Overview

- Centre for Next Generation Localisation CNGL
- What is Localisation?
- Language Technology and Localisation: Translation Memories (TMs)
- Machine Translation (MT - different approaches: RBMT, SMT, EBMT, TM, Hybrid)
- MT in Localisation: Symantec and Microsoft
- MT Research at CNGL
Overview

- Centre for Next Generation Localisation (CNGL)
- What is Localisation?
- Language Technology and Localisation: Translation Memories (TMs)
- Machine Translation (MT - different approaches: RBMT, SMT, EBMT, TM, Hybrid)
- MT in Localisation: Symantec and Microsoft
- MT research at CNGL
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, Traslan, Speech Storm
- 100 people (mostly research staff: PhD students and PostDocs)

http://www.cngl.ie/
Overview

- Centre for Next Generation Localisation CNGL
- **What is Localisation?**
- Language Technology and Localisation: Translation Memories (TM)
- Machine Translation (MT - different approaches: RBMT, SMT, EBMT, TM, Hybrid)
- MT in Localisation: Symantec and Microsoft
- MT research at CNGL
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: information dissemination and sharing
Localisation (traditionally ICT)

- Software
- Interfaces
- Web Pages
- User Manuals
- Technical Documentation
Localisation (more generally)

- Product Documentation
  - Automotive
  - Pharma
  - Consumer electronics

- Services
  - Commercial
  - Governmental
  - Educational

- Media (Global Content Distribution Industry)
  - Print/Text
  - Audio/Video
Example: Services

Philadelphia ab 469 €
und über 40 weitere Ziele für Frühbucher

⇒ Lob=Lufthansa Ticket Service Change

Grenzenlose Möglichkeiten:
Urlaubs buchen auf
lufthansa.com

Lufthansa bietet in Kooperation mit
DERTOUR jetzt auch komplette
Urlaubsangebote, z.B. an die
Westküste der USA.

Buchen Sie Flüge für Ihre
Familie und Freunde

Sie können über unsere Online
Flugbuchung auch dann Flüge
buchen, wenn Sie selbst gar nicht
mitfliegen.

Fluginformationen

Wir informieren Sie zum aktuellen Flugbetrieb.
Example: Services

Booking  Check-In
Flight  Hotel  Car

Japan  English  Other countries  Help & Contact

Lufthansa

Example: Services

99 € ~

条件を詳しくご確認する
優雅な旅路の予約
アワード予約

クイックリンク
タイムテーブル
機内食
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
日本のご案内
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
企業の旅行のご案内
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

betterFly - ヨーロッパ往復のWEB価格
スーパー便利

¥510,000 ~

条件を詳しくご確認する
優雅な旅路の予約
アワード予約

クイックリンク
タイムテーブル
機内食
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
日本のご案内
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
企業の旅行のご案内
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

ルフトハンザ・コンシュルン
企業のご案内
ルフトハンザの航空

サイトマップ
Example: Software
Example: Software

**What's in the Box**

- AC power cord
- 45W MagSafe Power Adapter

**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.

---

**Chapter 1  Ready, Set Up, Go**
Example: Magazines
Localisation is Important

- Large part of Microsoft world wide revenue 2008 from Ireland (large part localisation, marketing and sales, customer support)
- 51% of Google’s revenue in Q1 2008 from international sources
- Facebook 52.2 (2007) → 132 (2008) Million visitors … German, Spanish, French
Overview

- Centre for Next Generation Localisation CNGL
- What is Localisation?
- Language Technology and Localisation: Translation Memories (TMs)
- Machine Translation (MT - different approaches: RBMT, SMT, EBMT, TM, Hybrid)
- MT in Localisation: Symantec and Microsoft
- MT research at CNGL
A “Typical” Localisation Task

- Microsoft - “Bulk”/“Enterprise” Localisation
- Microsoft Office 2007
- Volume
  - Software:
    - 2 million words (help files, menus, …)
  - Content:
    - 3 million words end user (manuals, …)
    - 4 million words developer (technical documentation, …)
- Languages
  - 30+ full localisation
  - Up to 100 with local language programme
- Release schedule: 10 languages sim-ship
- And that’s just Office, doesn’t include Office Online 2007
Localisation Workflows are Complex
Today’s Localisation Industry is Highly Automated

