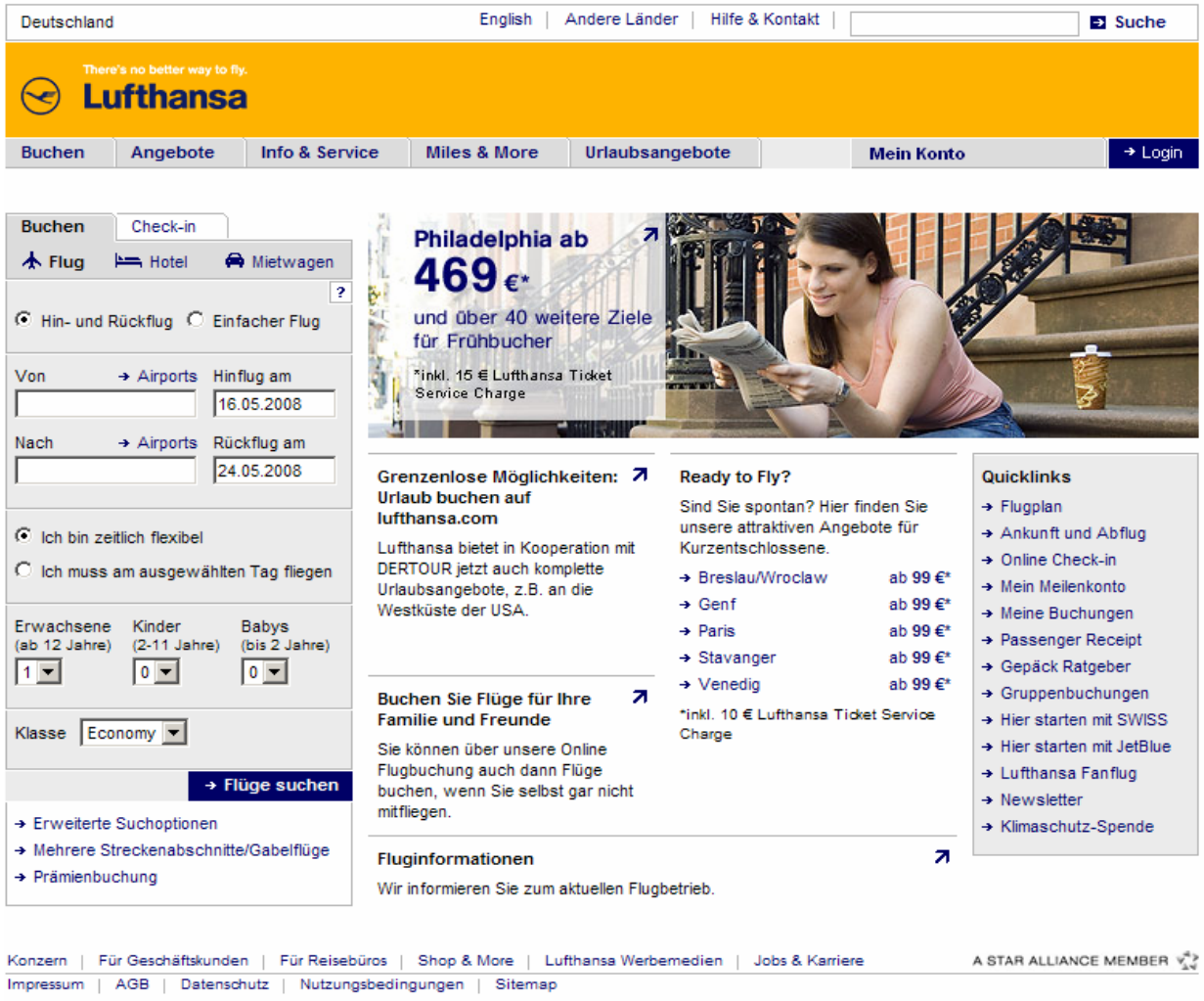
- 33M euro, ~17M SFI + ~13M Industry Partners + Overheads
- 100 people (mostly research staff: PhD students and PostDocs)

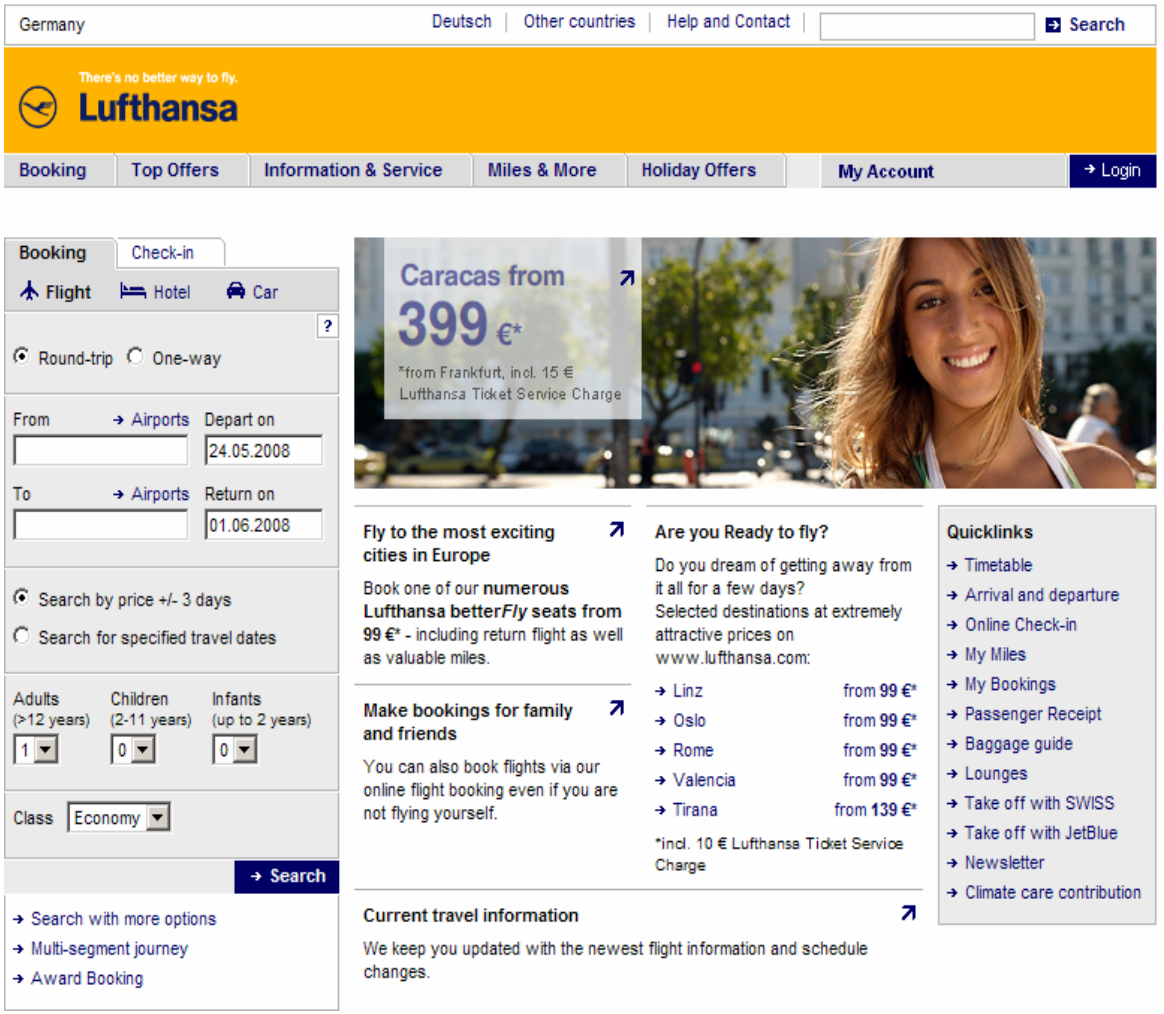
- http://www.cngl.ie/

Centre for Next Generation Localisation

The Economic Imperative

Simple Example: I want to fly to Berlin




Germany
Deutsch | Other countries | Help and Contact
Search
There's no better way to fly.
Lufthansa
Booking
Top Offers
Information & Service
Miles & More
Holiday Offers
My Account
Login
Booking
Check-in
Flight
Hotel
Car
Round-trip
One-way
From
Airports
Depart on
24.05.2008
To
Airports
Return on
01.06.2008
Search by price +/- 3 days
Search for specified travel dates
Adults (>12 years)
Children (2-11 years)
Infants (up to 2 years)
Class
Economy
Search
Search with more options
Multi-segment journey
Award Booking
Caracas from
399 €*
*from Frankfurt, incl. 15 € Lufthansa Ticket Service Charge
Fly to the most exciting cities in Europe
Book one of our numerous Lufthansa betterFly seats from 99 €* - including return flight as well as valuable miles.
Make bookings for family and friends
You can also book flights via our online flight booking even if you are not flying yourself.
Are you Ready to fly?
Do you dream of getting away from it all for a few days? Selected destinations at extremely attractive prices on www.lufthansa.com:
Linz from 99 €*
Oslo from 99 €*
Rome from 99 €*
Valencia from 99 €*
Tirana from 139 €*
*incl. 10 € Lufthansa Ticket Service Charge
Current travel information
We keep you updated with the newest flight information and schedule changes.
Quicklinks
Timetable
Arrival and departure
Online Check-in
My Miles
My Bookings
Passenger Receipt
Baggage guide
Lounges
Take off with SWISS
Take off with JetBlue
Newsletter
Climate care contribution

日本 English | Other countries | ヘルプ & コンタクト | 検索

There's no better way to fly.
Lufthansa

フライト予約 | お得な料金 | Info & Service | Miles & More | マイ アカウント | → ログイン

Booking | Check-In

Flight | Hotel | Car

往復 / 片道

出発地 → 都市の検索 往路搭乗日 2008.05.16

目的地 → 都市の検索 復路搭乗日 2008.05.24

料金を優先
日程を優先

大人（12才以上） 1 | 小児（2~12才） 0 | 幼児（2才未満） 0

クラス エコノミー

→ 空席状況を調べる

→ 条件を絞り込んで検索する
→ 複雑な旅程の予約
→ アワード予約

ビジネスクラスなら
『WEB得 ビジ』

ヨーロッパ往復

¥510,000 ~

betterFly - ヨーロッパ内往復 スーパー割引

99 € ~ *

* 税・手数料込み

お食事とお飲み物

今月のスターシェフのご紹介をはじめ、ルフトハンザ自慢の機内食とワインについてのご案内です。

『WEB得28』 - 春夏のヨーロッパへ、ルフトハンザで。

→ フランクフルト	¥79,000 ~
→ パリ	¥79,000 ~
→ ニース	¥84,000 ~
→ ローマ	¥79,000 ~
→ ボルドー	¥84,000 ~
→ フィレンツェ	¥95,000 ~
→ ベニス	¥84,000 ~
→ ロンドン	¥79,000 ~
→ その他の都市	

クイックリンク
→ タイムテーブル
→ 発着案内
→ オンラインチェックイン
→ 自動チェックイン
→ マイルの確認について
→ 日本のお客様へ
→ ルフトハンザモバイル
→ メールニュース申し込み
→ お問い合わせ
→ 企業の皆様へのご案内
→ コーポレート予約
→ プレスリリース
→ ルフトハンザについて

お客様へのお知らせ

ご出発前に必ずご一読ください。
「機内におけるシートベルト着用義務に関する新規定」「燃油追加料金徴収について」「機内持ち込み手荷物に関する新規定」

ルフトハンザ コンツェルン | 企業の皆様 | 旅行会社の皆様 | Lufthansa WorldShop | Lufthansa Ambient Media | A STAR ALLIANCE MEMBER
サイト管理者 | ご利用に関する規定 | 個人データの保護 | ご利用条件 | サイトマップ

Is it just a matter of translating content?

No: Adapt imagery to local/cultural norms

Adapt date format, currency, etc. to locale

What is Localisation?

