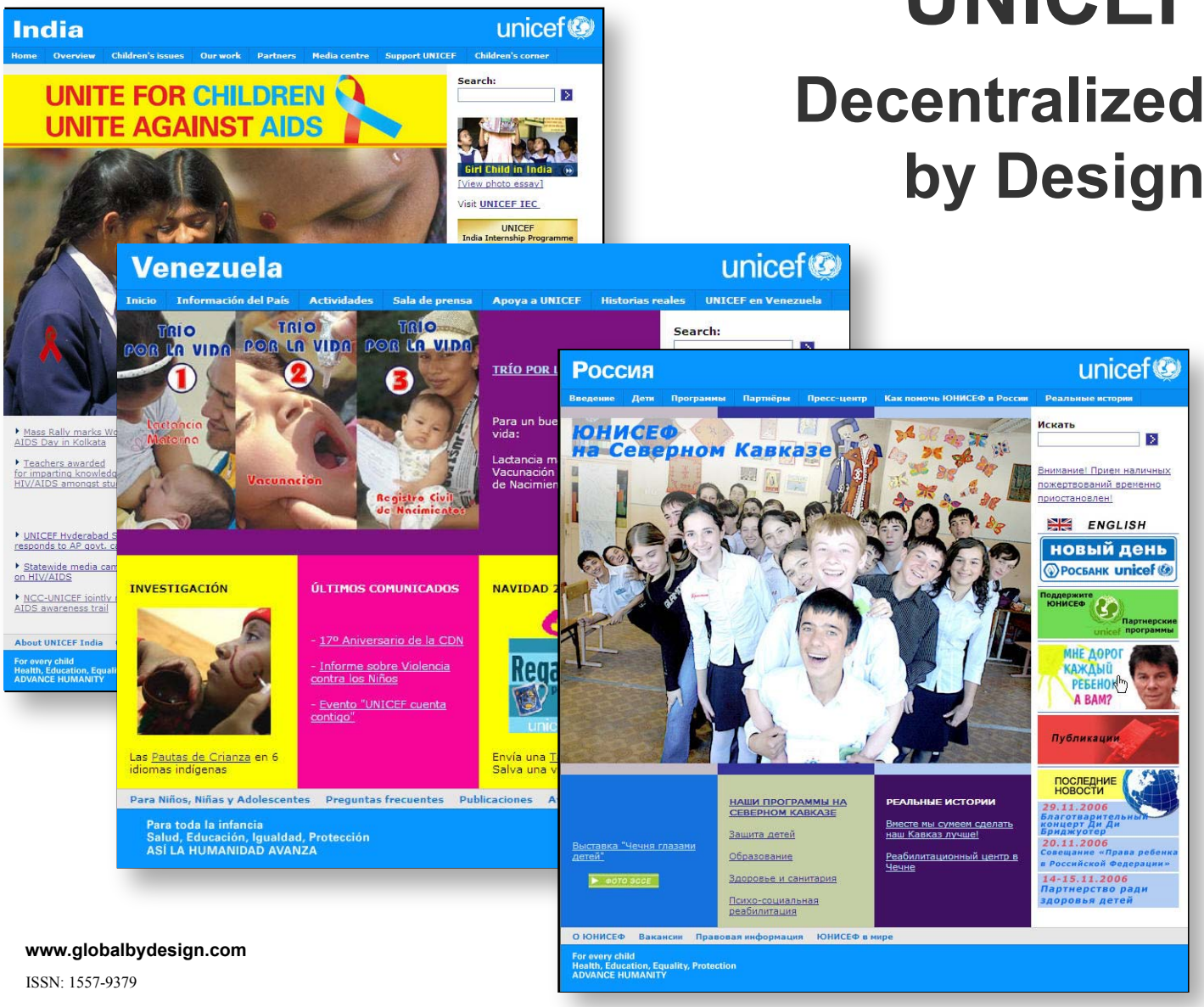
Global*by*Design

Web Globalization Intelligence

December 2006

UNICEF

Decentralized by Design



www.globalbydesign.com

ISSN: 1557-9379

GlobalbyDesign

Web Globalization Intelligence

December 2006

Did you know…

One of Japan's largest automakers is now using statistical machine translation to accelerate workflow?