- Global Team Management Systems
- Global Content Management Systems
- Global Workflow Systems
- Language Technology
- Quality Control & Testing
- Taking Apart and Building (Target) Applications
Language Technology and Today’s Localisation

- Text encoding standards (Unicode, UTF, …)
- Font management
- Mark-up standards
- Tokenisation
- Word counting
- Alignment
- Term extraction
- Terminology management
- Electronic dictionaries/thesauri
- Translation Memories (TMs)
- Controlled language (checkers)
- Translators’ workbenches (Computer-Aided Translation CAT tools)
- Spell checkers
- Grammar checkers
Language Technology and Today’s Localisation

- Text encoding standards (Unicode, UTF, …)
- Font management
- Mark-up standards
- Tokenisation
- Word counting
- Alignment
- Term extraction
- Terminology management
- Electronic dictionaries/thesauri
- Translation Memories (TMs)
- **Machine Translation (MT)**
- Controlled language (checkers)
- Translators’ workbenches (Computer-Aided Translation CAT tools)
- Spell checkers
- Grammar checkers
Language Technology and Today’s Localisation

- Text encoding standards (Unicode, UTF, …)
- Font management
- Mark-up standards
- Tokenisation
- Word counting
- Alignment
- Term extraction
- Terminology management
- Electronic dictionaries/thesauri
- Translation Memories (TMs)
- Machine Translation (MT)
- Controlled language (checkers)
- Translators’ workbenches (Computer-Aided Translation CAT tools)
- Spell checkers
- Grammar checkers
Translation Memories

- Database of previously translated text
- Sentence by sentence

- Compare and merge presentations ⇔ Präsentationen vergleichen und zusammen fügen
- Spelling and Grammar ⇔ Rechtschreibung und Grammatik
- ... ⇔ ...

```
Input S
If <S,T> in TM
    Then return T
Else human translator
```
Translation Memories

- When you translate a sentence, the TM checks whether the sentence has been translated and retrieves the translation.

- “Simple” look-up
- Needs sophisticated search …!
- Saves time and money
- Recycling – Leveraging
- Up to 30% in technical domains (software: Mozilla Firefox version 1.6 to version 2, ...)

```
Input S
If <S,T> in TM
  Then return T
Else human translator
```
Translation Memories

- Full match in context: +++ $S_{i-1}, S_i, S_{i+1} \Leftrightarrow T_{i-1}, T_i, T_{i+1}$
- Full match: ++
- Partial match: +
- Repetition: + $S_1 \ldots S \ldots S \ldots S_1 \ldots S_n$

- Update TM: add new translations!

- TM is a simple Example-Based MT method
- Operates on the level of segment (sentence, string …)
Translation Memories

- Can be very effective
- Very valuable ⇒ core technology in Localisation
- Usually combined with other technologies:
  - Terminology management systems etc.
Overview

- Centre for Next Generation Localisation CNGL
- What is Localisation?
- Language Technology and Localisation: Translation Memories (TMs)
- Machine Translation (MT - different approaches: RBMT, SMT, EBMT, TM, Hybrid)
- MT in Localisation: Symantec and Microsoft
- MT research at CNGL
Machine Translation

- RBMT
- SMT
- EBMT
- TMs
- Hybrid
Machine Translation

- RBMT: Rule Based MT
- SMT: Statistical MT
- EBMT: Example Based MT
- TMs: Translation Memories
- Hybrid: A mix … motivated by the advantages and disadvantages of the above
RBMT: Rule-Based MT

- **Definition (to a first approximation):**
  - Linguistic knowledge (lex, morph, syn, sem) coded in translation (analysis, generation) rules
  - Vauquois triangle: direct, transfer, interlingua