- industrial process of adapting digital content to culture, locale and linguistic environment
- at high quality, speed, volume and low cost
- key enabling, value adding, multiplier component of global manufacturing, services, software and content distribution industry
- drives products into global markets
- social, cultural and political dimension

Example – Microsoft Office 2007

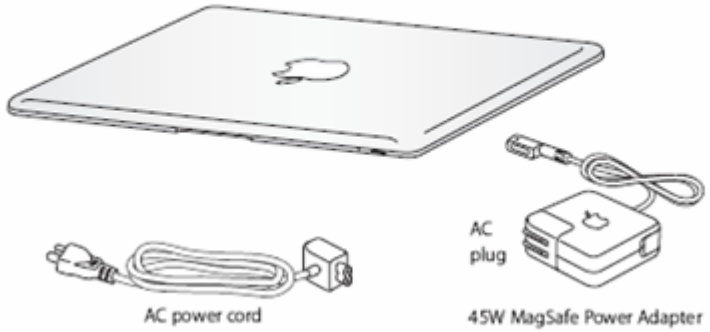
Introduction: Microsoft Office 2007 as a localisation project

- Scope
  - Volumes
    - Software: 2 million words
    - Content: 3 million words end-user, 4 million developer
  - Languages
    - 30+ languages full localization
    - up to 100 with Local Language Program
  - Release schedule
    - 10 languages sim-ship

[Source: Microsoft]

And not just a Recent Need: e.g. Product Information

What's in the Box



Important: Remove the protective film covering the 45W MagSafe Power Adapter before setting up your MacBook Air.

About Optical Discs

Although your MacBook Air doesn't have an optical disc drive, it does include DVD discs with important software. You can easily access this software, as well as install applications and access data from other optical discs, using the optical disc drive on another Mac or Windows computer and the Remote Disc feature on your MacBook Air. You can also use the optional MacBook Air SuperDrive, an external optical disc drive.

包装箱内的物品

【重要事项】安装 MacBook Air 之前，请先撕去包着 45W MagSafe Power Adapter 的保护膜。

关于光盘

MacBook Air 未配置光盘驱动器，但是附带了含有重要软件的 DVD 光盘。通过在另一台 Mac 或 Windows 电脑上使用光盘驱动器并在 MacBook Air 上使用 "远程光盘" 功能，您可以轻松使用该软件，以及从其他光盘安装应用程序和访问数据。您还可以选购 MacBook Air SuperDrive 驱动器，并将其用做外置光盘驱动器。

Example: Magazines – Content, Locale, Culture…

6,912 Living Languages.

Language Statistics

http://www.aboutlanguageschools.com/language/statistics/

The following table shows each one of the continents with their respective language statistics.

A **Living Language** is defined as one that has at least one speaker for whom it is their first language; these not includes extinct languages and languages that are used only as a second language.
In this count, each language is counted only once so that the total count at the bottom of the table shows the total of living languages in the world. A language that is spoken in more than one country is counted under the area of its primary country.

Number of Speakers is the total number of people who use these languages as their first language, regardless of where in the world they may live.

Distribution of Languages by Area

Area	Living Languages		Number of Speakers	
	Count	Porcent	Count	Porcent
Africa	2 092	30.3	675 887 158	11.8
Americas	1 002	14.5	47 559 381	0.8
Asia	2 269	32.8	3 489 897 147	61.0
Europe	239	3.5	1 504 393 183	26.3
Pacific	1 310	19.0	6 124 341	0.1
Totals	6 912	100.0	5 723 861 210	100.0

Languages of the European Union

Language	Official in	Since
Bulgarian	Bulgaria	2007
Czech	Czech Republic	2004
Danish	Denmark	1973
Dutch	Netherlands and Belgium	1958
English	Ireland, Malta and United Kingdom	1958
Estonian	Estonia	2004
Finnish	Finland	1995
French	Belgium, France and Luxembourg	1958
German	Austria, Belgium, Germany and Luxembourg	1958
Greek	Cyprus and Greece	1981
Hungarian	Hungary	2004

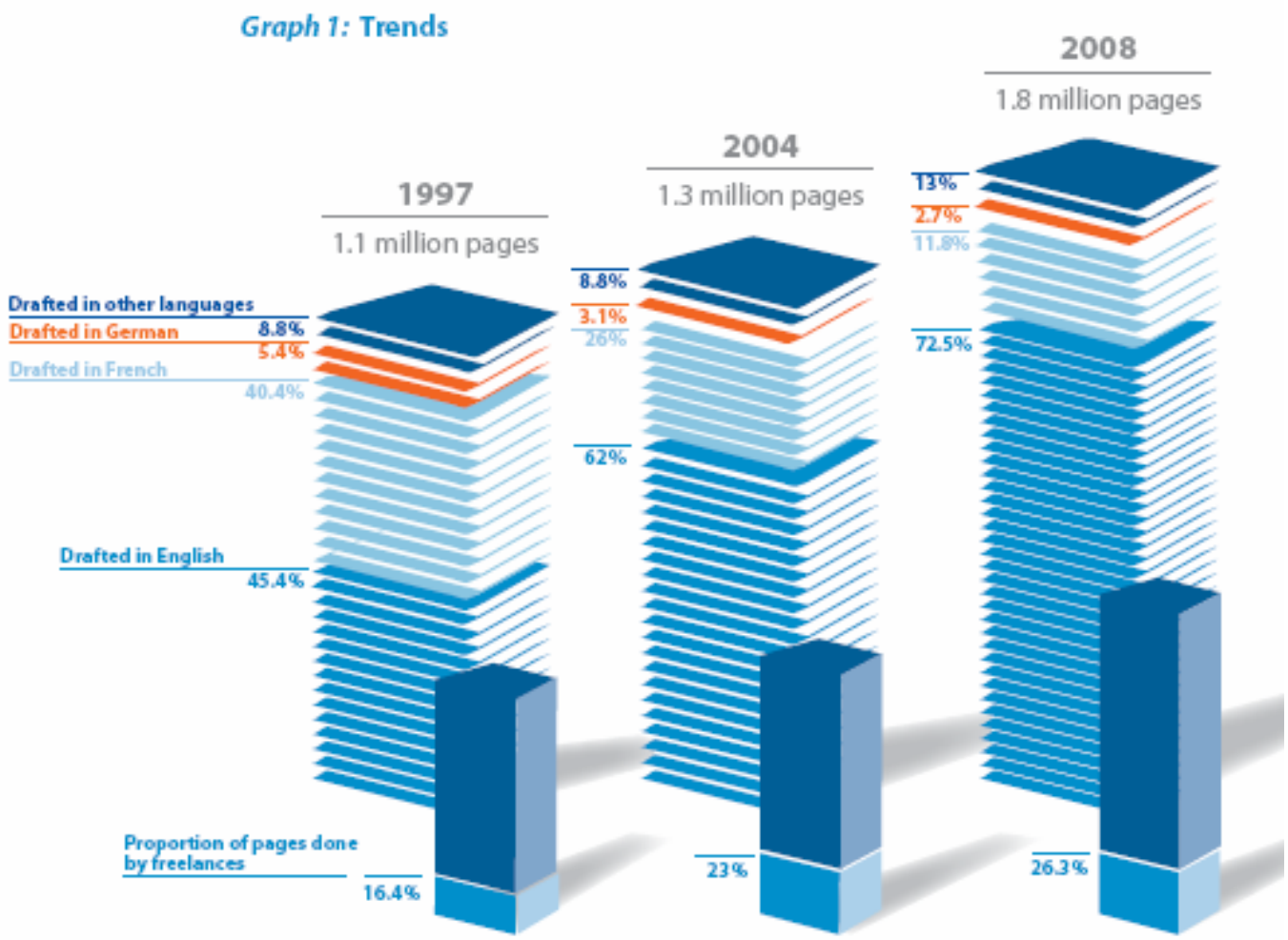
Language	Official in	Since
Irish	Ireland	2007
Italian	Italy	1958
Latvian	Latvia	2004
Lithuanian	Lithuania	2004
Maltese	Malta	2004
Polish	Poland	2004
Portuguese	Portugal	1986
Romanian	Romania	2007
Slovak	Slovakia	2004
Slovene	Slovenia	2004
Spanish	Spain	1986
Swedish	Finland and Sweden	1995

http://en.wikipedia.org/wiki/Languages_of_the_European_Union

Translation in the European Union

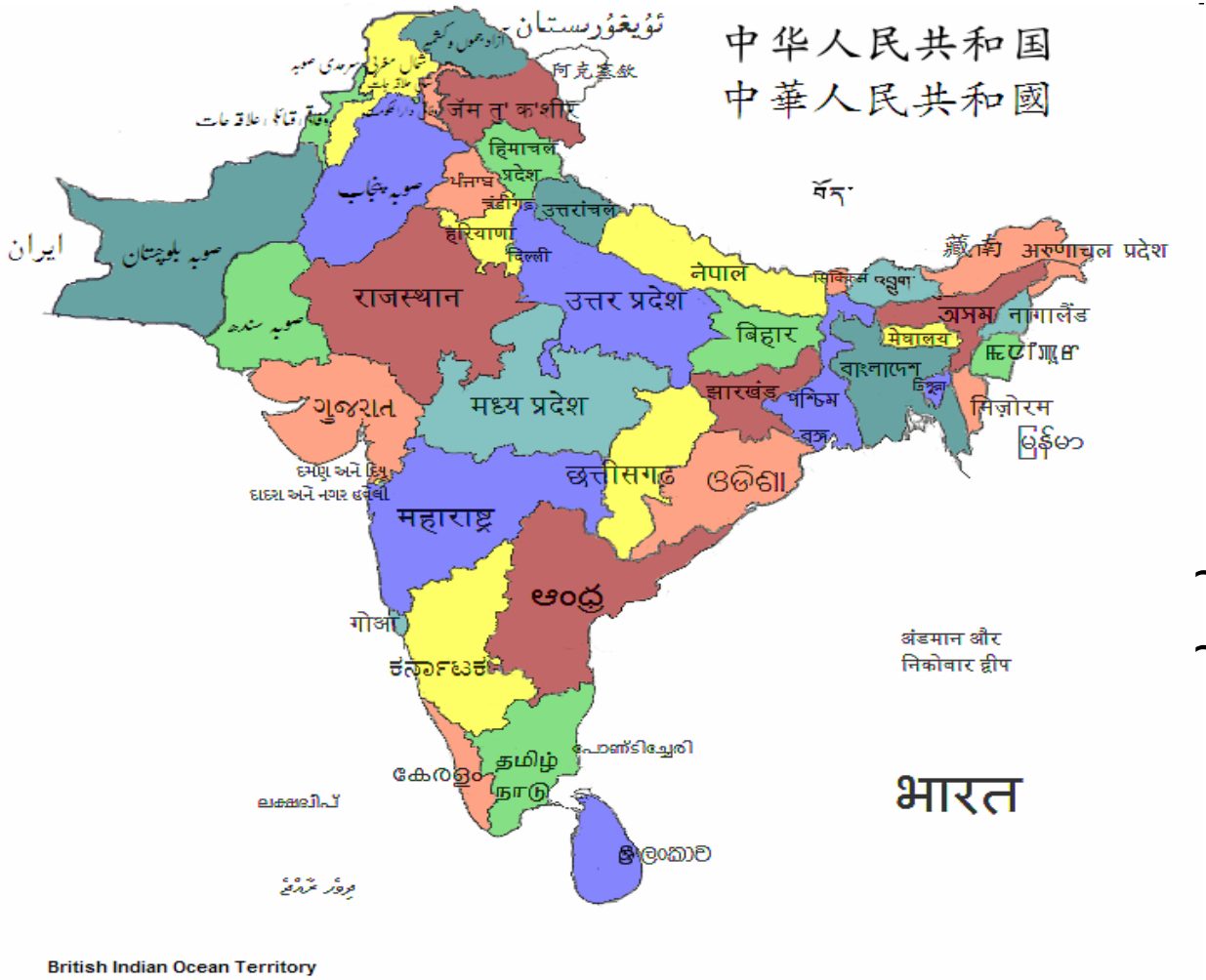
- European Commission Directorate General for Translation.
- DGT divided along language lines with separate department/unit for each of 23 official languages plus three additional support groups.
- 1,750 translators working full-time plus 600 support staff.
- In 2008 DGT translated 1,805,689 pages.