According to a recent Translation Automation User Society survey, more than 60% of companies expect to use machine translation within the next 18 months?

CONTENTS

Web Globalization from the Bottom Up

Why the best local Web sites may come from surprising places

Sometimes the best way to help your local offices develop successful Web sites is to give them the tools, the training, the vision, and then get out of their way.

Don't overcomplicate things. Just create a framework that will allow them to start developing a local Web presence. You may be surprised by which local offices rise to the challenge.

Over the past few years, UNICEF, profiled in this issue of *Global by Design,* has been helping its field offices around the world develop local Web sites. Instead of a top-down strategy that mandated that certain offices go live before others, UNICEF let the offices that wanted to go first do so.

And the first local field office to go online with a Web site was Somalia, a field office operating in one of the most challenging environments in the world.

Not all local offices are equally motivated to develop local Web sites. So consider working first with those offices that are most motivated. All you'll need to do is give them the tools and direction, and then get out of their way.

John Yunker
jyunker@bytelevel.com

About Us

Every month, *Global by Design* explores the art and science of Web and content globalization. We cover emerging trends, real-world Web sites, the latest technologies, and innovative vendors.

Editor

John Yunker
jyunker@bytelevel.com
+1 760.317.2001

To Subscribe

www.globalbydesign.com

Global by Design
ISSN: 1557-9379
Published by:
Byte Level Research LLC

bytelevel research

Statistical Machine Translation Gets Real

The Language Weaver Profile

In August 2005, we first profiled Language Weaver (www.languageweaver.com), a developer of statistical machine translation (SMT) software. At the time, the company supported a handful of language pairs and was only just beginning to target the corporate market.

A lot has happened in a year. The company now supports 27 language pairs and recently announced its first translation agency customer: ITP (www.itp.co.jp/english/index.html), one of the larger translation and publishing agencies in Japan.

To view white papers and demonstrations of Language Weaver software, login to your *Global by Design* members portal at: **www.globalbydesign/members/home.php** and look under "Bonus Deliverables."

ITP-Europe purchased the Language Weaver French and Spanish language modules to support a client, one of the world's largest automobile manufacturers. By applying Language Weaver's SMT software to its client's large database of translation memories, the agency was able to accelerate translation workflow by 30%. And we're talking about really large workflow projects – of up to three million words per week. ITP plans to next add the German module in the first quarter of 2007 with additional European languages coming later in the year.

If you think SMT is something that only a large translation agency can appreciate, think again. Language Weaver is currently working on pilot projects with a few large corporations, and we hope to see their names announced sometime in 2007. When this does happen we expect Language Weaver, and the emerging field of SMT, to receive a lot more attention from the translation industry as well as their customers.

Don't believe that SMT is ready for prime time? Consider Google, which this year went live with SMT engines for Chinese and Arabic. Russian is rumored to be next in line.

Although we believe that SMT is a disruptive technology that will transform the translation industry, it's also a complex technology and it has its limitations.

So we set out to learn more about SMT and Language Weaver.

Our Visit to Language Weaver

Last month, we traveled to the Language Weaver offices in Los Angeles and met with senior staff of the company, including Bryce Benjamin, CEO, and Kirti Vashee, VP of Sales and Marketing. The offices were recently expanded to accommodate a staff that now includes 40 employees.

In this article, we explain how SMT has evolved, how it works, and why it poses such great potential for companies and their agencies.

A Brief History of Machine Translation
Machine translation (MT) – the process of using computers to automatically translate text from source text to target text without human intervention – has been around for half a century.

To "teach" a computer to translate a sentence, a computer first needs a dictionary of source and target terms for a specific language pair (such as English <> Chinese). This dictionary is then paired with a software engine that applies linguistic and grammatical rules, such as: "A verb comes before the noun." This type of MT is known as "rules-based" MT.