- **Requires:**
  - Rules, transfer components, parsers, generators

- **Pros/Cons:**
  - Promises to deliver high quality translation (sophisticated linguistically motivated analyses)
  - Traditionally (!) hand-crafted: knowledge acquisition bottleneck
  - Difficult to scale to real data, many language pairs
  - Noise tends to multiply: analysis $\Rightarrow$ transfer $\Rightarrow$ generation
  - Direct method insufficient, interlingua elusive …
SMT: Statistical MT

Definition (to a first …):
- Noisy channel model + Bayesian inversion
- Uses only information from strings (pure, original version, …)

Requires:
- Bitext, alignments, translation model, language model, decoder

Pros/Cons:
- Simple, easy to scale, software open domain
- Bitext not always available
- Limited reordering
- Alignment and decoding intractable: heuristics
- Performance plateau …?!
EBMT: Example-Based MT

- **Definition (to a first …):**
  - Remember previously seen translation examples/units
  - Recombine them during translation (lazy learner – MBL)

- **Requires:**
  - Bitext, procedure to establish translation units, alignments, translation model, decoder (≈ linguistically motivated SMT)

- **Pros/Cons:**
  - Translation units linguistically motivated
  - Simple, easy to scale
  - Software not widely available
  - Bitext not always available
  - Sometimes limited reordering
  - Performance =< SMT (fewer translation unit types in EBMT)
TMs: Translation Memories

Definition (to a first …):
- Database of previously translated sentences/strings
- Look-up

Requires:
- Bitext, efficient storage and retrieval

Pros/Cons:
- Effective for certain text types (technical documentation, …)
- Simple, easy to scale
- Software widely available
- Bitext not always available
- Will not perform well on varied unseen data
Hybrid MT

Aim:
- Combine advantages of (some) of RBMT, SMT, EBMT, TM …
- Avoiding the disadvantages …

Hybrid Approaches Centred on SMT
- System Combinations (Multi-Engine MT)
- Learning Resources for (Statistical) RBMT
Hybrid Approaches Centred on SMT

- Enrich SMT models with linguistically motivated information (over and above what is in the surface strings)
  - POS-tag, morphological, syntactic and semantic information in alignment and language models
  - Factored models (Koehn et al.)
  - Phrase-based SMT (Och et al.)
  - Hierarchical Phrase-Based SMT (Chiang)
  - Reordering (Collins et al.)
  - Syntactic units (from e.g. EBMT, parsing etc.) for alignment and as phrase types in translation table (Way et al.)
- ...
System Combinations

Aim:
- Combine advantages of (some) of RBMT, SMT, EBMT, TM ...
- Two ways:
  - treat systems as black boxes
  - Use translation units from one in other

- Multi-engine MT (parallel): combine multiple in to single output

- Multi-engine MT (sequential): statistical post-editing (RBMT $\Rightarrow$ SMT, or TM $\Rightarrow$ MT, …)

- Multi-unit/-phrase MT: SMT, EBMT, parser-based phrases in e.g. PB-SMT
Aim:

- Address knowledge acquisition bottleneck …
- Use sophisticated linguistic knowledge
- Disambiguation (select best translation among many, similar to parse selection in statistical parsing)

- Learn probabilistic grammars (HPSG, LFG, CCG) parsers, generators and transfer rules from treebanks and parse-annotated bitext
- Riezler & Maxwell, Alon Lavie, Bojars, Sanches-Martinez

- Stochastic transduction grammars (Dekai Wu)
- Language models to chose between (traditional) RBMT output
Overview

- Centre for Next Generation Localisation CNGL
- What is Localisation?
- Language Technology and Localisation: Translation Memories (TMs)
- Machine Translation (MT - different approaches: RBMT, SMT, EBMT, TM, Hybrid)
- MT in Localisation: Symantec and Microsoft
- MT research at CNGL
Symantec

- Data-Security and Integrity
- Anti-Virus Software (Norton and others…)

- Language Services Research and Development Dublin
- Localise into FIGS
  - French
  - Italian
  - German
  - Spanish
  - Japanese
  - Simplified Chinese
Symantec

