Tools for Translation
State of the Practice,
Established Tools and
Capability Gaps

Jennifer DeCamp
MITRE Corporation
jdecamp@mitre.org
Agenda

• Fundamental Questions
• Types of Translation
• Translation Workflows and Current Tools
• Gaps
• Tolerance/Preferences
  – Human Translator
  – Post-Editor
  – MT User
• Conclusions
• Recommendations for MT Summit
Fundamental Questions

1. How can you get better accuracy in MT/HT?
   Communicative value (Tapling 2008)
   Fluency, fidelity, etc.

2. How can you make a C-level translator into a B-level translator?

3. How can you make human translators more productive?
Types of Translation

- Translation
- High Quality Translation
  - Literature, marketing material, important briefings
  - Nuance, style
- Gists/summaries
- Getting information to answer specific questions
- Sorting—figuring out the language, subject, and difficulty in order to route the material
- Multimedia
  - Television, etc.
- Embedded MT
- Multi-language source material
- Different dialects and registers
- MT embedded in chat and search tools
- Term translation
- Speech-to-speech translation
Some Major Distinctions

- Publisher centric vs. User centric
- Human Translation (HT) vs. Machine Translation (MT)
  – Increasingly mixed
- Cautious vs. whatever
Combinations

- Publisher-Centric
  - HT
  - MT
  - Combination

- User-Centric
  - HT
  - MT
  - Combination
  - With user tools?
Publisher-Centric HT

Data Owners

Translation

End Users
Publisher-Centric MT
User-Centric HT

End Users

Translation
User-Centric MT

End Users

Free Internet MT
- Bablefish
- Altavista
- Systransoft
- Systranet
- Google Translate
- Free Translation (SDL)
- Microsoft Windows Live Translator
- Licensed Internet MT
- Licensed In-House MT
Translation Memory

• Examples
  – Products: Across, Bee-Text, Multicorpora, SDL TRADOS, WordFast, LingoTek
  – Open Source: GlobalSight, Omega-T

• Based on MT research

• Align source and translated texts

• Presents past translations to translator

• Carried to extremes, it provides only text that has not been translated
GlobalSight Open Source TM
Terminology Management

• Examples:
  – MultiTerm, Terminotix, Etc.

• Dictionaries of approved terms
  – Content-oriented
  – Sometimes multiple languages

• Often developed by expert translators/terminologists

• Usually results in clear single right term
  – Potential for MT?
Research of Terms and Contexts

• Search
  – Google, Yahoo, etc.
  – TransSearch
  – Wikipedia
  – Multiple dictionaries
  – Chat rooms
  – Translator site bulletin boards
  – Blogs

• Communicate with
  – Language and language domain experts
  – Authors of text being translated
  – Developers
  – Subject Matter Experts
  – Users (focus groups)
  – Foreign marketing offices
  – Surveys

• Consult
  – Transliteration standards
  – Abbreviation and acronym conventions

• Disambiguate terms
• Document!
Machine Translation with Human Pre-Editing and/or PostEditing
Pre-Editing

• Examples
  – Authoring systems
    • Xerox, Smart’s MaxIT, Acrolynx, and AuthorIT
  – Translatability ratings
    • Bernth and Gdaniec (2001)

• Feedback to authors of text on how to make their writing more machine-translatable
  – Short sentences
  – Full sentences
  – Punctuation
  – Approved vocabulary

• Range from writing tips to controlled language
Post-Editing

• Examples
• Markup for translators
  – Areas needing work
  – Areas not needing work
• With no markup, translators must read source and target text. Source text is sometimes not read.
• Interesting editing and QA tools
  – TransType (RALI)
George Bush reconnait avoir autorisé des écoutes

Le président américain George W. Bush a admis samedi qu'il avait autorisé l'une des agences de renseignement les plus secrètes, l'Agence pour la sécurité nationale (NSA), à mettre sur écoutes des gens "connus notamment pour leurs liens avec Al-Qaïda".

"Dans les semaines qui ont suivi les attentats [du 11 septembre 2001] contre notre pays, j'ai autorisé la NSA, conformément au droit américain et à la Constitution, à intercepter les communications internationales de personnes connues pour leurs liens avec Al-Qaïda et des organisations terroristes qui y sont liées", a déclaré M. Bush dans son allocution radio-télévisée hebdomadaire.

George Bush acknowledges having authorized tapping

The American President George W. Bush admitted Saturday he had authorized an intelligence agencies more secret, the agency for national security (NSA), eavesdrop people "known for their links with Al-Qaïda."

