LANGUAGE TECH NEWS

ata A Publication of the Language Technology Division
of the American Translators Association.

From the Editor:

I would like to welcome all of you. This is the first issue on which I have worked independently as the editor. The earlier editor, Lisa Carter, was a great help with the last issue and a valuable sounding board when I discussed a few ideas with her. The newsletter has grown under her editorship and I hope I can be a worthy successor.

This issue is bulkier than the earlier ones. I hope future issues will be able to fulfill its promise. I have been looking after the listserv as well and I have noted that most of our members haven't joined the LTD listserv. Please do so and participate in the fruitful discussions between members. I would also like members to suggest topics on which they might like to see articles published and contribute one themselves! As editor, I would like to be proactive and involve as many members as possible.

I am particularly thankful to the contributors Tuomas Kostiainen, Ravishankar Shrivastava, and Kirti Vashee for their contributions. Without Laurie Gerber, we wouldn't have had the LTD minutes recorded at the ATA Annual Conference. I would also like to express my thanks to Thierry Fontenelle for letting me reprint his article. Without ideas from Dierk Seeburg, the LTD Administrator, and without Assistant Administrator Naomi Sutcliffe de Moraes's tolerance of my constant banter and queries, and without Cindy's formatting work, Language Tech News would not have seen the light of day.

Any errors and shortcomings that you see here are mine. Please send feedback to roomynaqvy@gmail.com and roomynaqvy@yahoo.com.

I hope the new year brings great tidings for language technology enthusiasts! □

—Roomy Naqvy

VOL. 2, NO. 4 / FEBRUARY 2009

IN THIS ISSUE:

From the Assistant Administrator:

Welcome to the winter edition of Language Tech News.

I hope many of you were able to go to the great ATA annual conference in Orlando, FL, and attend some of the language technology talks and the LTD wine and cheese reception. It is not too soon to submit a proposal to speak at the next annual conference, the 50th, in New York City, which promises to be very well attended. The submission deadline is March 9, 2009, which is fast approaching! Go to http://www.atanet.org/conferencesandseminars/proposal.php .

The LTD also suggests speakers for the conference, especially non-translators who have expertise in an area related to language technology. Please let us know if you have any suggestions for topics or individuals or both!

I would also like to remind you that we have a blog on the main LTD website at http://www.ata-divisions.org/LTD/. You can just stop by occasionally or sign up to receive updates through the RSS feed.

Naomi J. Sutcliffe de Moraes

Assistant Administrator, Language Technology Division

LTD NEEDS YOUR CONTRIBUTIONS!

Become a Presenter

It is not too soon to submit a proposal to speak at the next annual conference, the 50th, in New York City, which promises to be very well attended. The submission deadline is March 9, 2009, which is fast approaching! Go to www.atanet.org/conferencesandseminars/proposal.php to submit a proposal.

Call for Reviewers

Are you using software that helps you in your day-to-day work as a translator and/or interpreter? Tell us about it!

Send reviews of your new, favorite, even most-hated language technology software to roomynaqvy@gmail.com and roomynaqvy@yahoo.com by the submission deadlines (see page 5.) Your colleagues will thank you!

Language Tech News

Vol. 2, No. 4 February 2009

Copyright © 2009

American Translators Association
225 Reinekers Lane, Suite 590
Alexandria, VA 22314
Telephone (703) 683-6100
Fax (703) 683-6122
ata@atanet.org
www.atanet.org

Editor
Roomy Naqvy
roomynaqvy@gmail.com

Proofreader:
Barbara Guggemos
Naomi J. Sutcliffe de Moraes

Contributors to this Issue:
Thierry Fontenelle
Laurie Gerber
Tuomas Kostiainen
Naomi J. Sutcliffe de Moraes
Roomy Naqvy
Dierk Seeburg
Ravishankar Shrivastava
Kirti Vashee

Layout:
Cindy Gresham
illustrator@cgresham.com

LTD is the
Language Technology Division
of the American Translators
Association

LTD Administrator:
Dierk Seeburg
info@icotext.com

LTD Assistant Administrator:
Naomi J. Sutcliffe de Moraes
www.justrightcommunications.com

Can I Remove a Word from Office's Speller Dictionary?

By Thierry Fontenelle

The other day, I was discussing a number of suggestions to improve Office's spell-checker. A customer was suggesting we should allow users to delete individual items from Word's spell-checker lexicon. This feature is already available, in fact: if you want to specify a preferred spelling for a word and to exclude a given spelling from the main lexicon used by the Office speller, you need to use an "exclusion dictionary." Your speller comes with an empty exclusion dictionary and you can add words to it if you want them to be permanently red-squiggled.

I am thankful to Thierry Fontenelle for letting me reprint his article. His blog post has been reproduced from the Natural Language Group Blog at MSDN, which can be found at:

http://blogs.msdn.com/naturallanguage/archive/2008/09/11/can-i-remove-a-word-from-office-s-speller-dictionary.aspx

—*Editor*

You first need to locate your exclusion dictionary, which, if you use Vista and Office 2007, can be found in the following folder:

```
C:\Users\UserName\AppData\Roaming\
Microsoft\UProof\
```

Each language has a specific dictionary whose name starts with "ExcludeDictionary", followed by the language code (EN for English, FR for French, SP for Spanish, GE for German…), followed by the LCID (locale identification number). The extension is .lex. For instance:

English: `ExcludeDictionaryEN0409.lex`

French: `ExcludeDictionaryFR040c.lex`

You can open the file with Notepad or WordPad and add a word which you want the speller to flag as misspelled. Save and close the file. You are done!

You can type "exclude dictionary" or "exclusion dictionary" in the Office help to get more information about this feature.

Of course, caution should be exercised when you decide to remove a word from your Office speller. If you decide to remove the word manger because you frequently type program manger instead of program manager, you should not be surprised when your speller flags manger in a sentence like "Jesus was born in a manger". This is why we have introduced a contextual speller, which tries to identify words which exist but are misspelled in a given context.

To give another example where contextual spelling might be preferred over exclusion, consider the user who had contacted the Word newsgroup to find out how to exclude the word "ahs" from the main speller lexicon. This user kept typing ahs instead of has. The new context-sensitive speller in Office 2007 flags a number of contexts where "ahs" should not be used, however, which should address this user's problem without having to remove the word altogether from the lexicon. You will see a blue squiggly line under "ahs" if you write something like "He ahs never done it before", for instance. But you will not get any flag under "ahs" if you write "we definitely got oohs and ahs all around when we launched this product". ▢

Thierry Fontenelle is a Senior Program Manager with Microsoft's Natural Language Group where he works among others on the creation of dictionaries for proofing tools and natural language processing. Before joining Microsoft, he worked as a translator for NATO and the European Commission Translation Service. He can be contacted at: thierryf@microsoft.com

ATA Language Technology Division Annual Meeting

Minutes taken by Laurie Gerber

Dierk Seeburg opened the meeting at 4:05 p.m.

Dierk reviewed the accomplishments of 2008:

- There are 15 language technology sessions at the conference
- The division put out 4 newsletter issues under the editorship of Lisa Carter (the maximum number, on par with established divisions)
- The division maintains a listserv for division news and other topical discussion
- An LTD networking event is being held at ATA for the first time

Discussion:

Some meeting participants were not aware of the newsletter and the listserv. Naomi collected names and emails to add attendees to the list.

Attendees were reminded that the newsletter is posted on the LTD division webpage, accessible from the ATA website. The division website and newsletter is publicly accessible, and Naomi encouraged attendees to spread the word about this information resource to their colleagues inside and outside of ATA. New issues of the LTD newsletter are announced in the general ATA email update, as well as on the LTD listserv.

Goals of the listserv include discussion and information-sharing on tools, conferences and trends of interest.

LTD's role in developing the LT session program consists largely of inviting and proposing presenters and encouraging division members to submit presentation proposals. Selection of sessions from the submissions and proposals is done by the TAC (Translation and Computers) committee chaired by Alan Melby.