- Use a number of technologies:
  - Controlled Language
  - Terminology Management
  - Translation Memories
  - Machine Translation
    - RBMT: Systran
    - SMT: PBSMT (Moses)
    - Statistical post-editing
  - Technical Support + Products (Norton AntiVirus)
Symantec (Controlled Language - CL)

- CL is a subset of a natural language that uses a restricted grammar and a restricted vocabulary (technical domain)

- Makes source content clearer and less ambiguous

- Improves comprehensibility and (machine-) translatability of source content

- Example: Caterpillar Technical English in the mid 1990s (142 rules and more than 70000 terms)
Symantec (Controlled Language - CL)

- Rule 54: Check the spelling.
- Rule 22: Do not coordinate verbs or verbal phrases when the verbs do not have the same transitivity.
- Rule 19: Do not use pronouns that have no specific referent.
- Rule 35: Avoid unusual punctuation.
- Rule 41: Avoid embedded clauses introduced by commas or dashes.
- Rule 1: Avoid ambiguous coordinations by repeating the head noun, or by changing the word order.
- Rule 4: Do not omit hyphens in Noun + Adjective or Noun + Past Participle structures.
- Rule 5: Do not use more than 25 words per sentence.
- Rule 48: Use a question mark only at the end of a direct question.

Most effective ...
Symantec (Controlled Language - CL)

- Acrolinx CL management

- acrocheck is used to ensure that source content is compliant
  - Compliance with grammar rules
  - Compliance with terminology
  - Compliance with rules based on corporate guidelines (20+)
  - Compliance with MT-specific rules (6+)
Symantec (Controlled Language - CL)

- MT specific CL rules (Systran)
  - Do not use slashes to list lexical items
  - Repeat the head noun in ambiguous coordinated structures
  - Use a hyphen to indicate the first part of a compound
  - Use articles in specific contexts (for disambiguation)
  - Keep both parts of a two-part verb together
  - Use "could" with "if"
  - Avoid parenthetical expressions in the middle of a sentence
Symantec (Controlled Language - CL)

- Keep both parts of a two-part verb together

- This document gives directions to turn email scanning on or off.
  - Dieses Dokument gibt Richtungen zum Umdrehung E-Mail-Prüfung an oder weg.
  - Ce document donne des directions à l'analyse du courrier électronique de tour en fonction ou hors fonction.

- This document gives directions to turn on or turn off email scanning.
  - Dieses Dokument gibt Richtungen, E-Mail-Prüfung zu aktivieren oder zu deaktivieren.
  - Ce document donne des directions pour activer ou désactiver l'analyse du courrier électronique.
Symantec (Machine Translation)

- Machine Translation
  - RBMT: Systran (heavily tuned and optimised to domain & terminology)
  - SMT: PBSMT (Moses)
  - Statistical post-editing (SPE architecture)
Symantec (Machine Translation)

- Overall productivity increases
  - Controlled language
  - Terminology
  - Systran
  - PBSMT (SPE)
- Depends on language
- 50% for French, Spanish and Brazilian
- 30% for Japanese and simplified Chinese
- Compared to without

Main challenges:
- Sloppy source
- Getting terminology right (right pain in human post-editing)
Microsoft

- Office/Office Online
- Localisation for EMEA
  - Europe
  - Middle East
  - Africa

- Localisation Language Technology Dublin
Microsoft

Use a number of technologies:

- Controlled Language (lightweight + integrated into authoring)
- Terminology Management
- Translation Memories
- Machine Translation
  - MSR-MT: syntax-enhanced SMT (dependency treelets, Quirk et al. 2005)
  - Trained on MT TMs

- Technical Support + Products (Office, Office Online)
Microsoft

- Recycling (TMs) - 35%
- Export no-match to TMX
- Machine translate TMX
- Post-edit (human) or use raw (for some technical documentation and SDKs, where source/MT is better than just source)
Microsoft

Productivity Gains 5% - 30%:

- Czech 6%
- French 14%
- Brazilian 20%
- Danish 28%

Main challenges:

- Terminology
- Mark-up, internal formatting
- Quality of source text – controlled language
- Translators and post-editing
- Standards (lack of …)
Overview