Translation Volume at EU-DGT



http://ec.europa.eu/dgs/translation/bookshelf/brochure_en.pdf

Let's Take a Look at Multi-Lingual India



22 'Official' languages

~60m Internet users
~5% penetration

About 10% English speakers

Rank	Country	% English Speakers	Total Population	Total English Speakers	As First Language	As an Additional Language
1	United States	95.81%	262,375,152	251,388,301	215,423,557	35,964,744
2	India	10.66%	843,900,000	90,000,000	226,449	65,000,000 *second* language speakers. 25,000,000 *third* language speakers

Source: Wikipedia

- Only about 10% of the Indian population speaks English.
- Mostly as a second or third language.

Lets Look at China

中国互联网络信息中心
China Internet Network Information Center

China now has Largest Internet Population

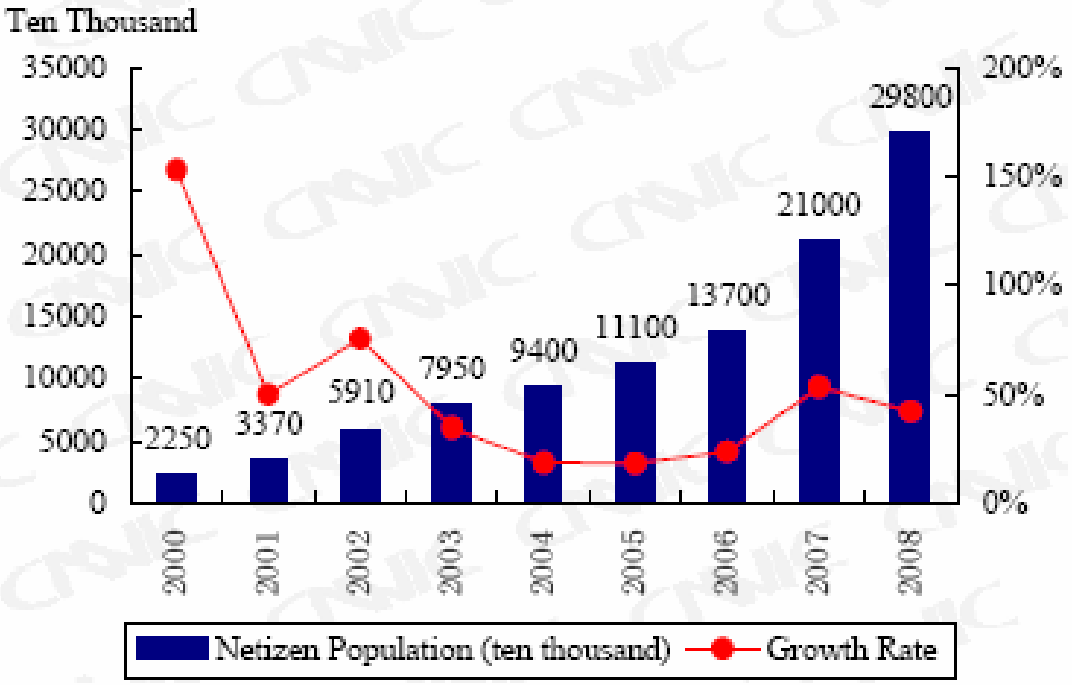


Figure 1 Size of Netizens and Growth Rate in China from 2000 to 2008

- China 298 million Internet users.
- USA 248 million Internet users.

…With only 22% Internet Penetration!

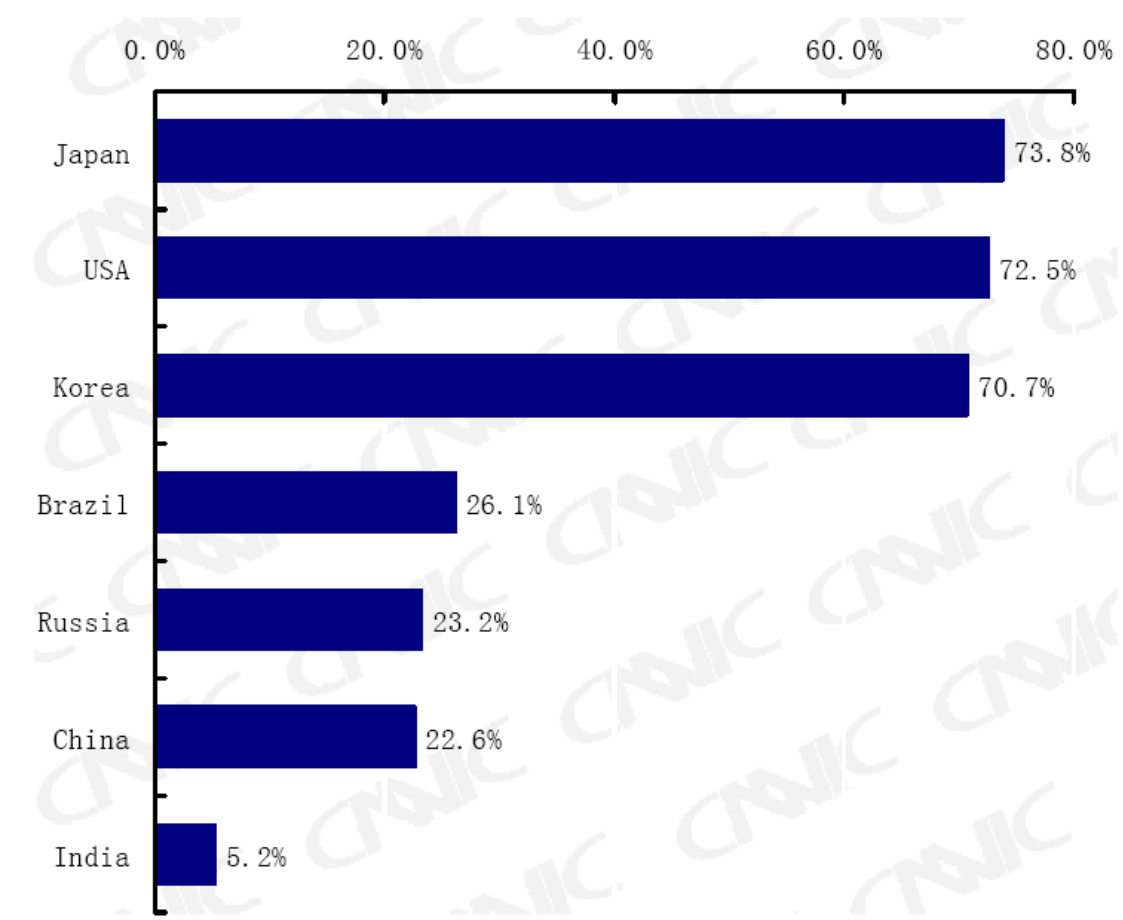


Figure 2 Internet Penetration Rates of Some Countries

中国互联网络信息中心
China Internet Network Information Center

And less than 1% Population speaks English!

Rank	Country	% English Speakers	Total Population	Total English Speakers
18	China	0.77%	1,300,000,000	10,000,000

Source: Wikipedia

- Much statistics/analysis of internet participation in China:
  - "Statistical Survey Report on the Internet Development in China"
  - CNNIC – China Internet Network Information Centre
  - http://www.cnnic.cn/uploadfiles/pdf/2009/3/23/131303.pdf

Global Mega-Trends: Power Shift East

The rebalancing of power

The global economic landscape is changing, and the emerging markets are playing an increasingly significant role. Economic power is moving from developed to emerging economies – from West to East and North to South. Emerging economies accounted for 44% of global GDP in 2007[1]; while projected GDP growth rates for major developed markets in 2009 are now predicted to lie between -0.2% and 0.5%, emerging markets are expected to grow at 6.1% on average, with China (9.3%) and India (6.9%) performing even better.[2] This may be less than was projected before the financial crisis, but emerging markets still demonstrate considerably stronger growth than the developed world. The financial crisis may have undermined decoupling theories, with emerging markets also suffering from issues of liquidity, investor confidence and over-valued assets, but their hunger for growth (alongside their rapidly industrializing economies and growing populations) should set them on the path to recovery more quickly. In China and Russia's case, their huge accumulated reserves (China with US$1.9 trillion[3] and Russia with US$560 billion[4]) will no doubt ease the pain.

ERNST & YOUNG
Quality In Everything We Do

http://www.ey.com/Publication/vwLUAssets/Global_megatrends_2009/$FILE/Global_megatrends_2009.pdf

Centre for Next Generation Localisation

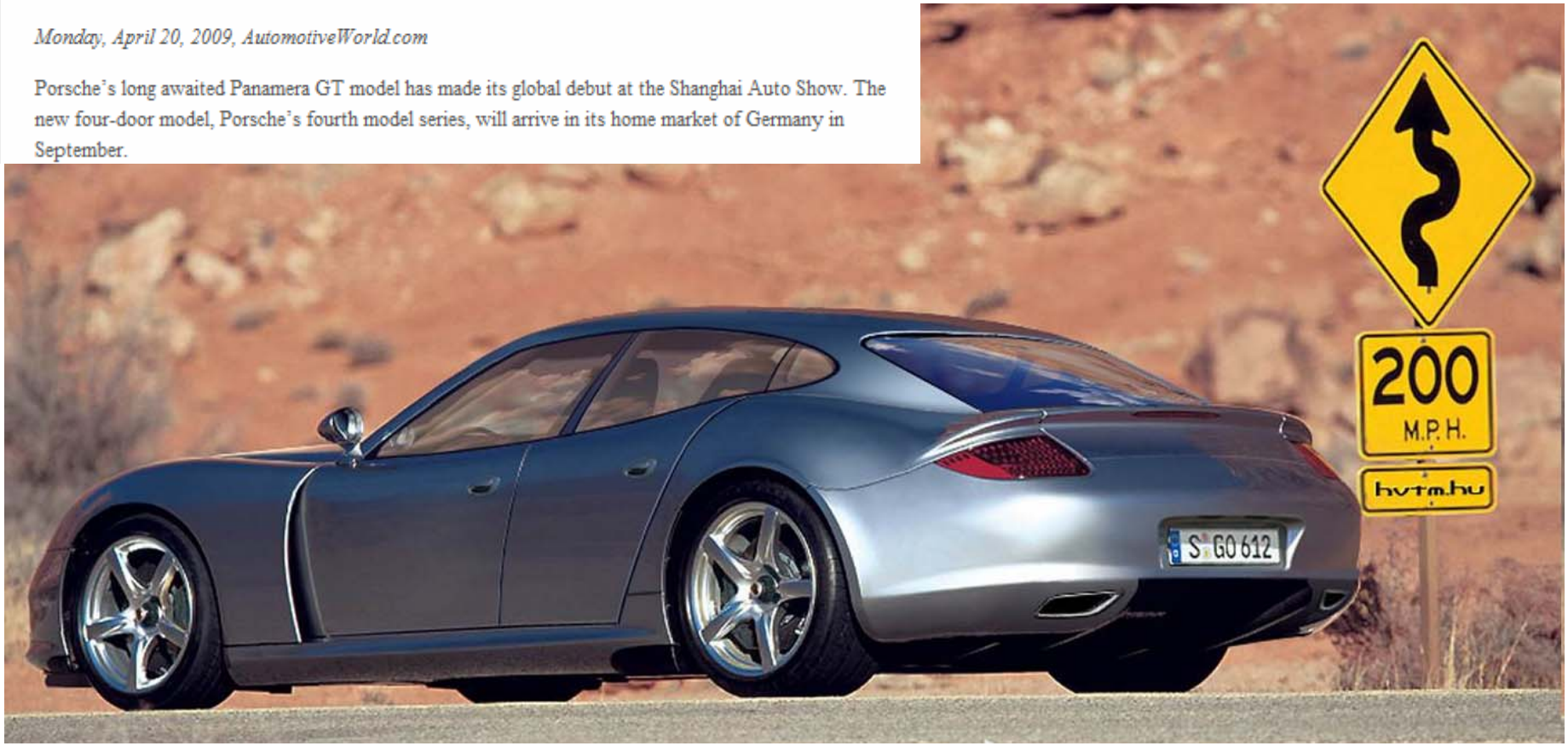
Introducing the Porsche Panamera

You are here: Home > News & Analysis > Emerging Markets

China: Porsche Panamera debuts in Shanghai

Monday, April 20, 2009, AutomotiveWorld.com

Porsche's long awaited Panamera GT model has made its global debut at the Shanghai Auto Show. The new four-door model, Porsche's fourth model series, will arrive in its home market of Germany in September.