Yet even the most comprehensive sets of rules may fall short when faced with words that may have multiple usages. For example, "bank" may be a noun or a verb, a subject or an object, all depending on context. And context means everything in machine translation.

That's why human translators are generally required to assist in pre-editing or post-editing the text, or both. Companies and organizations that have had the best success with MT generally establish strict guidelines for the source text. For example, shorter, less complex sentences are easier to translate. And consistent use of terminology is critical. You don't want to refer to an airplane as a "jet" in one sentence and "plane" in the next. The many style and terminology guidelines a company develops will result in something commonly referred to as a "simplified" or "controlled" language. There is even software available to help companies enforce the use of a consistent terminology and writing style, such as Acrocheck (www.acrolinx.de).

Of course, developing this workflow and getting buy-in from writers is no trivial task. And there is still the widely held perception that MT just isn't ready for prime time, a perception largely fueled by the popularity of free translation portals powered by Systran.

Systran Takes MT to the Masses
The odds are pretty good that you have used a Systran (www.systransoft.com) product by now, even though you may not be aware of having done so.

For example, you may have used the free translation Babel Fish portal, shown here.

The Babel Fish Translation Portal

Source: AltaVista

Or perhaps you have translated a Web page while using Google or Yahoo!, two Web sites that license Systran software. Systran, founded in 1968 in La Jolla, California, provides a range of rules-based MT solutions – from software for personal use all the way up to enterprise products used by governments and multinationals. But Systran is best known for its free "gisting" software, available in 20 language pairs.

Because the free MT engines are not customized to a specific industry or company and because the input is anything but controlled, output is at best only going to give you the gist of the meaning. But the gist is better than nothing, and Systran has proven extremely popular in recent years.

According to Dimitris Sabatakakis, CEO of Systran, his company's software is used to translate more than 37 million pages of text every day – which means that machine translation is already translating more text daily than all the world's human translators combined**.**

And MT software can be used to provide translation that goes beyond mere gisting. When Systran is customized to a specific organization within a specific industry, and a strict style guide is enforced, the results can be highly effective. The European Union, for example, has been successfully using a customized version of Systran for more than two decades, as have many departments of the US government. The corporate world is also using Systran effectively. Ford and DaimlerChrysler use Systran to manage translation of English documentation into a number of target languages. A *controlled English* is used on the front end, and translators edit the work on the back end.

But despite the success of rules-based MT, adding a new language pair or customizing an existing language pair is highly labor intensive. Which is where SMT enters the picture.

The Next Generation: SMT
As the Internet makes available huge libraries of source and target content, software engineers have begun leveraging this content to develop MT software that takes a completely different approach to translation.

Enter Language Weaver. Founded in 2002, the company got its start thanks to funding from DARPA (one of the research arms of the US government) the National Science Foundation, In-Q-Tel (the venture arm of the CIA), and private VCs. Although the company is now aggressively marketing itself to translation agencies and their customers, most of the company's revenues today come from the US government.

The company's products are used to provide real-time monitoring of broadcasts, Internet chat, and Web sites, as well as multilingual search. The software also can be installed in desktop and handheld computers to provide real-time, in-field translation – and is being used right now in Iraq.

SMT is a data-driven translation technology. Rather than relying on a dictionary of translations and rules, it starts with data in the form of lots and lots of source and target text, known as "parallel text corpora." The statistical process involves analyzing this data and identifying patterns. By analyzing millions and millions of words, the software gets pretty good at *guessing* how to translate a given text string. "We're not really translating," said Language Weaver CEO Bryce Benjamin. "What we're really doing is a probability forecast."

Sources of this corpora have come from The United Nations, news organizations, IT product support Web sites – even from translations of *Gone With the Wind*. Language Weaver secures permission from publishers to capture the relationships between the source and target data, not the data itself.