- Centre for Next Generation Localisation CNGL
- What is Localisation?
- Language Technology and Localisation: Translation Memories (TMs)
- Machine Translation (MT - different approaches: RBMT, SMT, EBMT, TM, Hybrid)
- MT in Localisation: Symantec and Microsoft
- MT research at CNGL
Three Challenges

Volume:

- rapidly increasing amount of digital content to be localised into more languages

Access:

- print/paper, full screen/keyboard vs. mobile devices, hands-busy, eyes-busy, access on the go
- off-line vs. instant/real-time

Personalisation:

- coarse-grained “locale” vs. individualised information
The Localisation Cube

- Access
- Personalisation
- Volume
“Traditional” Enterprise Localisation

Access

Personalisation

Volume
Vision Next Generation Localisation CNGL

Access  →  Personalisation  →  Volume
Vision Next Generation Localisation CNGL

Access

- Modality
- Language Technology, Speech

Personalisation

- Digital Content Management
- Adaptive Hypermedia, IR

Volume

- Automation, Language Technology, MT
- Crowd Sourcing
Organisation of CNGL Research

Adaptive Hypermedia IR

Design Standards
Workflows
Crowd Sourcing
Evaluation

Next Generation
Localisation

Personalised Localisation

Enterprise Localisation

Digital Content Management

Integrated Language Technologies

MT Speech NLP

Systems Framework

Software Eng.
HCI
Demonstrators Evaluation

Software Eng.
MT Research at CNGL

ILT Integrated Language Technologies .................................................................................. 35
ILT1 Core MT Systems (Prof. Andy Way) ........................................................................ 35
ILT1.1 Core MT Engines: Incorporating Syntax into SMT Systems .................................... 35
ILT1.2 Core MT Engines: Improved Hybrid Systems ......................................................... 35
ILT1.3 Core MT Engines: Scaling More Linguistically Sophisticated Systems .................. 36
ILT1.4 Core MT Engines: Probabilistic Transfer ............................................................... 36
ILT1.5 Core MT Engines: Tuning Machine Translation ..................................................... 37
ILT1.6 Core MT Engines: General Alignment Models ....................................................... 37
ILT1.7 Core MT Engines: Automatic MT Evaluation ......................................................... 38
ILT1.8 Core MT Engines: Controlled Language .................................................................. 38
ILT1.9 Core MT Engines: Patients with Limited English .................................................... 38
ILT2 Speech Technologies and MT (Prof. Julie Berndsen) .................................................. 39
ILT2.1 Speech & MT: Phonetic Parameters/Voice Characteristics ...................................... 39
ILT2.2 Speech & MT: Muster Speech Engine: Synthesis ..................................................... 39
ILT2.3 Speech & MT: Muster Speech Engine: ASR .......................................................... 40
ILT2.4 Speech & MT: MT/Synthesis Integration ............................................................... 40
ILT2.5 Speech & MT: MT/Recognition Integration ........................................................... 41
ILT3 Text Classification and Automatic Labelling (Dr. Carl Vogel) ..................................... 41
ILT3.1 Text Classification and Automatic Labelling for Next Generation Localisation: Automatic Dependency Annotation ........................................................................ 41
ILT3.2: Automatic Localisation Metadata Annotation ....................................................... 42
ILT3.3: Multilingual Text Type and Genre Classification .................................................... 42
## MT Research at CNGL