ATA Language Technology Division Meeting Attendees:

Dierk Seeburg
Naomi Sutcliffe de Moraes
Cathi Rimalower
Brooks Haderlie
Evan Cohen
Rosana Wolochwianski
Michael Wahlster
Jesus Garrido Muro
Laurie Gerber
Alan Melby

Updates and new business:

Newsletter

Newsletter editor Lisa Carter has handed over the reigns to our new editor, Roomy Naqvy. The layout person Cindy Gresham will continue in that role.

Division Election

The term for the current division officers will expire at next year's annual ATA conference. The division needs to prepare for officer elections by forming a nominating committee by March 2009. Attendees were encouraged to volunteer for the nominating committee, and Evan Cohen did so.

Alan Melby reported on this year's Tools forum session. The session was divided into two parts—the first session was aimed at individual translators, the second at project managers. Alan held a lunch with the tool vendors today. Many of the attendees at the meeting requested copies of the handouts. These may be made available on the division website.

Alan also reported on discussions with tool vendors about training sessions at the ATA conference. Last year, Trados held a one or two-day hands-on training session. This year Wordfast did the same. Naomi suggested that the ATA provide the time and space for all tool vendors to conduct hands-on training sessions with their tools, in which translators install the tools on their own computers.

Various logistical complications were discussed:

- Who would pay for the meeting room rental
- How much members would be charged to participate
- Whether it is practical to have participants install and set up tools during a short (half day or less) session

- How to get the vendors to cooperate/agree to a schedule
- That it may require participants to come a day earlier or stay later at the conference with the associated expense and time away from home

Alan will look into the availability and the cost of arranging the space.

The LTD was asked to poll members to assess the level of interest in such sessions.

Alan reported on the technology survey he conducted with Jost Zetsche. The survey was sent to ATA members and 7 FIT member associations internationally. There were 800 responses from ATA members and 400 from the FIT member associations. This survey was presented as "Phase 1." Jost is compiling the results. The last question of the survey was whether people would agree to participate in a more detailed "Phase 2" survey. Two-thirds of the respondents agreed. Alan would like to solicit feedback on a draft of the Phase 2 survey early next year before it is circulated.

ATA 2009

Dierk encouraged attendees to propose language technology sessions for next year's conference—the 50th ATA conference. Approximately two-thirds of those present do plan to attend next year's conference.

Division Website

ATA makes a $500 honorarium available to divisions for division website maintenance. Dierk suggested that the division focus on website maintenance, such as posting reports or other information. Discussion on the listserv may be a good way to develop ideas that can turn into such postings. The group suggested topics for such postings: tools, security standards, and industry trends.

Webmaster Michael Wahlster mentioned that articles are tagged by topic on the website so that they are easy to find. Moreover, they are accessible and indexed by search engines, so ATA members and others can find them.

Dierk suggested that the division needs to publicize the website and its resources in a better way.

Michael mentioned that there were an average of 375 unique hits per month on the LTD website, which is considerably higher than the German language division website with 150 hits per month. However, the German Language Division listserv has an average of 660 messages per month, so this is where information exchange happens.

AMTA collocation in 2010?

Laurie Gerber reported that the AMTA (Association for Machine Translation in the Americas) is considering holding its conference in/near the ATA conference in 2010. Feedback was sought on whether this was a good idea.

Adjournment

Dierk asked for a motion to adjourn. Motion was made, seconded, and voted on unanimously. Meeting adjourned at 6:25pm. □

Language Tech News Submission Guidelines

Submissions should be no longer than 2000 words. (If you have a longer piece, contact Roomy Naqvy at roomynaqvy@gmail.com and roomynaqvy@yahoo.com to discuss options.) Send all articles and items of interest to Roomy Naqvy at either of the above addresses. Articles should be in Word (.doc or .rtf) form, and should include a short bio at the end, ideally containing your e-mail or website contact. Do copy all communications to both the emails and do mention 'ATA LTD Newsletter January 2009' in the subject of the email.

We also gratefully accept reprints of articles you may have published elsewhere, so be sure to offer those as well.

Issues	Submission Deadline	Planned Publication Date
Vol. 3/No. 1	Friday, February 20	March 2009
Vol. 3/No. 2	Wednesday, May 20	June 2009
Vol. 3/No. 3	Thursday, August 20	September 2009
Vol. 3/No. 4	Friday, November 20	December 2009

Trados Tip

by Tuomas Kostiainen

Using MultiTerm With Trados

Part 1: Did You Forget MultiTerm?

MultiTerm is very easy to forget. First, you need to remember to download it, then to install it. Then at the end you also need to remember to set it up for use with Trados Workbench. Somewhere in between you would also need to figure out how and where to get those MultiTerm glossaries. However, once you have done all that and have started using it, I promise you won't forget it anymore. It will become an essential part of your Trados translation process. MultiTerm is one of the best features in Trados, but unfortunately it is also one of the most underutilized ones.

In this first MultiTerm article, I will give a little background information and explain how to use MultiTerm glossaries as part of the interactive translation process with Trados. In subsequent articles, we'll cover additional topics, such as converting glossaries from other file formats into MultiTerm format and adding new terms during translation, or any other MultiTerm-related topic you would like to see.

MultiTerm is a terminology management program that is part of the SDL Trados package and is included in the Trados purchase price. MultiTerm can be used as a stand-alone program for creating and managing termbases or together with Trados through the Term Recognition feature of Trados Translator's Workbench.

It is important to note the difference between Translator's Workbench and MultiTerm. Workbench searches for translations of the entire source segment from a translation memory, whereas MultiTerm searches for translations for individual words/phrases from one or more MultiTerm termbases. Note also that the termbases do not grow automatically during the translation process as translation memories do. However, you can add new terms manually or "semi-automatically."

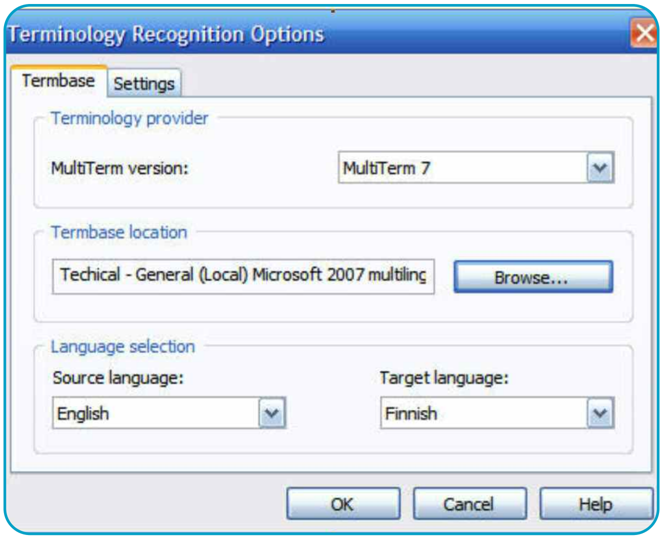
If you are one of those people who got frustrated with MultiTerm version 5 and have not tried it since, you should take a look at the newer versions. The new versions are much more functional, much easier to use and they have made it very simple to add new terms to a MultiTerm glossary directly from Word or TagEditor while translating.

Using Trados Term Recognition Feature

1. *Connecting MultiTerm Termbases to Translator's Workbench*

First you need to let Trados know which termbases you want to use:

- Open **Translator's Workbench** first if it is not open.
- Select **Options** > **Term Recognition Options**.
- Select the correct MultiTerm version from the drop-down list and click **Browse** to open the **Open Termbases** dialog box (see Figure 1).