Global Brands Moving East…

Why global brands now rise in the east

Print

By John Gapper

Published: April 22 2009 19:30 | Last updated: April 22 2009 19:30

Not long ago, Joanna Seddon, a marketing executive, lost a button on her Louis Féraud suit and looked for a store in New York or London at which to get it replaced.

Ms Seddon, an executive vice-president of Millward Brown, was out of luck: the late French designer's New York store on Madison Avenue had closed. She had to turn to China, where Féraud has 11 outlets. A brand made popular in the US in the 1980s by the soap operas *Dallas* and *Dynasty* had gone east.

The realigning of Louis Féraud from the US to China is an unusual story but it is becoming more common. As it does, our postwar assumption that the US is the place where most global consumer brands get launched before being spread around the world is being undermined.

This week, **Porsche** chose the Shanghai motor show to **launch its Panamera** four-door saloon, the fourth Porsche line after the 911, the Boxster/Cayman and the Cayenne (a US-oriented sports utility vehicle).

This time, there is no mistaking the Asian influence on Porsche's product development. The Panamera is a global model but its length – nearly as long as the stretched Series 5 that BMW made for China – tells the story. Rich car buyers in China prefer to be driven by chauffeurs.

The car industry is a leading indicator. The US slump has led to **China** turning into the world's largest car market this year, accentuating a long-term shift towards Asia.

Done

This week, **Porsche** chose the Shanghai motor show to **launch its Panamera** four-door saloon, the fourth Porsche line after the 911, the Boxster/Cayman and the Cayenne (a US-oriented sports utility vehicle).

This time, there is no mistaking the Asian influence on Porsche's product development. The Panamera is a global model but its length – nearly as long as the stretched Series 5 that BMW made for China – tells the story. Rich car buyers in China prefer to be driven by chauffeurs.

The car industry is a leading indicator. The US slump has led to **China** turning into the world's largest car market this year, accentuating a long-term shift towards Asia.

China is a lone bright spot for **General Motors**, which has chronic problems in the US with devalued brands and sagging sales, and is attempting to shed Opel and Vauxhall in Europe. In China, not only Cadillac but even Buick, a dismal mid-market brand at home, have allure.

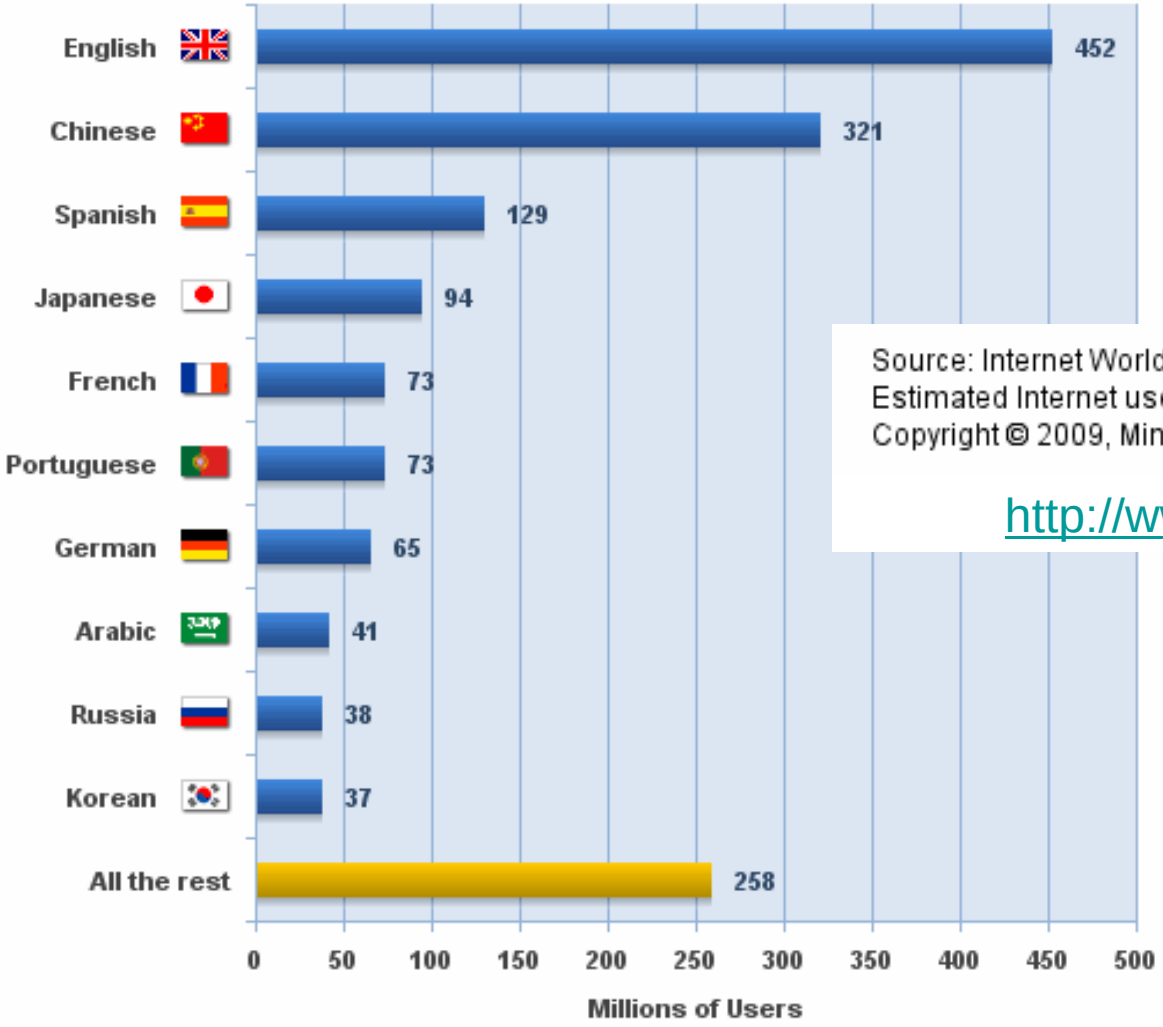
But it is not alone. The shift in consumer buying power towards emerging markets is not only giving Chinese brands a much better chance of breaking out of their domestic market but is also subtly altering how western companies develop and market global brands.

Published by the Financial Times
FT.com
Published: April 22 2009

Centre for Next Generation Localisation

The World Wide Web

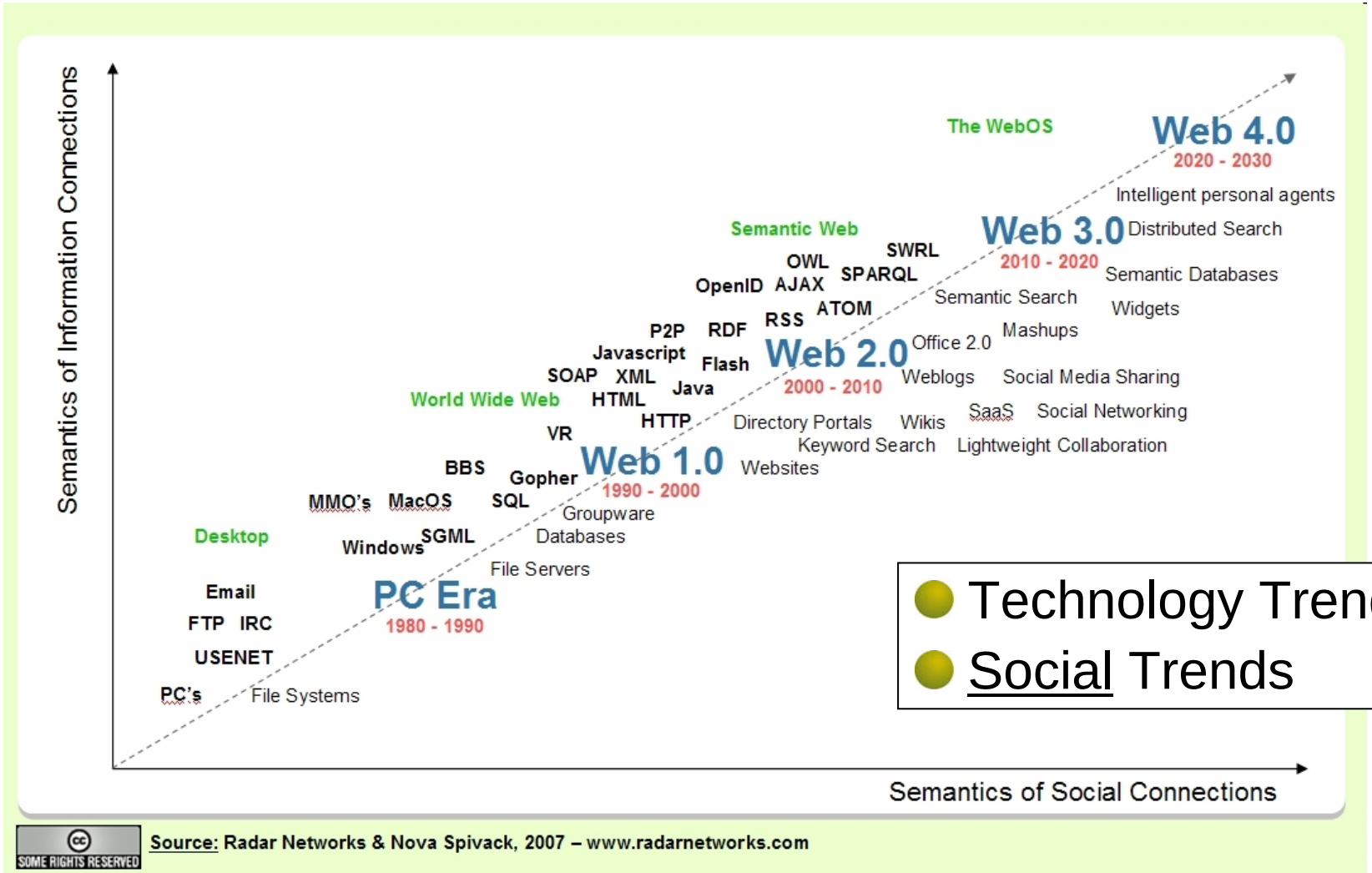
Top 10 Languages – Internet Users (Millions)