MT vs. SMT
So what makes SMT better than MT? Here are the key advantages:

> **Rapid Addition of Language Pairs**
> Because SMT does not require an exhaustive process of rules building for each language pair, new language pairs can be added relatively quickly. Language Weaver has added roughly 20 new language pairs in the past year alone. Assuming that a large corpus of source and target text is available, getting an SMT engine up and running may take a matter of weeks instead of the years normally required to assemble dictionaries and rules.
>
> **SMT Gets Better with Age**
> SMT software has the ability to "learn" from its accurate translations. Mistakes can easily be tweaked with changes to the algorithm and new corpora added to further improve the quality of the engine.

Rapid Customization and Deployment
Language Weaver can get a company set up and running in just a few weeks. The customization process is simply a matter of inputting massive amounts of corpora – a process that the customer can even manage by itself.

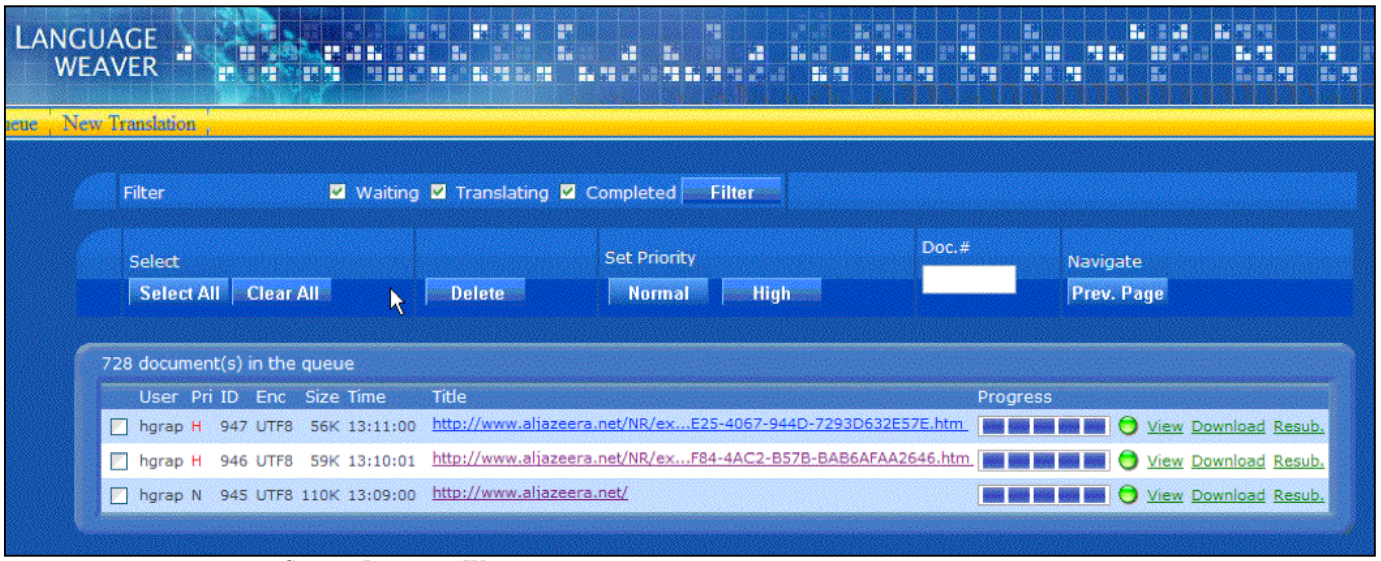
Where SMT Falls Short
Because SMT relies on previous translations to develop an accurate engine, the lack of corpora can be a major obstacle to quality. So it's not surprising that developing an SMT engine for, say, Portuguese <> Chinese is not going to be as easy to accomplish as developing the French <> English engine. But this obstacle isn't unique to SMT. Developing a rules-based MT engine for these more challenging language pairs would require bringing in linguistic experts in such language pairs – also no easy task.

So even though Language Weaver has launched a great many language pairs recently, we expect that rate of growth to slow as it tackles language pairs that feature a dearth of corpora.