MultiTerm version: MultiTerm 7

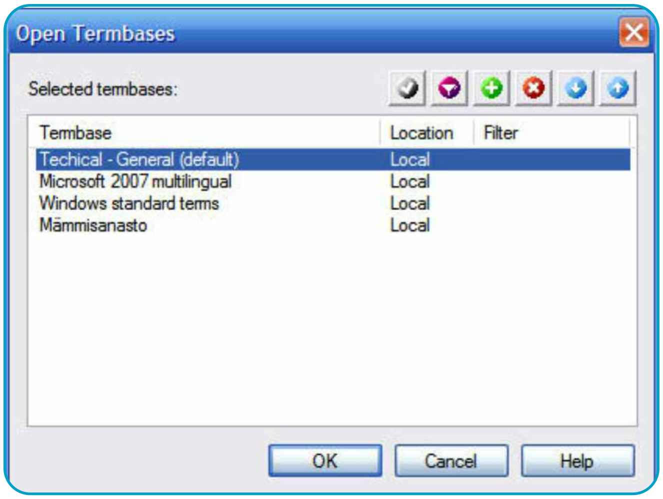
Techical - General (Local) Microsoft 2007 multilinc

Source language: English

Target language: Finnish

Figure 1. Terminology Recognition Options dialog box.

- To add a termbase, click the green **Add Termbase** button and select the termbase(s) from the list of available termbases (see Figure 2).
- Click **OK** to return to the **Open Termbases** dialog box when you have selected all the termbases you want to use.
- To delete a termbase from the list of Selected termbases, select the termbase and click the red **Remove Termbase** button. You can also change the order of the termbases with the blue **Move Down** and **Move Up** buttons if you have more than one termbase



Selected termbases:

Termbase	Location	Filter
Techical - General (default)	Local	
Microsoft 2007 multilingual	Local	
Windows standard terms	Local	
Mämmisanasto	Local	

Figure 2. Open Termbases dialog box listing currently selected termbases. Termbases can be added by selecting the green plus button and deleted by selecting the red X button. In this example, four termbases are selected.

selected and you would like to have the termbases searched and the results displayed in a specific order. Note that one of the selected termbases is listed as "default." A default termbase is always searched first, and it is also the termbase into which new terms are entered in TagEditor. (If you are using Word you need to specify the target termbase separately in Word.) You can change the default termbase with the grey **Set as Default** button.

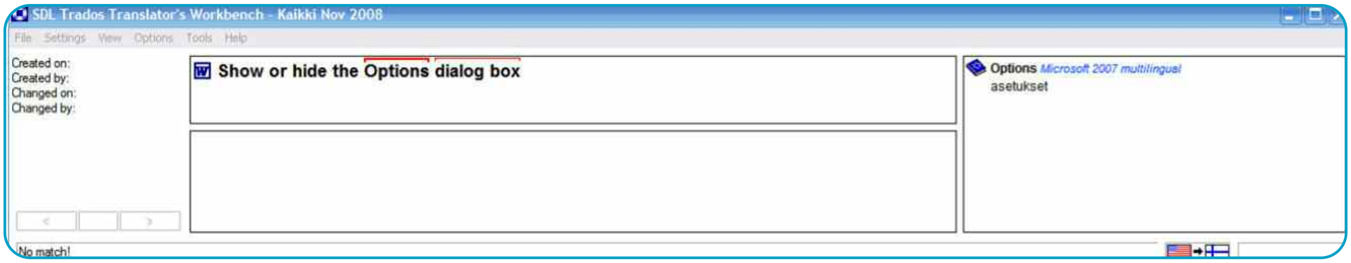
- When you have finished selecting termbases, click **OK** to close the **Open Termbases** dialog box. This brings you back to the **Terminology Recognition Options** dialog box. Click **OK** to close it.
- The terminology window should now be open in Workbench on the right. If not, check that **Options** > **Term Recognition** is selected. This is the window where the automatic termbase search results will appear during translation.

2. Using Term Recognition During Translation

So, now you have selected the termbases to be used for automatic term recognition. The recognition will take place automatically but you still need to select which terms you want to use and to get them inserted into your target field in your translation. This is done as follows:

- Verify that the term recognition feature and the correct termbases have been selected in Translator's Workbench (see above).
- Start translating normally in Word or TagEditor. When you open a sentence for translation, Trados automatically searches the selected termbases for any of the terms in the source segment. The terms that are found are indicated by red overhead brackets in the source window and the first term with its translation is shown in the Terminology window on the right-hand side of the Workbench window. Note also that the first term has a thicker bracket than the other terms. This indicates that the term is the "current" term—the term whose translation is in the Terminology window.
- You can easily insert a translation of any of the found terms into the target field in Word or TagEditor by clicking one of the **Get Term** buttons (**Get Previous**, **Get Current**, and **Get Next**) in the Trados toolbar, or by the corresponding keyboard shortcuts (Alt+arrow left/down/right) without having to retype it.
- Unfortunately, only one source term at a time is displayed in the Terminology window. However, you can quite easily change the current term and bring any of the other known (red-bracketed) terms to the Terminology window by clicking the bracket of the term, or by using the Alt+arrow left/right keyboard commands while in Workbench (Workbench has to be the active program in order for the keyboard commands to work).

Figure 3. Trados Term Recognition feature in action. Terms "Options" and "dialog box" have been found in the active termbase. Note the thicker bracket above "Options," which indicates that it's the current term. The English term, its Finnish translation and the name of the active termbase are shown in the Terminology window on the right. Pressing Alt+arrow down (Get Current Term) will insert "asetukset" into the target translation field. Pressing Alt+arrow right (Get Next Term) would insert the translation for "dialog box."



Created on:
Created by:
Changed on:
Changed by:

Show or hide the Options dialog box

Options Microsoft 2007 multilingual
asetukset

No match!

- The Terminology window displays the source term and its translation from each of the termbases where the term was found and the name of the termbases. It does not show any of the other information you might have included in the termbases, such as context or source information. However, double-clicking the dictionary symbol in the Terminology window opens a separate MultiTerm window that includes all the additional information about the term.
- You can also do additional MultiTerm searches by selecting a word or a phrase in the source or translation memory window in Workbench, right-clicking it and then selecting **Search in MultiTerm** or **Fuzzy Search in MultiTerm** from the context menu. This will open a separate **MultiTerm** window with the search results if a matching term has been found. You can also do this type of additional search in Word or TagEditor by using the MultiTerm toolbar.

That's it. As you can see, using MultiTerm with Trados is quite simple after you have figured out the setup steps outlined above. □

Tuomas Kostiainen (tuomas@jps.net) is an English to Finnish translator and Trados trainer, and has given several Trados workshops and presentations. For more Trados help information, see www.finntranslations.com/tradoshelp.

Calendar of Upcoming Events

For more information, go to http://atanet.org:80/calendar/

DATE	TITLE
February 18–21	**National Association for Bilingual Education (NABE)** 38th Annual Conference Austin, TX For more information, visit http://atanet.org/calendar/
March 14	**American Translators Association (ATA)** Professional Development Seminar: Translation Tools San Francisco, CA For more information, visit http://www.atanet.org/pd/tools/

Share Your Knowledge

Language Tech News is a great way to get your name out to your colleagues, to share your expertise, and to give back to your division and your association. Send in an article and share the wealth of knowledge you have!

Sil Converter : A Freeware, Universal World Font Converter

By Ravishankar Shrivastava

Prior to Unicode, there were a number of arbitrary, stop-gap, language-specific arrangements through which documents with non-Latin characters were created in various electronic forms. In some cases—for example, such as in Hindi—there were more than a hundred different sets of proprietary legacy fonts that had been introduced by an equal number of vendors. Further, these fonts were not compatible with each other. As these documents were created in non-standardized fonts, they remained unsearchable and hence defied the very purpose of their electronic format. Now, Unicode has solved all these problems and has become the de-facto standard font for the Internet as well as for modern multilingual documents. However, there is a large amount of old data and ample content in legacy fonts that needs to be converted to Unicode (and, at times, vice-versa). Further, some old systems simply cannot be upgraded and made Unicode-enabled, and hence they still produce data in legacy fonts—often with complex formatting. The only option here is to use a good, efficient font converter.