Source: Internet World Stats - www.internetworldstats.com/stats7.htm
Estimated Internet users is 1,581,571,589 for 2008
Copyright © 2009, Miniwatts Marketing Group

http://www.internetworldstats.com/

Bear in mind Web Evolution



Source: Radar Networks & Nova Spivack, 2007 – www.radarnetworks.com

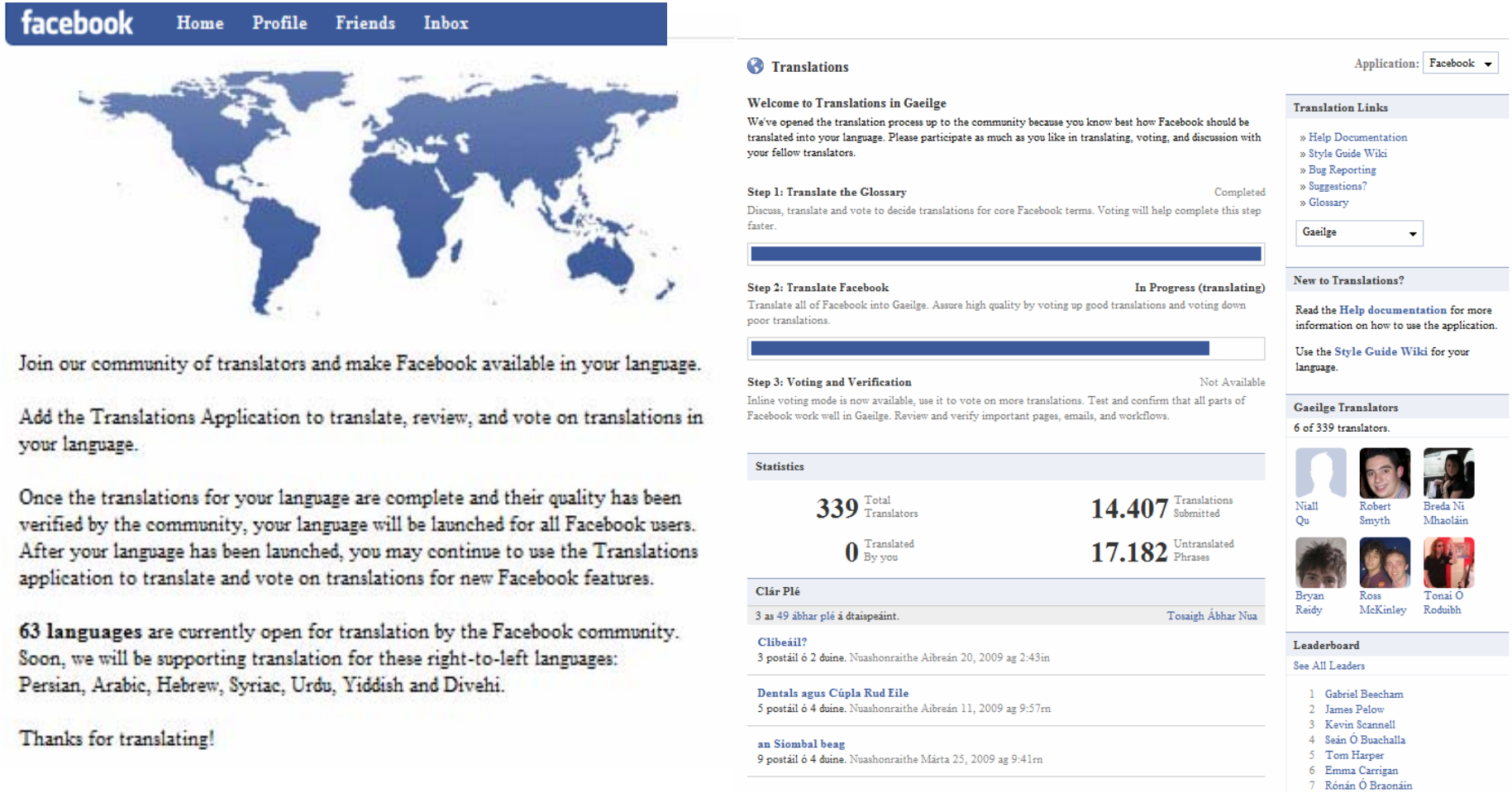
Facebook – 40 Languages

San Francisco - Five years after it was founded in a Harvard dorm room, the online social networking site Facebook has registered its 200 millionth user, the site confirmed Thursday. "Growing rapidly to 200 million users is a really good start, but we have always known that in order for Facebook to help people represent everything that is happening in their world, everyone needs to have a voice," said Facebook founder and CEO Mark Zuckerberg.

Facebook says that 100 million users visit the site every day, and users have an average of 120 "friends" on their networks.

The service is available in 40 languages, with 70 per cent of its users outside the US.

Now 63 Languages! Community Translation



http://www.facebook.com April 26th 2009

The Multi-Lingual Wikipedia

Wikipedia in 265 Languages (>100 Articles)

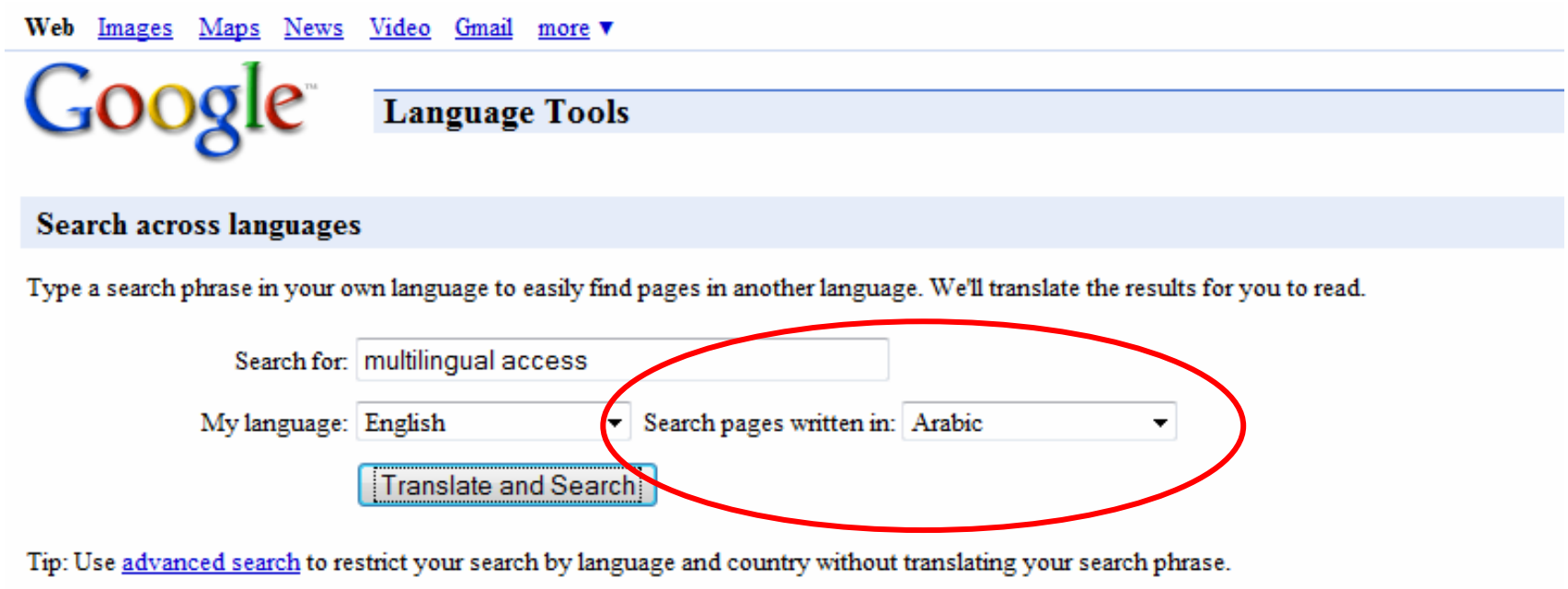
- http://meta.wikimedia.org/wiki/List_of_Wikipedias

- This page contains a list of the official Wikipedias under the auspices of the Wikimedia Foundation for various languages.

- All **265** Wikipedias are ordered by article count .

80	West Frisian	Frysk	fy	11,410
81	Walloon	Walon	wa	11,108
82	Swahili	Kiswahili	sw	11,060
83	Cantonese	粵語	zh-yue	10,728
84	Quechua	Runa Simi	qu	10,396
85	Samogitian	Žemaitėška	bat-smg	10,264
86	Urdu	اردو	ur	10,153
87	Chuvash	Чăвашла	cv	10,143
88	Ripuarian	Ripoarisch	ksh	10,040

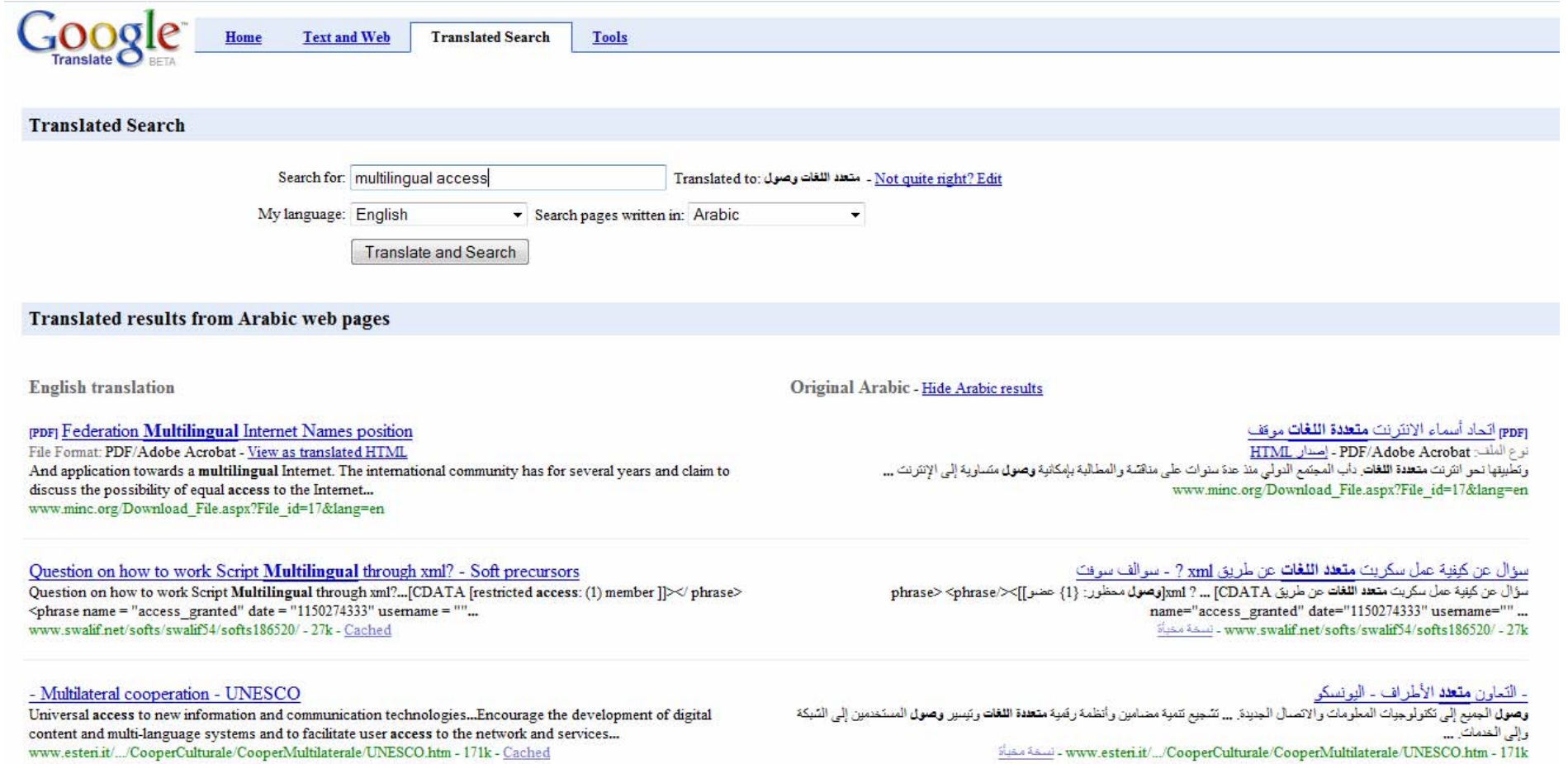
Multi-Lingual Search online?



- 57+ (?) Possible language combinations:

  http://www.google.com/language_tools

Results Automatically Translated.



Just Announced - Google Translator Toolkit

The Official Google Blog

Insights from Googlers into our products, technology, and the Google culture.