"In the weeks following the attacks [of 11 September 2001] against our country, I authorized the NSA, pursuant to the American law and the Constitution, to intercept communications international people known for their links with Al-Qaïda and terrorist organisations which are related," said Mr. Bush in his speech radio-television weekly.
Tools for GIL

- Globalization, Internationalization, and Localization (GIL)
  - Globalization, Internationalization, and Localization Technology (GILT)
  - i18n, l10n, g11n
  - Locale

- Tools
  - Strip out text and provide it to translators
  - Help ensure consistency of usage
  - Provide information/means of dealing with differences between countries and regions
    - Language/dialect, voltage info, plug info, font preferences, color and design preferences (Common Locale Data Repository), spelling differences
    - Term differences (terminologies)
      - Useful for MT?
User-Centric MT with Tools?

End Users

Post-Editing
Rewriting of Input
Gaps

1. Use MT approaches to help HT
2. Use HT approaches to help MT
How do we better determine and provide term translations?

• What can we do besides provide dictionaries in electronic format?
  – WordNets
  – Translation Memory with context
  – TransSearch (RALI)
  – Morphological Analyzers (right approach?)

• How do we make the right word with the right conjugation appear in the user’s translation with the least number of keystrokes?

• How can we share dictionaries and standardize terminology across tools?
  – ISO Lexical Markup Framework, OLIF, etc.
How do we build better dictionaries and terminologies?

• How do we find new terminology?
  – OOV / NFW searches?
  – Other?

• How do we develop definitions?
  – Can we automate more of the process?
    • Examples, larger term, distinguishing term, etc.

• How do we standardize terms?
  – Frequency is not always the best solution

• How do we provide easy conversion between terms in a translation?
  – Transliterated terms
  – Different terms

• How do we alert users that there may be synonyms?

• How do we disambiguate terms?
  – Frequency is not always the best solution
  – Cycle through by frequency?
How can we use MT to help translators?

- MITRE Research
- Each source sentence is visually aligned with usually two distinct machine translations of that sentence
  - Sentence boundaries are identified automatically
  - MT output can be copy + pasted into translation/gist authoring pane
- Provides context and overview
- Provides triangulation
How else can we use MT technologies?

• **TransType**
  - EC’s Information Society Technologies (IST) Program
  - Lapalme, Langlais, Macklovitch, Gandrabur, Nguyen, Nie
  - Partners
    – Atos Origin
    – Lehrstuhl für Informatik VI, Computer Science Department, RWTH Aachen - University of Technology
    – Instituto Tecnológico de Informática, Universidad Politécnica de Valencia
    – RALI Laboratory - University of Montreal
    – Celer Soluciones
    – Société Gamma
    – Xerox Research Centre Europe

• **Text prediction**
  – Saves keystrokes
  – May increase speed and accuracy of translators
TransType

Thank you for purchasing the Xerox DP350CF/DP700CFDDP500CF/DP1000CFD continuous forms laser printer.

The printer is designed for ease of use, but to use your machine to its fullest potential, take some time to read this Operator Guide.

Audience

This DP350CF/DP700CFDDP500CF/DP1000CFD Operator Guide is intended to be used by the Operators and Key Operators of the printer in order to run and define production jobs.

It is assumed that all users understand the peripheral equipment that may be used in the print line. It is also assumed that suitable operation manuals will be available for this peripheral equipment.

Document organization

The User's Guide contains the following sections:

* Chapter 2 - provides a Product Overview which describes the main components of the DP350CF/DP700CFDDP500CF/DP1000CFD and provides a product specification.

* Chapter 3 - Duplex Printing describes a description of the duplex printing process and how it is achieved on the DP700CFDDP1000CFD.

Acerca de este manual

Gracias por adquirir la impresora láser Xerox DP350CF/DP700CFDDP500CF/DP1000CFD de formularios continuos.

La impresora está diseñada especialmente para facilitar aprovechar todo el potencial de la máquina, dedique algún tiempo a leer esta guía se indique que lo haga.
Do not touch the drum surface.

Touching the drum could reduce the print quality.

WARNING

This warning alerts you to areas of the product where there is the possibility of personal injury.

Si toca el tambor, puede hacer que se reduzca la calidad de impresión.
Puede hacer que se reduzca la calidad de impresión.
Puede hacer que se reduzca el calidad de impresión.
Use of MT and TM to Spot-Check Quality

• Compare source and translated texts to identify problems with
  – Omissions
  – Numerical expressions
  – Source language interference (Macklovitch)
    • E.g., False cognates
      – Library and librairie

• Check consistent use of terms
• Test translators
Use of Tools to Get Bilingual Concordances of Terms

• To
  – Check context
  – Research alternatives
  – Develop terminologies and/or new dictionaries

• Translation Memory

• Search
  – TransSearch (RALI)
  – In house or restricted high-quality search
  – Internet search

• Many opportunities for analysis and development
vital and creative communities.