Languages which the Sil Converter installs by default:

Devnagari (Hindi)
Bengali
Gujarati
Gurmukhi
Kannada
Malayalam
Oriya
Tamil
Telugu
Arabic
Cyrillic
Greek
Han
Hangul
Hebrew
Hiragna
katakana
Jamo
Pinyin
Tibetan

Many font converters are available now and some of them can be effectively used for font conversions. The result thus obtained may be from 80-100% accurate. Most of the font converters were designed for conversion between specific sets of fonts. Similarly, most font converters were efficient in converting fonts, but they were simply unable to preserve formatting. Most of them worked in an odd, plain text environment. So, if your document was rich in formatting and you wanted to preserve it after font conversion, your choices were severely limited and you might have needed to reformat the entire document.

Sil Converter was designed as a universal font converter with the ability to preserve document formatting. It can convert between any given set of fonts with a simple mapping file. According to Sil Converter's Help file, an overview of Sil Converter's capabilities is given below:

This package provides tools through which you can change the encoding, font, and/or script of text in Microsoft Word documents, XML documents, and SFM text and lexicon documents. It also installs a system-wide repository to manage your encoding converters and transliterators (TECkit, CC, ICU, Perl, or Python-based, as well as support for adding custom transduction engines).

For developers, it provides a simple COM interface to select and use a converter from the repository. It is easy to use from VBA, C++, C#, Perl, Python or any .NET/COM enabled language. This package is fully integrated with SIL FieldWorks, Adapt It, and the forthcoming SpeechAnalyzer software, providing the same system-wide registry of installed and available encoding converters for all of these user programs.

Additionally the package includes some extra utilities such as a clipboard converter for manipulating text between cut-and-paste operations.

Sil Converter works independently as a standalone application, but if you have Microsoft Word installed on your computer, it also installs a Word Macro called **Data Conversion** that you can access from the **Tools** menu. Conversion using the Microsoft Word macro is easy and efficient. You simply need to select all or a part of the document that needs conversion and click on **Tools** > **Data**

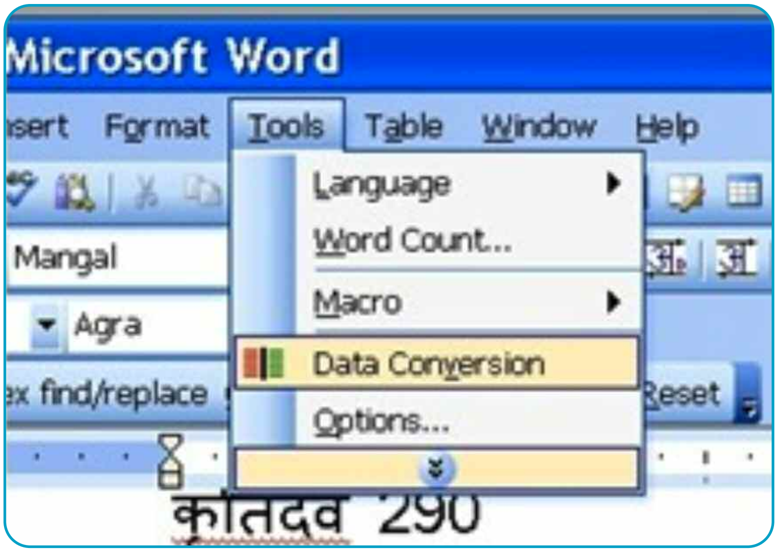
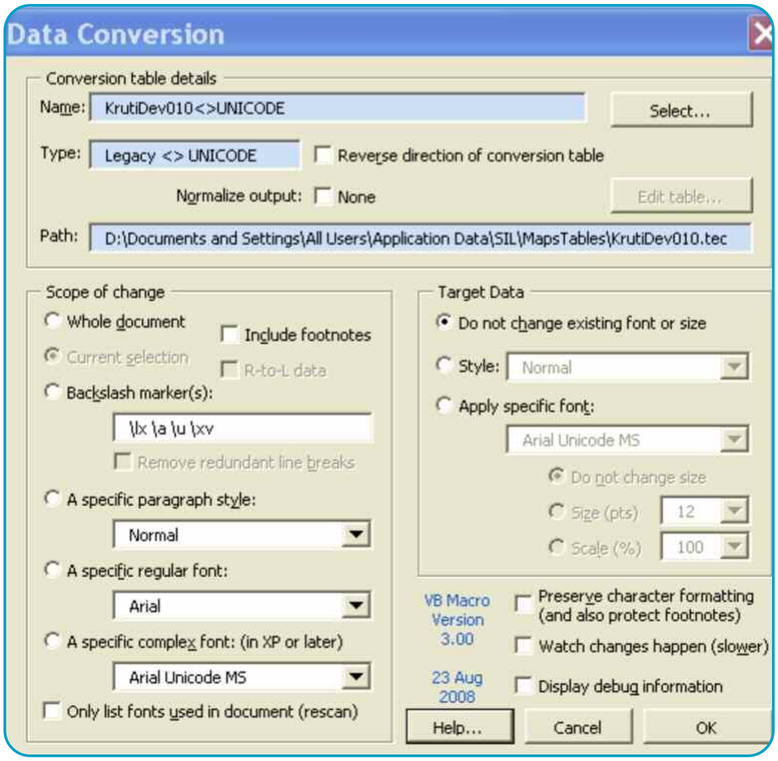


Figure 1: Sil Converter Data Conversion Macro MS Word Menu



Data Conversion

Conversion table details

Name: KrutiDev010<>UNICODE

Type: Legacy <> UNICODE ☐ Reverse direction of conversion table

Normalize output: ☐ None

Path: D:\Documents and Settings\All Users\Application Data\SIL\MapsTables\KrutiDev010.tec

Scope of change

○ Whole document ☐ Include footnotes

● Current selection ☐ R-to-L data

○ Backslash marker(s): \lx \a \u \xv

☐ Remove redundant line breaks

○ A specific paragraph style: Normal

○ A specific regular font: Arial

○ A specific complex font: (in XP or later) Arial Unicode MS

☐ Only list fonts used in document (rescan)

Target Data

● Do not change existing font or size

○ Style: Normal

○ Apply specific font: Arial Unicode MS

● Do not change size

○ Size (pts) 12

○ Scale (%) 100

VB Macro Version 3.00

23 Aug 2008

☐ Preserve character formatting (and also protect footnotes)

☐ Watch changes happen (slower)

☐ Display debug information

Figure 2: Interface of Sil Converter MS Word Macro

Conversion > **OK** (provided you have already set parameters such as conversion table details, etc.). Sil Converter can convert text from within the Clipboard—which means that you can copy/cut text in one Font and paste it into the other converted font.

Installation and Use:

Download the Sil Converter installer self-extractor file from this link: http://downloads.sil.org/EncodingConverters/3.0/WithExtras.exe . Double-click WithExtras.exe to extract file to a suitable folder (Say, `C:\converter`) on your hard disk. Within that folder, all the necessary files will be extracted under the **Sil Converters** sub-folder. In the **Sil Converters** sub folder, you will find **SetupEC.msi** file. Run this file to install Sil Converter (the Setup.exe file in parent folder may give you installation errors, hence it is safer to install with the *.msi file).

Now, open a word document (it works equally well in other file formats such as SFM & XML documents, too) in Microsoft Word and click on the Tools menu. You will find **Data Conversion** (in Microsoft Word 2007—it is available under the **Add-Ins Menu**) in the dropdown menu that appears. Click on it to open the configuration window. Click on the **Select** button under **Conversion Table** details, then click on the **Converter Installer** in the window that appears. Here, you will see dozens of converters. Select the desired ones and click **OK** twice to close this configuration window.