<table>
<thead>
<tr>
<th>ILT Integrated Language Technologies</th>
<th>35</th>
</tr>
</thead>
<tbody>
<tr>
<td>ILT1 Core MT Systems (Prof. Andy Way)</td>
<td>35</td>
</tr>
<tr>
<td>ILT1.1 Core MT Engines: Incorporating Syntax into SMT Systems</td>
<td>35</td>
</tr>
<tr>
<td>ILT1.2 Core MT Engines: Improved Hybrid Systems</td>
<td>35</td>
</tr>
<tr>
<td>ILT1.3 Core MT Engines: Scaling More Linguistically Sophisticated Systems</td>
<td>36</td>
</tr>
<tr>
<td>ILT1.4 Core MT Engines: Probabilistic Transfer</td>
<td>36</td>
</tr>
<tr>
<td>ILT1.5 Core MT Engines: Tuning Machine Translation</td>
<td>37</td>
</tr>
<tr>
<td>ILT1.6 Core MT Engines: General Alignment Models</td>
<td>37</td>
</tr>
<tr>
<td>ILT1.7 Core MT Engines: Automatic MT Evaluation</td>
<td>38</td>
</tr>
<tr>
<td>ILT1.8 Core MT Engines: Controlled Language</td>
<td>38</td>
</tr>
<tr>
<td>ILT1.9 Core MT Engines: Patients with Limited English</td>
<td>38</td>
</tr>
<tr>
<td>ILT2 Speech Technologies and MT (Prof. Julie Berndsen)</td>
<td>39</td>
</tr>
<tr>
<td>ILT2.1 Speech &amp; MT: Phonetic Parameters/Voice Characteristics</td>
<td>39</td>
</tr>
<tr>
<td>ILT2.2 Speech &amp; MT: Muster Speech Engine: Synthesis</td>
<td>39</td>
</tr>
<tr>
<td>ILT2.3 Speech &amp; MT: Muster Speech Engine: ASR</td>
<td>40</td>
</tr>
<tr>
<td>ILT2.4 Speech &amp; MT: MT/Synthesis Integration</td>
<td>40</td>
</tr>
<tr>
<td>ILT2.5 Speech &amp; MT: MT/Recognition Integration</td>
<td>41</td>
</tr>
<tr>
<td>ILT3 Text Classification and Automatic Labelling (Dr. Carl Vogel)</td>
<td>41</td>
</tr>
<tr>
<td>ILT3.1 Text Classification and Automatic Labelling for Next Generation Localisation: Automatic Dependency Annotation</td>
<td>41</td>
</tr>
<tr>
<td>ILT3.2: Automatic Localisation Metadata Annotation</td>
<td>42</td>
</tr>
<tr>
<td>ILT3.3: Multilingual Text Type and Genre Classification</td>
<td>42</td>
</tr>
</tbody>
</table>
MT Research at CNGL

ILT Integrated Language Technologies ................................................................. 35
ILT1 Core MT Systems (Prof. Andy Way) ................................................................ 35
ILT1.1 Core MT Engines: Incorporating Syntax into SMT Systems ......................... 35
ILT1.2 Core MT Engines: Improved Hybrid Systems ............................................... 35
ILT1.3 Core MT Engines: Scaling More Linguistically Sophisticated Systems ........... 36
ILT1.4 Core MT Engines: Probabilistic Transfer ................................................... 36
ILT1.5 Core MT Engines: Tuning Machine Translation ........................................... 37
ILT1.6 Core MT Engines: General Alignment Models ............................................. 37
ILT1.7 Core MT Engines: Automatic MT Evaluation ............................................. 38
ILT1.8 Core MT Engines: Controlled Language ................................................... 38
ILT1.9 Core MT Engines: Patients with Limited English ........................................ 38
ILT2 Speech Technologies and MT (Prof. Julie Berndsen) ........................................ 39
ILT2.1 Speech & MT: Phonetic Parameters/Voice Characteristics ........................... 39
ILT2.2 Speech & MT: Muster Speech Engine: Synthesis ....................................... 39
ILT2.3 Speech & MT: Muster Speech Engine: ASR ............................................. 40
ILT2.4 Speech & MT: MT/Synthesis Integration .................................................. 40
ILT2.5 Speech & MT: MT/Recognition Integration ............................................... 41
ILT3 Text Classification and Automatic Labelling (Dr. Carl Vogel) ......................... 41
ILT3.1 Text Classification and Automatic Labelling for Next Generation Localisation: Automatic Dependency Annotation ......................................................... 41
ILT3.2: Automatic Localisation Metadata Annotation ........................................... 42
ILT3.3: Multilingual Text Type and Genre Classification ......................................... 42
Thank you! Questions?