Translating the world's information with Google Translator Toolkit

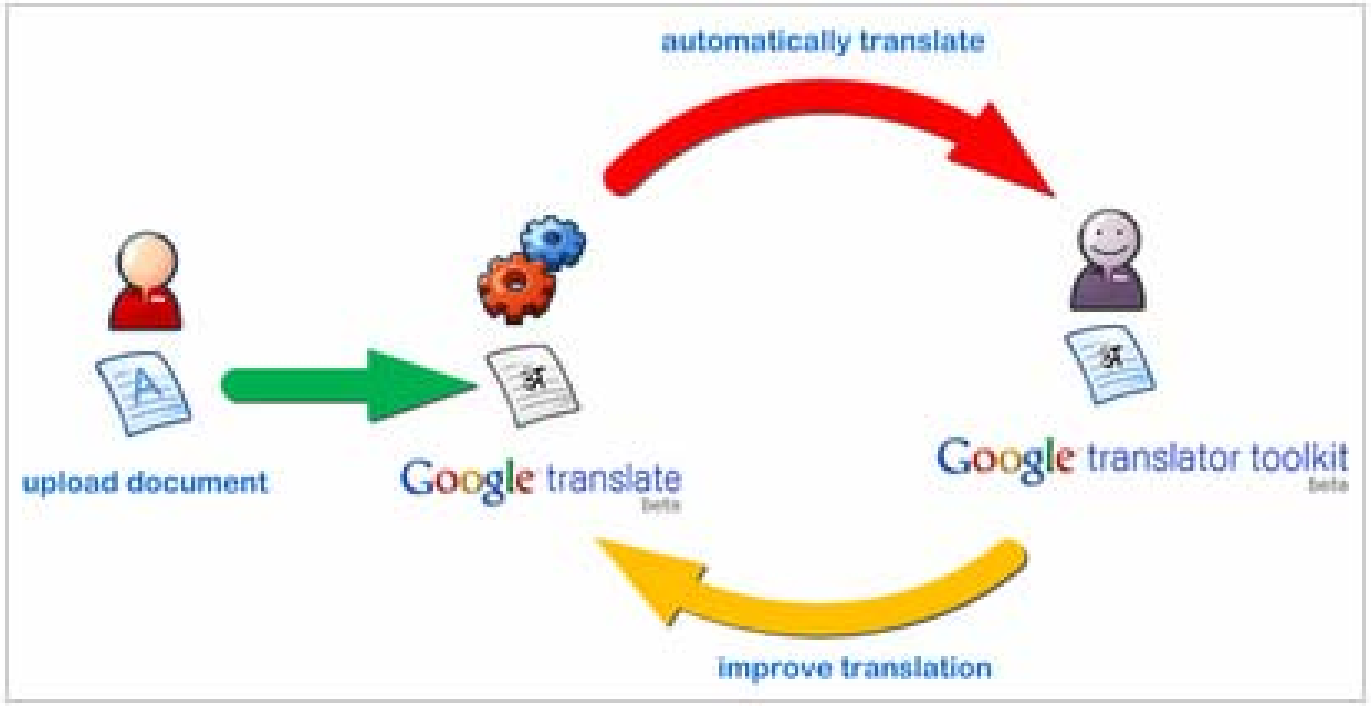
6/09/2009 09:00:00 AM

At Google, we consider translation a key part of making information universally accessible to everyone around the world. While we think Google Translate, our automatic translation system, is pretty neat, sometimes machine translation could use a human touch. Yesterday, we launched Google Translator Toolkit, a powerful but easy-to-use editor that enables translators to bring that human touch to machine translation.

For example, if an Arabic-speaking reader wants to translate a Wikipedia™ article into Arabic, she loads the article into Translator Toolkit, corrects the automatic translation, and clicks publish. By using Translator Toolkit's bag of tools — translation search, bilingual dictionaries, and ratings, she translates and publishes the article faster and better into Arabic. The Translator Toolkit is integrated with Wikipedia, making it easy to publish translated articles. Best of all, our automatic translation system "learns" from her corrections, creating a virtuous cycle that can help translate content into 47 languages, or over 98% of the world's Internet population.

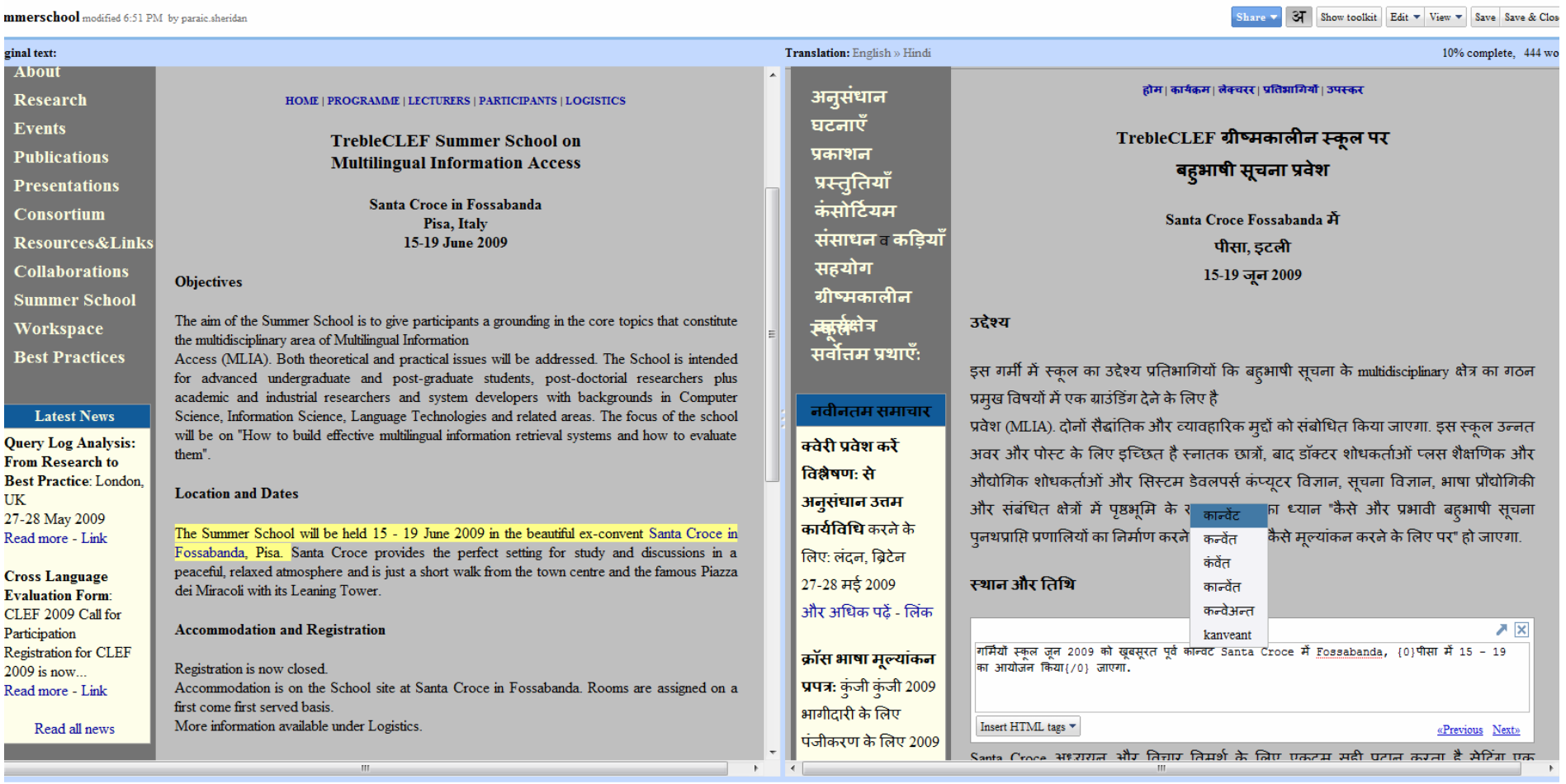
http://googleblog.blogspot.com/2009/06/translating-worlds-information-with.html

‘Learn’ from Corrections…



- Automatically translate document using Google MT
- Translator ‘Post-Edits’ (Corrects) Translation
- Google translation ‘learns’ from corrected sentences (SMT)

E.g. Translate & Correct Summer School Page



World Wide Web - Summary

- Much content/services being offered in multiple languages – community translation a big Web2.0 trend.
- Google cross-language search / Google Translate to find information in a foreign language.
- What about under-resourced languages/areas? African languages? [e.g. Translators without Borders]
- What are the specific domains where ‘generic’ translation is not good enough?

Translating to Overcome Information Poverty

- Since early 2007, **Asia Online** has been working to build what it believes to be the world's' Largest Literacy Project.

- Asia Online is addressing the issue of “information poverty” that millions of non-English speaking Asians face.

- The business delivers large portions of the world's English language Internet content to the whole of Asia in the respective local languages: Wikipedia, World Fact Book, open courseware from MIT, Carnegie Mellon University, etc.

http://www.asiaonline.net

Centre for Next Generation Localisation

Enterprise Search

Innovation & Consolidation in Enterprise Search

- ‘Enterprise Search’ is one part of the larger Enterprise ‘Knowledge Management’ ‘Information Management’ endeavour.
- Enterprise search market is a constant mix of small innovative companies with focused (often vertical) offerings and large companies trying to consolidate and extend their reach within enterprise customers.
- The leads to an active stream of acquisitions in the space… it’s often hard to keep up!

Cross-Language roll-up at SAP

- **Business Objects Acquires Inxight Software**

  Posted May 25, 2007

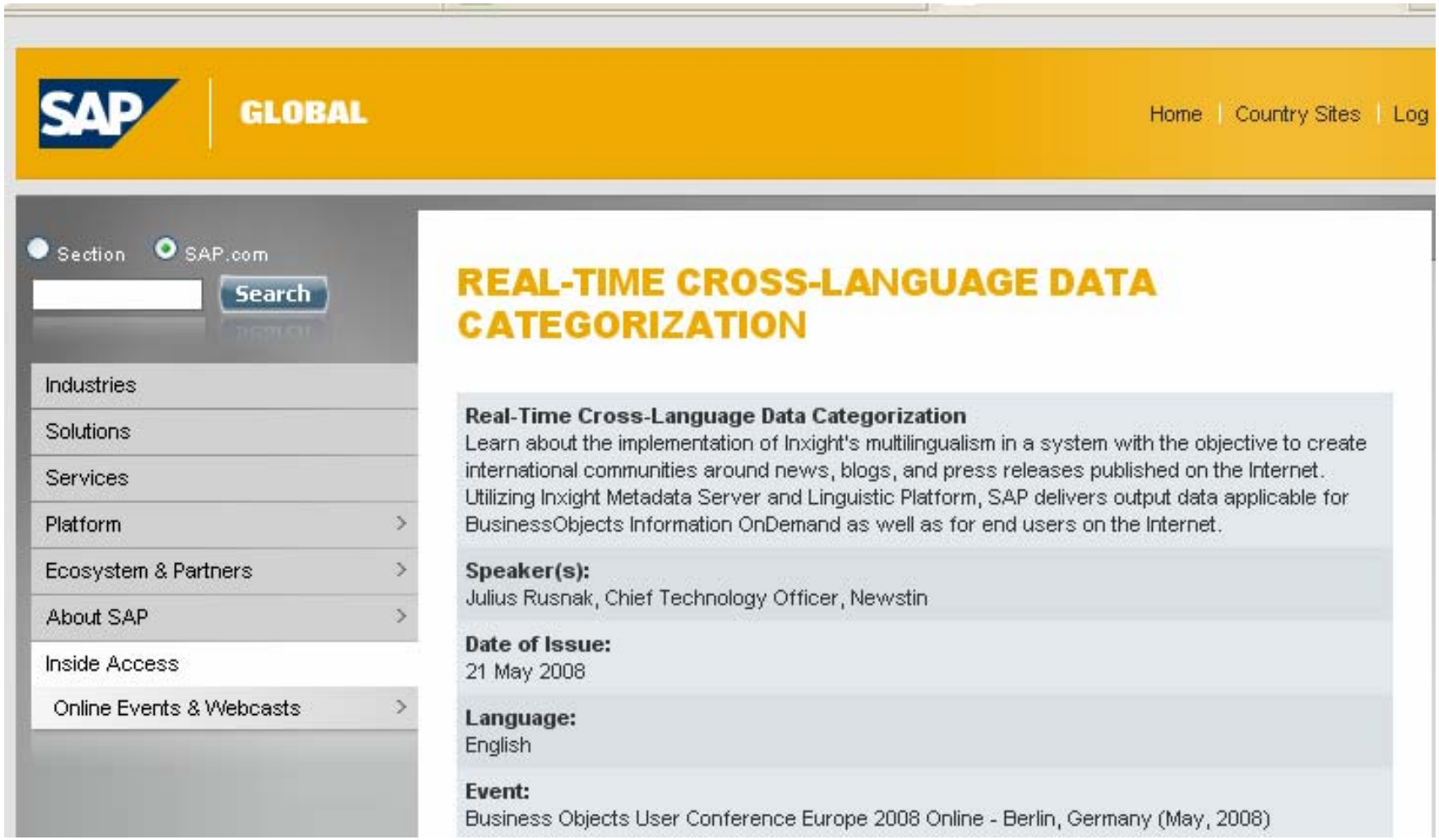
  Business Objects, a provider of business intelligence solutions, has entered into a definitive agreement to acquire Inxight Software Inc. Inxight is a provider of software solutions for unstructured information discovery, including text analytics, federated search, and data visualization. Financial details of the transaction, which is expected to close in July of 2007, were not disclosed.