Now, to convert a part of the text, simply select it, then click on **Tools**>**Data Conversion** then click the **Select** button. From the available options, select the font pair you want to convert and click **OK** twice. Your data will be converted within no time. You will notice that not only was your document's formatting well preserved, but the conversion was also error-free. Please note that long chunks and big files take more time to convert; thus, it is advisable to break large files in smaller sections for speedier conversion. Similarly, conversion of plain text without formatting also takes noticeably less time than conversion of formatted text. □

Ravishankar Shrivastava is a pioneer in translating Linux Operating System, KDE 3.2-4.0, Gnome 2.0, XFCE 4.0, Debian Installer, OpenOffice Help, PC BSD Installer etc., into Hindi. His witty Hindi Blog : Raviratlami ka Hindi Blog was named the Best Hindi Blog of 2006 by Microsoft BhashaIndia. He has won a number of awards for his translations. In Feb. 2008, his Hindi Team won the prestigious FOSS.IN award—sponsored by NRCFOSS. He can be contacted at: raviratlami@gmail.com

The Continuing Evolution of Automated Translation Technology: RbMT vs. SMT

By Kirti Vashee

After more than fifty years of empty promises and repeated failures, amazingly, interest in machine translation continues to grow. It is still something that almost everybody hopes will work someday. Is MT finally ready to deliver on its promise? What are the issues with this technology and what will it take to make it work? And why do we continue to try after 50 years of minimal success? This overview attempts to provide a lay perspective on the ongoing discussion in the evolution of the two main approaches to "machine translation" that are in use today, and attempts to answer these questions. While other technical approaches to MT do exist, this overview will only focus on Rule-based MT (RbMT) and Statistical Machine Translation (SMT), as these approaches underlie virtually all the production MT systems in use today.

Kirti Vashee is Vice President of Enterprise Translation Sales, for Asia Online. He is a seasoned IT sales and marketing executive and SMT (statistical machine translation) evangelist who was previously responsible for the worldwide business development strategy of SMT pioneer, Language Weaver. At Asia Online, he is working with a team that plans to bring hundreds of millions of pages of new translated content, in various Asian languages, to the web over the coming year. He has established successful sales operations for several companies in Europe and the Asia-Pacific region, and has extensive experience developing motivated and effective distribution channels and partner networks. He can be contacted at kirti.vashee@asiaonline.net or k.vashee@gmail.com.

Why it matters

We live in a world where knowledge is power and information access has become a human right. In 2006, the amount of digital information created, captured, and replicated was 1,288 x 1018 bits. In computer parlance, that's 161 exabytes or 161 billion gigabytes …

This is about 3 million times the information in all the books ever written!

Between 2006 and 2010, the information added annually to the digital universe will increase more than six fold from 161 exabytes to 988 exabytes.It is likely that the bulk of this new information will originate in just a few key languages. So, are we heading into a global digital divide in the not-so-distant future? Two references testify to these facts:

- http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/
- http://www.universityofcalifornia.edu/news/article/17949

Peter Brantley (2008) at Berkeley says it quite eloquently:

For the Internet to fulfill its most ambitious promises, we need to recognize translation as one of the core challenges to an open, shared and collectively governed internet. Many of us share a vision of the Internet as a place where the good ideas of any person, in any country, can influence thought and opinion around the world. This vision can only be realized if we accept the challenge of a polyglot internet and build tools and systems to bridge and translate between the hundreds of languages represented online.....Mass machine translation is not a translation of a work, per se, but is rather a liberation of the constraints of language in the discovery of knowledge.

While stories of MT mishaps and mistranslations abound (we all know how easy it is to make MT look bad), it is becoming increasingly apparent to many that it is important to learn how to use and extend the capabilities of this technology successfully. While MT is unlikely to replace human beings in any application where quality is really important, there are a growing number of cases that show that MT is suitable for:

- Highly repetitive content where productivity gains with MT can dramatically exceed what is possible with just using TM alone
- Content that would just not get translated otherwise
- Content that cannot be affordably translated by human translators

- High-value content that is changing every hour and every day
- Knowledge content that facilitates and enhances the global spread of critical knowledge
- Content that is created to enhance and accelerate communication with global customers who prefer a self-service model
- Content that does not need to be perfect, but just approximately understandable

How They Work

As with all engineering problems, automated translation starts with a basic goal: to take a text in a given language (the source language) and convert it into a second text in another language (the target language), in such a way as to preserve the meaning and information contained in the source. While several approaches have been tried to date, two approaches stand out: Rule-based MT and Statistical Machine Translation (SMT). Some will argue that Example-based MT (EBMT) is also important, and many say this is the approach that underlies translation memory technology.

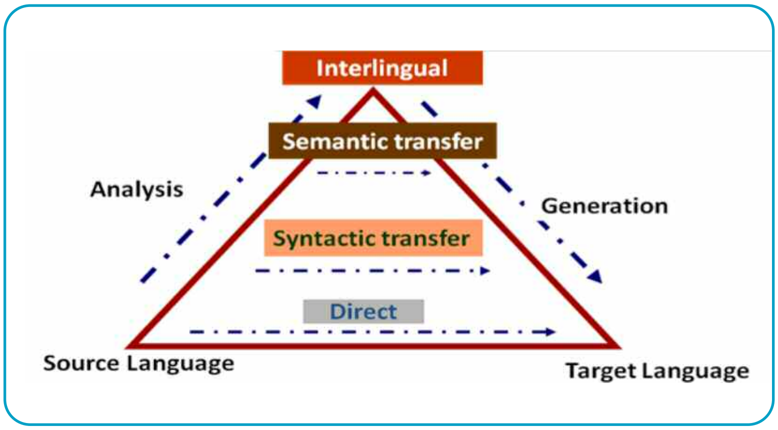


Figure 1

Rule-Based Machine Translation: RbMT

The foundation for rule-based systems is relatively easy to understand intuitively. Languages can be considered to have two foundational elements:

1. The meaning of the words—the semantics, and,
2. The structure of how the words are put together—the grammar, syntax and morphology etc.

So basically, an RbMT system attempts to map these two elements of the source language to the target language. While this may sound simple on the surface, it quickly gets complicated. Developers of RbMT solutions combine the theories of traditional grammarians and linguists and attempt to convert this linguistic knowledge into systematic, encyclopedic sets of rules encompassing grammar, morphology, syntax and meaning across a language pair. Programmers encode this information into rule sets and dictionaries and try and get as much linguistic knowledge as possible into these rule sets. Linguistic knowledge refers to information about word structure (singulars and plurals, first, second and third person endings and so on), word meanings (dictionary definitions), grammar (word order, part-of-speech [POS], typical phrasing), and homonyms (e.g. ambiguous terms can have different meanings in different contexts). Very simply put, an RbMT system consists of a dictionary and a set of rules for the language combinations that the system can process.

The Vauquois triangle shown in Figure 1 is useful to describe the various approaches to the MT engineering problem, and to better understand how they might evolve. What we see is that the simplest approaches used by early RbMT systems were based on direct transfer approaches, with simple dictionaries and very simple rules to change word order in the target language e.g., [Subject-Verb-Object (SVO) to Subject-Object-Verb (SOV)]. These early systems were improved by switching to a more modular approach that separated the analysis of the source language sentence from the synthesis of the target language sentence by means of a transfer stage. Thus, they have evolved into systems where linguistic analysis is done on the source language, and then through a transfer process (more complex rules + dictionary) the sentence is reconstructed in the target language. In this analysis phase, the system will look at inflections and conjugations of words (morphology), word sequence and structure of sentences (syntax), and to a certain degree the meaning of words in context (semantics). In other words, it will "parse" the source language sentence, so that the so-called "part of speech" or word class is determined for each word or phrase in the sentence, by checking it in a comprehensive dictionary of the language. This is repeated sometimes for the target language. Through this approach, there is some possibility to create a generalized model rather than a specific model for a single language pair, as in the direct transfer mode. To the best of my knowledge, none of the systems have really evolved to a point where semantics are deeply incorporated into the automated translation process. This is an area of possible evolution.