The 2007 Ontario Budget included a $15 million annual funding increase to the Ontario Arts Council by 2009-10.

Support arts and culture in the 905 region with an additional $1.2 million provided to the McMichael Canadian Art Collection.

Increase annual funding to the Ontario Trillium Foundation by $20 million by 2010, to meet the needs of Ontario's diverse community-based charities and not-for-profit organisations.

Work with community partners to ensure the successful completion of 62 cultural capital projects approved for $146.9 million in funding from the Sports, Culture and Tourism Partnerships (SCTP).

Seventeen projects are still in progress.

Unit handled a 53% increase in the number of terms rendered into French.
TransSearch

Macklovitch, Lapalme, Gotti, 2009
TransSearch
20 Most Frequent 2-Word Queries

What could be done with this kind of data?

<table>
<thead>
<tr>
<th>Query</th>
<th>Freq.</th>
<th>Query</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>as such</td>
<td>1195</td>
<td>as of</td>
<td>533</td>
</tr>
<tr>
<td>over time</td>
<td>1046</td>
<td>most importantly</td>
<td>530</td>
</tr>
<tr>
<td>consistent with</td>
<td>743</td>
<td>more importantly</td>
<td>511</td>
</tr>
<tr>
<td>in turn</td>
<td>708</td>
<td>where appropriate</td>
<td>499</td>
</tr>
<tr>
<td>hard work</td>
<td>680</td>
<td>as per</td>
<td>493</td>
</tr>
<tr>
<td>along with</td>
<td>669</td>
<td>due diligence</td>
<td>486</td>
</tr>
<tr>
<td>subject to</td>
<td>655</td>
<td>build on</td>
<td>483</td>
</tr>
<tr>
<td>based on</td>
<td>649</td>
<td>in particular</td>
<td>478</td>
</tr>
<tr>
<td>as required</td>
<td>609</td>
<td>focus on</td>
<td>472</td>
</tr>
<tr>
<td>as appropriate</td>
<td>587</td>
<td>where possible</td>
<td>461</td>
</tr>
</tbody>
</table>
Use of Intelligent Templates

• Martin Kay, Xerox PARC
• Restricted options for translators
How can we get better HT Training?

• MT incorporated into subject-focused training
  – Egan 2009

• Provide automatically-generated lessons on highly relevant subject matter
How can we improve post-editing?

• Mark the problems for the post-editors
• Provide information on the specific problems
  – Kind of problem
  – Suggested change
  – Linked resources for making change/getting info
• Provide other info
  – Tips on getting better output
• Automate some post-editing (Doyen et al. 2008)
How can we improve pre-editing?

- Automate some parts
- Make “checkers” less intrusive
- Use source text analysis in additional ways
  - Translatability Ratings (Bernth & Gdaniec 2001)
  - Problem tagging (DeCamp 2009)
Use of MT, Entity Tagging, and Other Tools

- **Examples**
  - MITRE research
  - Day et al.
  - CACI Language Workbench

- Help routing
- Provide overview
- Call up resources?
How Can We Improve Retranslation?

• If a user requests a retranslation of a passage or document, how can we better understand what he/she needs?

• How can we quickly provide it?
  – What additional information can we send without going back to a translator?
How can we better understand translators?
How can we help translators use tools?

- Automated logging
- Recommender systems
- Examples
  - TransType
  - MITRE
- Testing difficult
How Can We Improve MT with HT Practices?

- Footnoting and annotation
- Multiple meanings with examples
Translator and User Acceptance

• Translators and Post-Editors
• Tolerance depends—and can change
  – Group (demographics, etc.)
  – Tasks and objectives
  – Training and exposure
  – Perception of helpfulness of tools
    • May be influenced by simple issues such as access
• Perception of their customers’ requirements
• Difficult to assess productivity gains due to so many factors in human translation
Conclusions

1. The term “translation” covers many tasks and requirements.

2. There are many means of meeting these requirements with HT, MT, and a combination, depending on needs for accuracy, fluency, etc.

3. There is much room for research and development
   - MT to HT
   - HT to MT
At the Technology Showcase

- **Friday, August 28**

- **Dictionaries**
  - Army Research Labs
  - WordScape

- **Translation Memory and/or Tools for Translators**
  - Basis Technology (particularly input methods)
  - Canadian Translation Bureau
  - MultiCorpora
  - PAHO
  - ProMT
  - SDL and LanguageWeaver

- **Document Triage and Handling**
  - CACI
  - Center for Applied MT