- **SAP to Acquire Business Objects in Friendly Takeover**

  October 7, 2007

  WALLDORF, Germany & PARIS--(BUSINESS WIRE)--SAP AG (NYSE:SAP) and Business Objects S.A. (Nasdaq:BOBJ) (Euronext Paris ISIN code: FR0004026250 – BOB) today announced that the companies have reached an agreement that will bring together two of the information technology industry's leaders, resulting in an unmatched offering The transaction volume taking into account the transaction costs will be **slightly above €4.8 billion**.

SAP Cross-Language offering (InXight)



Search (incl. CLIR) roll-up at Microsoft

- **FAST Acquires Convera's RetrievalWare Business**

  Posted On April 9, 2007

  Fast Search & Transfer (FAST; www.fastsearch.com) has announced its acquisition of Convera Corp.'s (www.convera.com) RetrievalWare business, a knowledge platform for information retrieval and search. The acquisition, **priced at $23 million**…

- **Microsoft Acquires FAST**

  Software giant Microsoft (News - Alert) (News - Alert) announced its plans for acquiring Norwegian based enterprise search engine company, Fast Search & Transfer (FAST). Microsoft proposed a deal of about **$1.2 billion in cash** and the companies announced the merger on Tuesday.

- **Microsoft: The Latest Disruptor In Enterprise Search**

  November 6, 2007

  Microsoft makes a very aggressive move into the enterprise search market by introducing the free download of Microsoft Search Server Express 2008 (MSSE), which is sure to disrupt the already confusing search space. Though free, MSSE has a substantial set of capabilities, most notably its connectivity to many intranet- and Internet-based resources, including Microsoft Exchange, IBM's Lotus Notes, Microsoft Office SharePoint Server 2007, and even EMC's Documentum and IBM's FileNet enterprise content management systems.

  (Source: Forrester)

Autonomy acquires Verity for Market Lead

"Autonomy's acquisition of Verity creates a formidable force in enterprise search and information retrieval, and will dwarf competitors such as Convera, Endeca, and FAST.

But the enterprise search market has failed to reach more than a billion dollars because competitors like Google, IBM, and Microsoft continue to **commoditize basic search**. If Autonomy can overcome key challenges integrating the two companies, it will be well-positioned in the high-end market for unstructured information retrieval and analysis.

However, this market will be contested by the likes of IBM, which hopes to capitalize on core search technology for solutions such as root-cause analysis, eDiscovery, fraud detection, compliance, market intelligence, and others. "

Source: Forrester

The State of Play 2008: Forrester

- **The Forrester Wave™: Enterprise Search, Q2 2008**
  Autonomy, Endeca, FAST, And Vivisimo Lead A Competitive Market

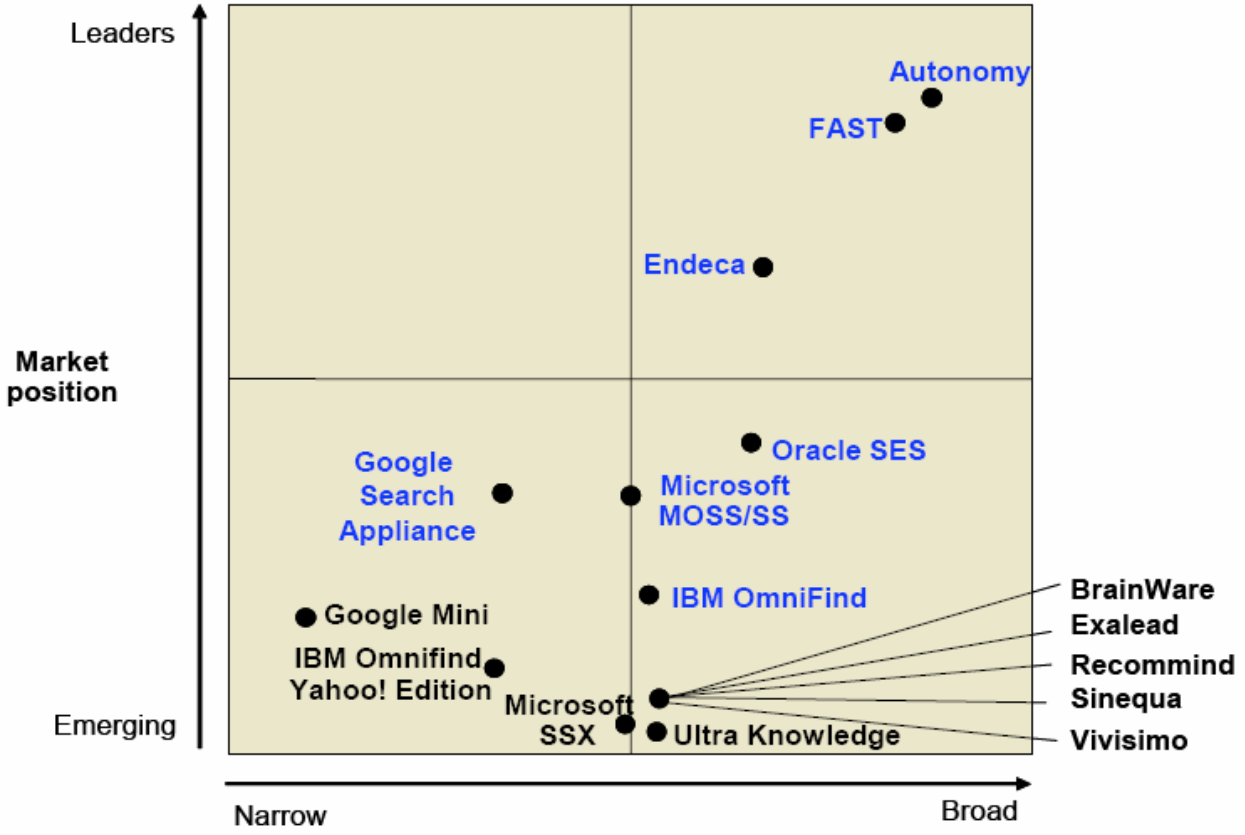
  Forrester evaluated enterprise search vendors against 147 criteria and found that Autonomy, Endeca, FAST (a Microsoft subsidiary), and Vivisimo lead the market with impressive technical capabilities and a diverse customer base pushing for further innovation.

  Over the past two years, IBM, Microsoft, and Oracle have all recognized their potential advantage in the enterprise search market. Not only do they have deep financial resources to support their products, but they also can win customers who seek to natively connect and integrate search with existing IBM, Microsoft, or Oracle applications and content repositories. Yet these vendors are still digesting acquisitions and building search features that customers want now.

  Meanwhile, Recommind and InQuira deliver distinctive solutions for targeted market segments — legal and customer service, respectively— while Coveo is repositioning itself due to industry shakeups.

The State of Play 2008: Ovum

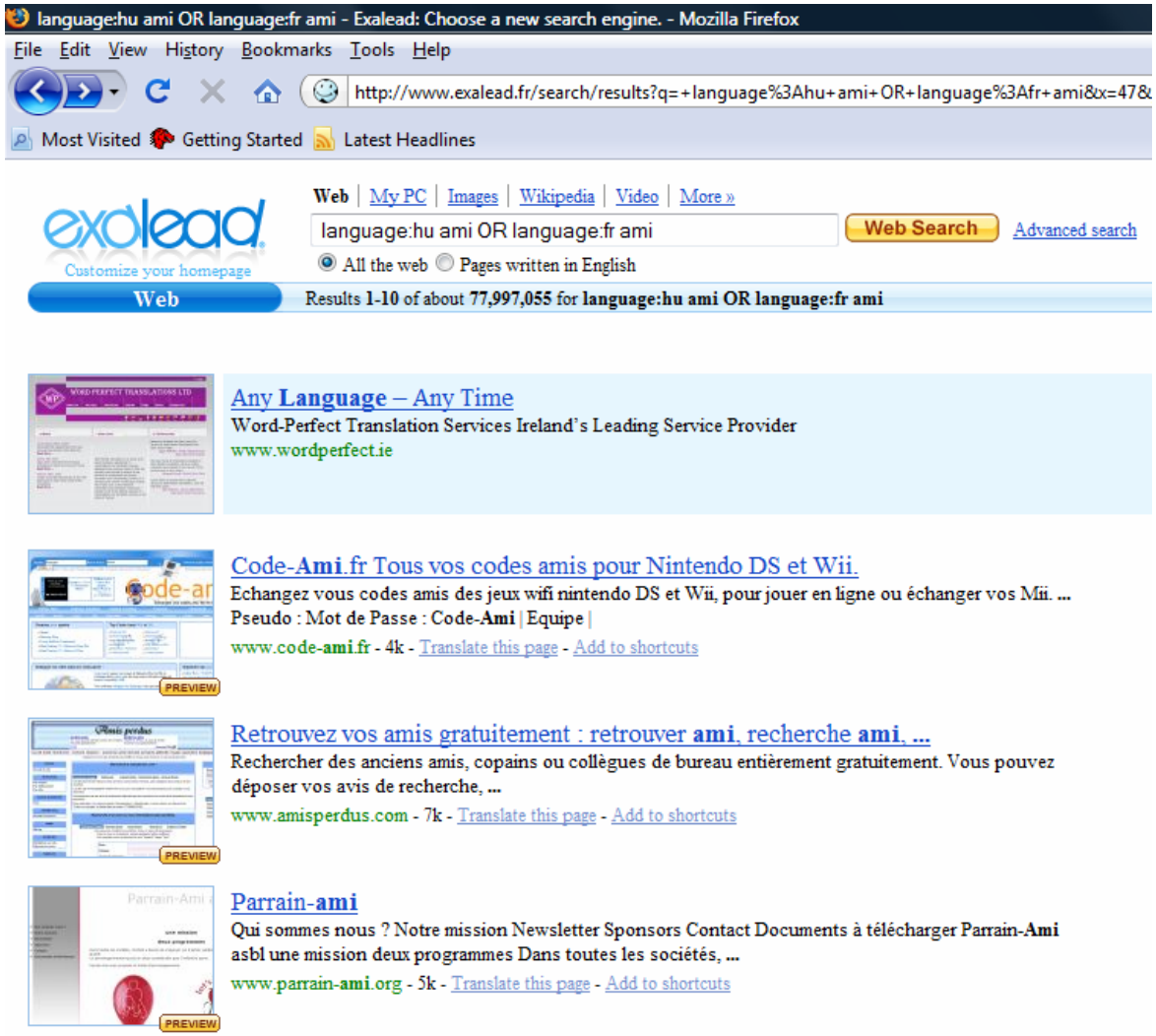
Figure 1 **Enterprise information access solutions**



Source: Ovum

'The future of search' Feb 2008 – www.ovum.com

Exalead Multilingual Search



Market Summary - Overall

- Enterprise Search is a highly competitive market space with several large players but a consistent presence from small focused companies.