> *Often, several options for character combinations could exist and computers have great difficulty knowing how to approach this without the context that humans can easily place on a sentence.*

There are several core problems for RbMT systems:

1. **Human language inconsistency**: Human languages are filled with exceptions that do not follow the rules.
2. **Disambiguation**: Ambiguity remains the core challenge of RbMT systems. Words may have different meanings depending on their grammatical and semantic references.
3. **Local sentence focus**: RbMT systems analyze and translate one sentence at a time. They have little understanding about the context or the broader corpus from which the sentence originates.
4. **Inherent system conflicts**: Dictionaries are the key mechanism used to tune and refine and improve RbMT systems. However, the lack of semantic features on words and expressions means that dictionary maintenance must be monitored very carefully. A new dictionary entry that improves the translation in one sentence may introduce an error in another context.
5. **Skills required**: The development of these dictionaries is expensive as it requires a very unique skill set. These lexical skills require persons conversant in linguistics, corporate terminology, and computer software technology and software programming. The lexical information required goes beyond bilingual word lists and requires knowledge of part-of-speech (POS) information and morphology. The developers must have linguistic skills as well as language fluency.
6. **Maintenance overhead**: The complex rule sets that drive these systems become cumbersome and difficult to maintain over a period of time. Often different teams develop different parts of the rule sets and it can be nearly impossible for developers to have a full understanding of the rules and their interactions. This problem is compounded when very specific rules are introduced to handle special cases of grammar or translation.
7. **Diminishing returns**: As human language is inherently inconsistent—or infinitely nuanced—the law of diminishing returns comes into effect. Eventually, modifying a rule to improve results in one context weakens or destroys results in other contexts. Most RbMT systems have hit a ceiling on achievable quality after a few years, so that further modifications introduce as much degradation as improvement. Once RbMT systems reach this plateau, improvements are slow even if they are possible.
8. **New language pairs**: Given the process and the difficulties described above, we see that the development of new language pairs is arduous and slow if any reasonable quality is desired. The effort requires development of grammars, lexicons, morphologies, transfer rules, and generation rules. The people involved must have highly specialized skills and deep knowledge of the languages involved.

Many Asian languages have the additional issue of segmentation to consider. How is a computer to decide what a word is in a continuous block of Chinese or Thai characters? Often, several options for character combinations could exist and computers have great difficulty knowing how to approach this without the context that humans can easily place on a sentence. This is further complicated by the sparse use of punctuation and other elements that are common in western languages.

Customization of RbMT

The “free” MT systems that we see on the web today can all be characterized as baseline systems. Successful enterprise use of this technology is, however, characterized by special tuning efforts to raise the quality above these free systems. All MT systems can be optimized to specific use cases and will tend to perform better when this specific optimization is done.

For RbMT systems there are two ways to tune the system for specific customer requirements:

1. Modify the rule sets
2. Expand and extend the existing dictionaries and vocabulary to match the needs of the language used in the customer’s target application

The primary means of tuning most RbMT systems in use today involves the development of dictionaries. The dictionary development in most RbMT systems today requires dictionaries in which terms are coded for inflections (morphology) or for determining their position in a sentence (part of speech).

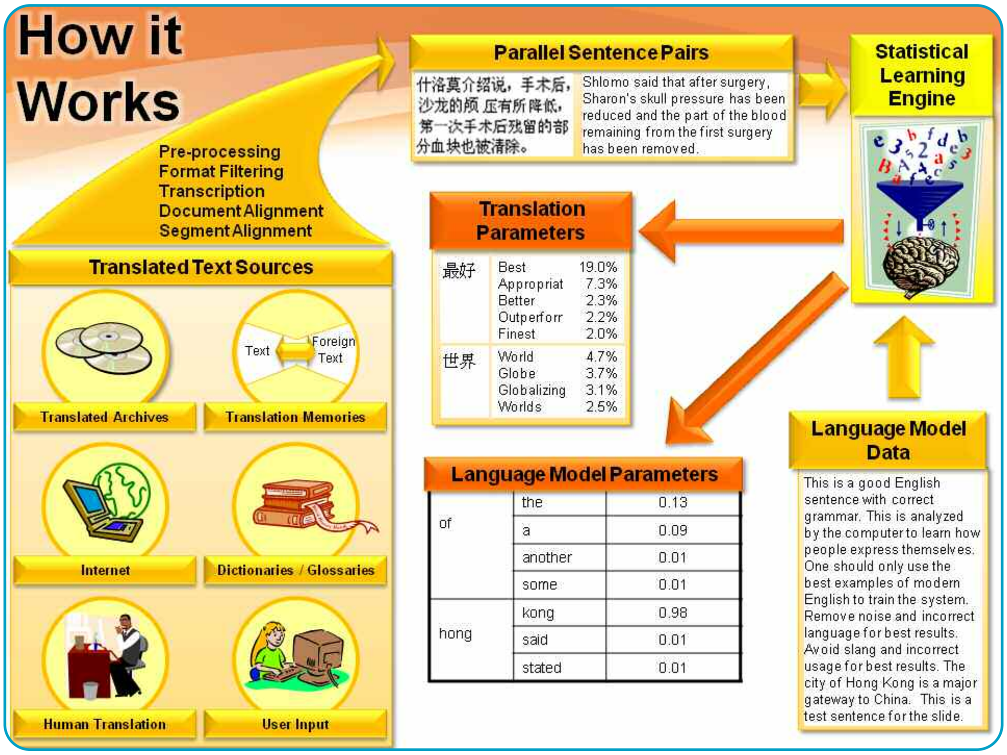
While it is theoretically possible to change the rule sets in RbMT systems, this is usually a very complex task that can only be done by very specialized development staff. Vendors of RbMT systems may be willing to do this for significant development dollars sometimes, but basically, this option is not available to general or even specialist users of the technology.

Statistical Machine Translation: SMT

Statistical machine translation approaches have been gaining considerable momentum since 2004. The premise of these new approaches is that purely linguistic knowledge is far less important than having large volumes of human- translated data to analyze and process. By analyzing large corpora of texts instead of just one sentence, the new systems attempt to simulate the way human translators work. Human translators have general knowledge of the everyday world, and they quickly grasp the context and the domain in which they operate. As computer storage and processing capacity increased exponentially, and as large digital bilingual corpora became available, researchers began to suggest that computers could “learn” and extract the systematic patterns and knowledge that humans have embedded into historical translations. SMT systems have in a few short years overtaken the RbMT systems in quality in most baseline systems available on the web today. (see Fig. 2)

These second-generation SMT solutions adopt a data and probability-based approach to translation and are also often

Figure 2



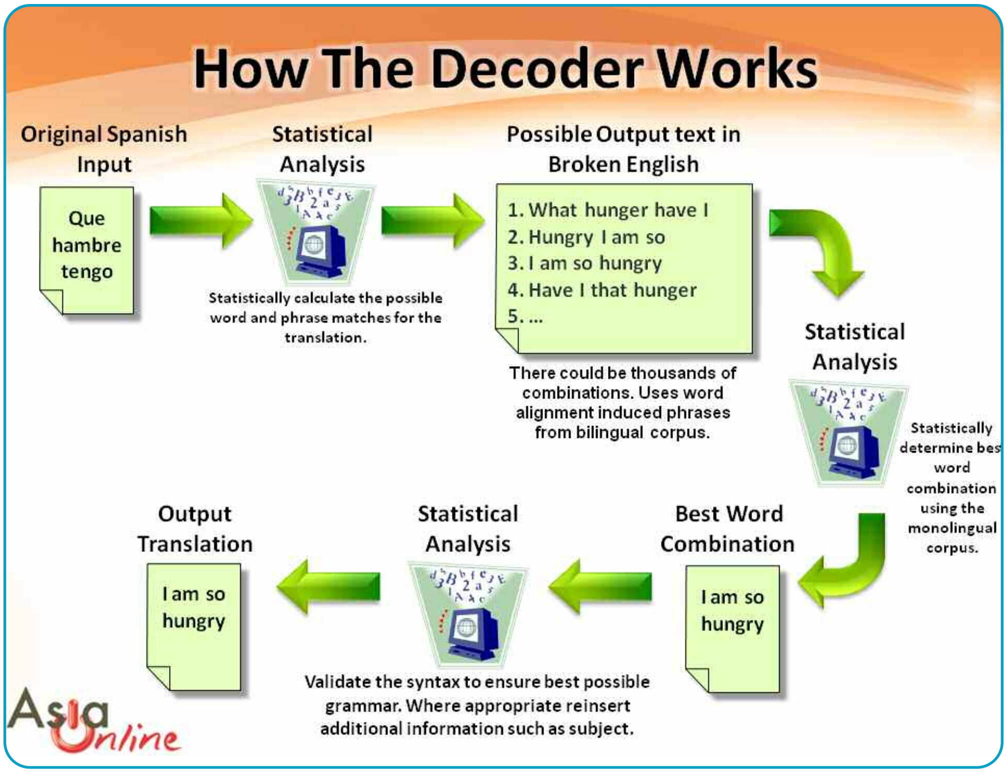
什洛莫介绍说，手术后，沙龙的颅 压有所降低，第一次手术后残留的部分血块也被清除。

Shlomo said that after surgery, Sharon's skull pressure has been reduced and the part of the blood remaining from the first surgery has been removed.

最好	Best Appropriat Better Outperforr Finest	19.0% 7.3% 2.3% 2.2% 2.0%
世界	World Globe Globalizing Worlds	4.7% 3.7% 3.1% 2.5%

of	the	0.13
	a	0.09
	another	0.01
	some	0.01
hong	kong	0.98
	said	0.01
	stated	0.01

This is a good English sentence with correct grammar. This is analyzed by the computer to learn how people express themselves. One should only use the best examples of modern English to train the system. Remove noise and incorrect language for best results. Avoid slang and incorrect usage for best results. The city of Hong Kong is a major gateway to China. This is a test sentence for the slide.



How The Decoder Works

Original Spanish Input

Que hambre tengo

Statistical Analysis

Statistically calculate the possible word and phrase matches for the translation.

Possible Output text in Broken English

1. What hunger have I
2. Hungry I am so
3. I am so hungry
4. Have I that hunger
5. ...

There could be thousands of combinations. Uses word alignment induced phrases from bilingual corpus.

Statistical Analysis

Statistically determine bes word combination using the monolingual corpus.

Best Word Combination

I am so hungry

Statistical Analysis

Validate the syntax to ensure best possible grammar. Where appropriate reinsert additional information such as subject.

Output Translation

I am so hungry

Figure 3

called "data-driven" approaches. Simply speaking, SMT systems are developed by computationally analyzing large bodies of parallel bilingual text, which they treat as strings of characters, determine patterns, and exploit these regularities by matching these patterns in new material that is presented for translation.

They learn by feeding on parallel corpora of texts, ingesting huge volumes of already translated content. SMT developers require large amounts (millions of TUs if possible) of both parallel corpora and monolingual corpora to train their systems. Parallel corpora consist of perfectly aligned texts in source and target languages, similar to translation memories. Typically,sentences and phrases are used rather than individual words or very large paragraphs. The statistical software picks out two-, three-, and four word phrases (ngrams) and so on usually up to eight-word phrases, in the source language that match the target language. These ngrams are then used to produce several hundred or even thousands of hypothetical translations for new sentences. The monolingual corpus in the target language is also used to "calculate" the best word and phrase combinations for these already translated text and help the system determine the best (statistically most significant) single translation to present to the user. This is done using the language model (LM), which is a statistical model encapsulating fluency and usage frequency in the target language. Disambiguation is much less of an issue for SMT systems as SMT systems have a much better sense of context. (see Fig. 3)

Thus, the internet is a natural source of material that can be used to train and build SMT engines. Also, this is essentially a technology that requires data processing and computational linguistic skills more than real bilingual competence in the development process. However, competent bilingual humans involved in vetting the training data and the output can add considerably to the quality of the systems. The EU has, in a few short years, generated 460+ SMT language pair combinations (23 core EU languages into 22 possible targets) to facilitate production of content into all the different EU languages. This Euromatrix project has also spawned a thriving open source SMT movement.

As with RbMT there are, however, several core problems with SMT:

1. **Data requirements**: SMT needs very large amounts of bitext (parallel bilingual data) to be able to build good systems. "Large" in this case can mean tens of millions of sentence pairs, which is often very difficult to find.
2. **Randomness of errors**: Often when dealing with such large amounts of data, SMT developers will scrape the web to collect data and introduce "noise" into the core training corpus. Noise refers to errors introduced by training on erroneous data such as mistranslations, wrong alignment,

and MT rather than human translations in the training material. This "dirty" data can produce many strange and unpredictable error patterns that are hard to eliminate. Clean data however, appears to dramatically reduce this problem.

3. **Transparency**: One does not have enough control and ability to tweak the system, and many complain it is too much of a black box. Users have complained that often the only refrain they hear from developers is to get more data (which is usually not available). This is changing with 2nd generation SMT vendors who are opening up the insides and are letting users see how the SMT engine comes to its translation conclusions. This allows the possibility of dynamic improvement and correction with translator feedback.

4. **Power of the language model**: The SMT Language Model can often produce very fluent and natural-sounding, but incorrect translations.

5. **System resources**: Many powerful computers are necessary for both training and running the final systems, so it can only be a server-based solution for the most part.

6. **Difficulty with unlike LPs**: Some language combinations work very well (FIGS, BP[1]), but language pairs that have large differences in morphology and syntax do not do well with the basic pattern-matching approach. While often more successful with Asian languages than RbMT, SMT too has a long way to go with Asian languages, which have many special issues.

7. **Lack of linguistics**: The first generation SMT systems are direct transfer systems that have no knowledge of linguistics, and thus have serious word order and morphology problems. As SMT systems evolve, they would begin to incorporate linguistics and produce even better quality. We are beginning to see the first of these emerging in 2008.

If enough data exists, it is possible to create a system out of a single customer's data.

Customization of SMT

SMT systems are very easy to tune to a specific customer's requirement if there is enough data available to do so. The most critical ingredient is a sizable amount of bitext comprising 100,000 or more TUs. In addition to monolingual content in the target language and domain, glossaries and other terminology assets can also be helpful for developing a language model. If enough data exists, it is possible to create a system out of a single customer's data.

In addition to the basic ease in getting a system customized, SMT systems can be designed to use and respond to real-time corrective feedback and are very well matched to massive online collaboration. SMT customization can involve frequent error modifications and corrections and systems that can actually get better over time as users use the system. Every correction adds to the "linguistic knowledge" of the system and so it continues to "learn". Theoretically it is conceivable that these systems can become quite compelling in their quality.

The Current Status Quo

Today, both approaches can claim success in many different kinds of applications. RbMT systems are in use in the EU, and at Symantec, Cisco, Fortis bank and many other places to enhance translation productivity and accelerate translation work. The PAHO (Pan American Health Organization) system is one of the most respected and actively used MT systems in the world, with a tightly integrated post-editing capability built into it. Several vendors offer RbMT solutions that can also run on the desktop and are sometimes (rarely) used by translators as a productivity tool. These vendors, which include Systran, ProMT and even SDL and Lionbridge, have relatively weak offerings. Additionally there are vendors who focus on regional languages like Apptek, Sakhr (Arabic, Middle Eastern) and Open Logos, BrainTribe and Linguatec (German). The

[1] FIGS- French, Italian, German, Spanish; BP-Brazilian Portuguese

Hindi MT: Google Translation Center and Matra, C-DAC

—Roomy Naqvy

Going to iGoogle, the Google Translation Center, and checking out the machine translation from English to Hindi is an interesting experience. The phrase ‘I like you’ translates perfectly into Hindi as मैं आपको पसंद करता / करती हूँ. On the other hand, if you enter the phrase ‘I like you’ at Matra, which has been developed by C-DAC, Mumbai, http://202.141.152.9/matra/index.jsp , it gives you the correct translation with a grammatical error in the Hindi sentence. Hindi is a gendered language, where the gender of the speaker would change the verb. There are other examples of Hindi being a gendered language, where nouns may take masculine or feminine gender and this can be quite illogical because ‘I cut my moustache’ certainly spoken by a male speaker would be translated as मैंने अपनी मूंछ काटी, where the noun moustache takes the long ‘ii’ sound, which is the feminine gender. Going back to the Google Translation Center, the phrase, ‘I like coffee’ gets translated as मैं कॉफी की तरह [I am like coffee] and the Hindi sentence is grammatically incomplete. However, if you go to Matra and enter ‘I like coffee’, it gets translated correctly if spoken by a female speaker. There is no male speaker equivalent given there. This is certainly an improvement over Google. Entering ‘I like tea’ at Google Translation Center gives you चाय की तरह मैं [tea is like I] with the Hindi syntax upside down. Matra translates the same sentence correctly but does not give the male speaker version.