- Enterprise Search is part of the larger Enterprise Information ecosystem – ERP, CRM, ECM etc. Customers expect that Search systems (incl. CLIR) will integrate with that broader system framework.

- Small/new players must find a niche area and show consistent competitive advantage to prosper (e.g. Endeca)

Market Summary - MLIA

- Multilingual Information Access **does** exist as an offering from all major Enterprise Search players, often from technology acquisitions from smaller companies.
  - Google Search + Google Translate (CLIR)
  - SAP (through InXight)
  - FAST (through Convera)
  - Autonomy / Verity
- **New** MLIA offerings in the Enterprise space must identify unmet user/customer needs and also provide a differentiated solution over the current offerings.

Enterprise MLIA – Typical Scenarios

- Global company ‘Knowledge Base’
  - Large global companies
  - Central ‘knowledge repositories’, best practices, case studies, FAQs, etc.
  - Multiple languages

- Mergers & Acquisitions
  - Company A from country 1 acquires Company B from country 2
  - All documentation, contracts, etc. from acquired company are in foreign language

Experiences with TextWise CINDOR search

- Use Case 1: Multinational CPG company. Bio-tech R&D group interested in Japanese Patent Search.
- Use Case 2: Hong Kong investment group interested in publicising investment opportunities to English speaking investors.
- Use Case 3: Service for search/discovery of business & financial news in ‘emerging markets’
- Use Case 4: Government customers.

Screenshots: CINDOR / Convera

Screenshots: CINDOR / Inxight

Lessons I learned…

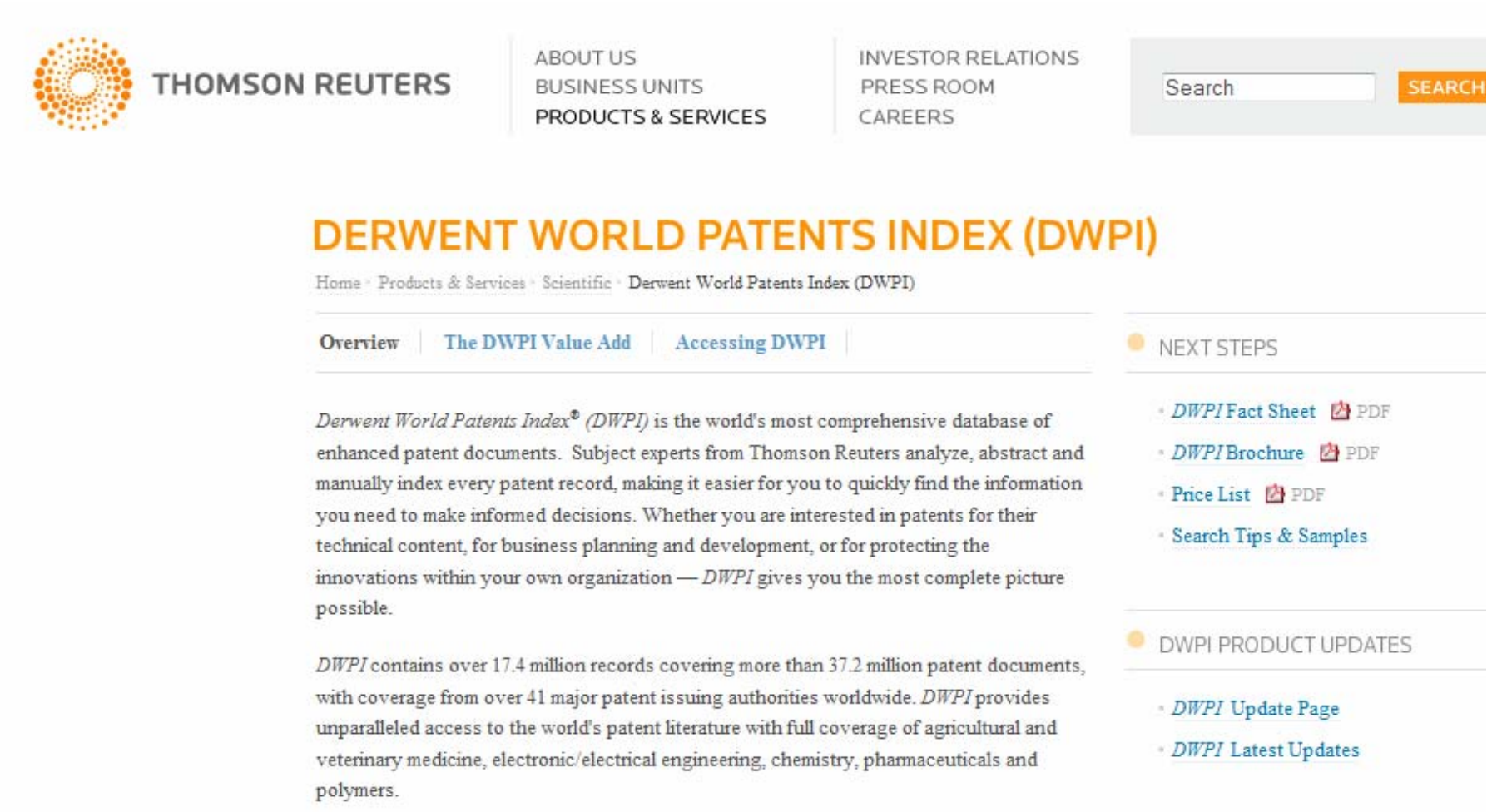
- FROM THE USER (CUSTOMER) PERSPECTIVE…
- Cross-language is a ‘feature’ – should not be a ‘separate’ system with separate user interface etc. Must be integrated with enterprise information/search environment.
- Finding foreign language documents is useless without also making those documents accessible to the user (translation)
- Sometimes bulk translation is the easier/required option.

What are some CLIR Applications now?

- Where is there high-value content in multiple languages that people need access to with specialised translation?

- E.G. Patent Search
  - Patent searches for ‘prior art’ must be thorough to establish novelty. A missed publication can later invalidate a patent.
  - Now, over 50% of all newly issued patents are in a CJK language (Chinese, Japanese, Korean).
  - Perhaps (?) a combination of MT/HT can efficiently translate foreign-language results… technical language/terminology.

Derwent – Manual indexing of world patents



THOMSON REUTERS

ABOUT US
BUSINESS UNITS
PRODUCTS & SERVICES
INVESTOR RELATIONS
PRESS ROOM
CAREERS

Search SEARCH

DERWENT WORLD PATENTS INDEX (DWPI)

Home › Products & Services › Scientific › Derwent World Patents Index (DWPI)

Overview | The DWPI Value Add | Accessing DWPI

Derwent World Patents Index® *(DWPI)* is the world's most comprehensive database of enhanced patent documents. Subject experts from Thomson Reuters analyze, abstract and manually index every patent record, making it easier for you to quickly find the information you need to make informed decisions. Whether you are interested in patents for their technical content, for business planning and development, or for protecting the innovations within your own organization — *DWPI* gives you the most complete picture possible.

DWPI contains over 17.4 million records covering more than 37.2 million patent documents, with coverage from over 41 major patent issuing authorities worldwide. *DWPI* provides unparalleled access to the world's patent literature with full coverage of agricultural and veterinary medicine, electronic/electrical engineering, chemistry, pharmaceuticals and polymers.

NEXT STEPS

- *DWPI* Fact Sheet PDF
- *DWPI* Brochure PDF
- Price List PDF
- Search Tips & Samples

DWPI PRODUCT UPDATES

- *DWPI* Update Page
- *DWPI* Latest Updates

http://www.thomsonreuters.com/products_services/scientific/DWPI

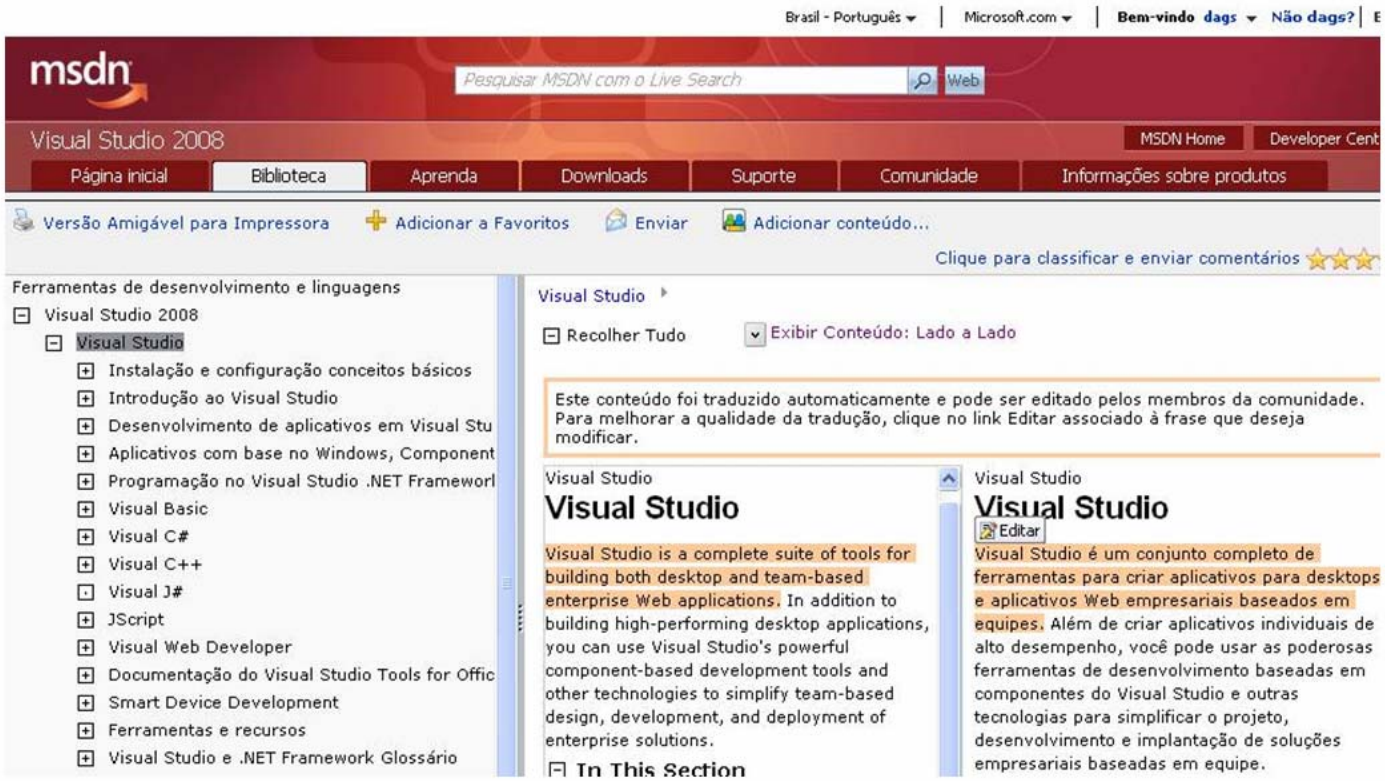
What are some CLIR Applications now?

- Where is there content/customers where (tailored) automated translation is good enough for use?
- Technical Customer Service / Forums?
  - Microsoft using MT for technical developer documentation online.
- Increasing interest in customer support applications.
  - Including search of knowledge bases, logs, faqs, etc.

Microsoft MT of Developer Documentation

Raw-MT piloting for Office Developer Documentation

- Translation wiki for SDKs



[Source: Microsoft]

Summary

- Major economic shifts East – massive populations in India & China who don't speak English.
- WWW increasingly multilingual. Combinations of technology and community to bridge language barriers.
- Technology reaching acceptance (e.g. MT) either
  - In very high value applications (e.g. patent search)
  - In lower barrier 'acceptance' scenarios – online tech help that wouldn't otherwise be available.
- Potential for novel combinations of technology/people to open up new translation markets!

Centre for Next Generation Localisation

Thank You! Questions?