Note: Notice the difference between ‘I like coffee’ [मैं कॉफी की तरह] and ‘I like tea’ [चाय की तरह मैं] in the Google versions in Hindi.

Japanese also have a whole suite of RbMT systems to their credit, with Toshiba and Fujitsu having the best reputations. Many RbMT companies have started and failed along the way. Systran is probably the best known name in the RbMT world and has the broadest range of languages available. However, the reputation of MT in general has been based on these systems, and several of the RbMT systems we see today are the result of 30+ years of effort and refinement. Many say now that RbMT systems have reached the limit of their possibilities and that we should not expect much more evolution in the future.

The SMT world is where most of the excitement in MT is today. Perhaps the most successful MT application in the world today, the Microsoft knowledge base, used by hundreds of millions of users across the globe, is mostly a SMT-based effort. Today, Google and the Microsoft Live free translation portals are powered by SMT. In just a few years, SMT systems have caught up in quality to RbMT systems that have decades of development efforts behind them. Given that we are just at the start of the SMT systems technology, which today are mostly “phrase-based SMT” (PBSMT) and really just simple direct transfer systems, there is real reason for optimism as these systems start to incorporate linguistics, add more data and get access to more computing power. SMT systems are also much better suited to massive online collaboration. Commercially, there are now several alternatives available from vendors like Asia Online, Alfabetics, ESTeam, Languagelens, Language Weaver, and probably many others will be available in the coming years. There is a growing open source movement (Moses) around the technology that is already outperforming the systems produced by SMT pioneers. Several automotive companies, Intel and others have implemented SMT-based translation productivity or technical knowledge base systems.

In a recent report on post-editing best practices, TAUS reported that, “In theory and also in practice, data-driven MT systems combined with machine learning systematically improve the output, reducing the post-editing load over continuous cycles of translation and machine learning.” They also state that “translation memory output is increasingly being seen as part of MT output and the two are being compared in terms of post-editing practices." In this report, they describe a test at Autodesk comparing quality and productivity of RbMT vs. SMT systems on the same data. Autodesk discovered that for its particular content, SMT outperformed RbMT on every indicator, especially post-editing productivity, where a rate of 1,340 words an hour (4 or 5 pages an hour or 30 pages a day) was achieved. However, there are also studies that show that post-editors often prefer the more systematic error patterns of RbMT output. So no definite conclusions can be drawn yet, but it is clear that SMT is on the march.

The real promise of SMT is yet to come. All the systems we are looking at today are essentially first generation systems. They will only get better and more robust in years to come.

What is a Hybrid System?

It has become very fashionable and desirable for developers to describe their approaches as hybrid. Many informed observers state that both RbMT and SMT need to learn and draw from each other to evolve. The expert opinion is that hybrid systems are the answer.

Systran 6.0 now uses a statistical language model approach to improve the fluency of its output. ProMT is also adding similar capabilities. Several SMT developers like Microsoft, Asia Online and LW are experimenting with syntax-based SMT. Thus, SMT is evolving from pure non-linguistic, pattern-matching techniques to the incorporation of grammar and parts of speech. Most SMT systems I am aware of already use some rules in pre- and post-processing.

COMPARATIVE OVERVIEW:	RbMT	SMT
Background	In development since early 60s and many systems have been around for 30+ years	Various systems based on IBM patents have emerged since 2004.
Language Coverage	40 to 50 language combinations after 50+ years of efforts. New LPs take at least 6 months to years to develop.	Over 600+ language combinations developed in less than 5 years. EU alone has 462 engines built out of Euromatrix project data. Development is possible wherever bilingual data is available.
Customizability	Based on complex dictionary and rule set modifications	Easily done when domain specific data is available.
Effort to Customize	Complex, long and expensive	Easy if data and computing resources are available
Quality Trends	Has essentially reached a plateau and has remained fundamentally the same for many years.	Most systems, especially Google systems have been improving rapidly and are expected to continue to improve over the coming years. Best quality approaches human draft quality.
Community Collaboration	Very little ability to incorporate community feedback except for dictionary contributions.	Microsoft, Google and Asia Online actively seek and incorporate massive "crowd" collaboration to correct raw MT and enable rapid improvements in quality. This practice may set this SMT quality apart from all previous MT.
Resource Requirements	Relatively low-end desktop installation is also possible. A single server can run 10+ language pairs.	Significant computing resources are required to both build and run SMT engines and not really suitable for single user desktop installation. Better suited to be a server, computing cloud-based solution.

This will continue as linguistic information enters the SMT development process and RbMT engines try to incorporate statistical methods around their rule structures. However, all the systems in active use today are predominantly one or the other.

The Future

As we look at MT technology today, we see that while there have been decades of failures, SMT appears to be the increasingly dominant way of the future. However, we still have some distance to go. Raw MT technology still falls short of any widely understood and accepted standard of quality.

The path to higher quality has to involve humans. Language is too complex, too varied and too filled with irregularities to be resolved just by algorithms and computers with lots of data. One of the most promising new trends is a movement to a more intensive man-machine collaboration in large-scale translation projects. The combination of SMT with massive online collaboration could bring us to a tipping point that really does help MT technology to become more pervasive. Microsoft has already led the way in showing how much value domain-focused MT can provide to a global customer base. They are now adding crowdsourcing into the mix and are having their best resellers manage crowd-sourced editing to improve the raw MT that is in the bulk of the knowledge base. Asia Online has embarked on a project to translate the English Wikipedia into several South Asian languages to reduce information poverty in these regions. Initially, internal staff linguists post-edit and help raise the quality of raw MT to a level where 100% comprehensibility is reached. This still-imperfect content is then released and people who use the content, or roaming bands of bilingual surfers, come and help to put finishing touches to the output. Some do it because they want to help and others because they might win prizes. These edit changes flow back into the learning systems and the systems continue to improve. Based on early results, it is conceivable that the translation quality will get to a level where students could use it directly in homework assignments with very minor grammatical adjustments. Even Google invites the casual user to suggest a better translation. These efforts will steadily drive SMT quality higher.

So, is MT possible in your future? It is important to note that the best MT systems will need to get the respect of real translators, professional and amateur. These translators will only be interested in using these systems if they have evidence that they can work faster, more efficiently and more effectively using the technology.

Also, real translators are among the most competent people to judge what is good, and what is not, on issues related to translation. Until MT vendors are willing to submit to their judgment and earn an approval or even an endorsement from them, the MT market will stumble along in the doldrums as it has for the last 50 years, making empty promises.

The combination of massive computing power, ever increasing volumes of clean bilingual text, and a growing band of motivated bilingual humans (not always professional translators) will be the key forces that drive this technology forward. The scientists may also produce better algorithms along the way, but their contributions will be relatively small. The most important driving force will be: the human need to know and understand what other humans across the globe are saying, the need to share, and the urge to learn. The breakthrough that will end (or at least reduce) the language barrier will be social, not technical. ▢

Note: I should state for the record that I am currently employed by Asia Online, which has developed a largely SMT based technology platform and that I have also previously been associated with other commercial SMT initiatives in the past. —*Kirti Vashee*

References:

Brantley, Peter .
http://blogs.lib.berkeley.edu/shimenawa.php/2008/11/02/losing-what-we-don-t-see-translation, 2008