How to Use Language Weaver Software
Language Weaver's flagship product, now in version 4.2, can translate 5,000 words per minute per CPU and is scalable to up to half a million words per minute. A standard server is roughly 4 Gigs, but can be scaled down to as low as 1 Gig for portability.

Below is what the basic interface looks like. Note the list of Web page URLs. The software can translate any type of text-oriented document, from Word file to HTML file.

Language Weaver Interface



Source: Language Weaver

When a company purchases the software it could, if it chooses, begin using the "out of the box" baseline system. However, most organizations will customize the software to their industry terminology and corporate style. Language Weaver has developed a Customizer software application that companies can use to do this work themselves – or let Language Weaver manage. Customization is when the software is fed all prior translation memories, dictionaries, and other available corpora.

Shown below is a source sentence and three translations of this sentence. Notice how the quality improves as the Language Weaver software is customized and improves even more if a second iteration is applied to the text.

Original:
Detects whether the parking brake lever is applied.

Language Weaver baseline system:
Detecta si el freno de estacionamiento palanca es aplicada.

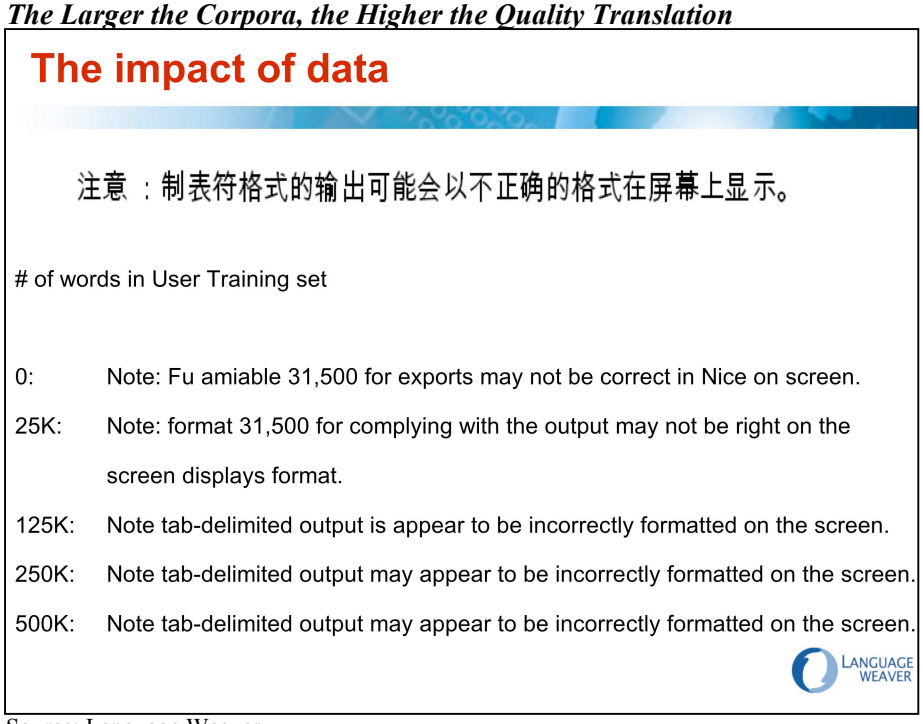
Language Weaver customized:
Detecta si la palanca del freno de estacionamiento se aplica.

Language Weaver customized - 2nd iteration:
Detecta si la palanca del freno de estacionamiento está accionada.

How Much Corpora Is Enough?
For SMT to be effectively customized to an industry and company, it requires a sizable database of previously translated text strings. As a general rule, this baseline corpora should include at least 5 million words. But this is a loose rule, as much depends on the industry and the language pair. Some language pairs, like Swedish <> English, require substantially less corpora than pairs such as Arabic <> English.

As shown here, you can see just how the number of words in the corpora affects the quality of the results.

The Larger the Corpora, the Higher the Quality Translation



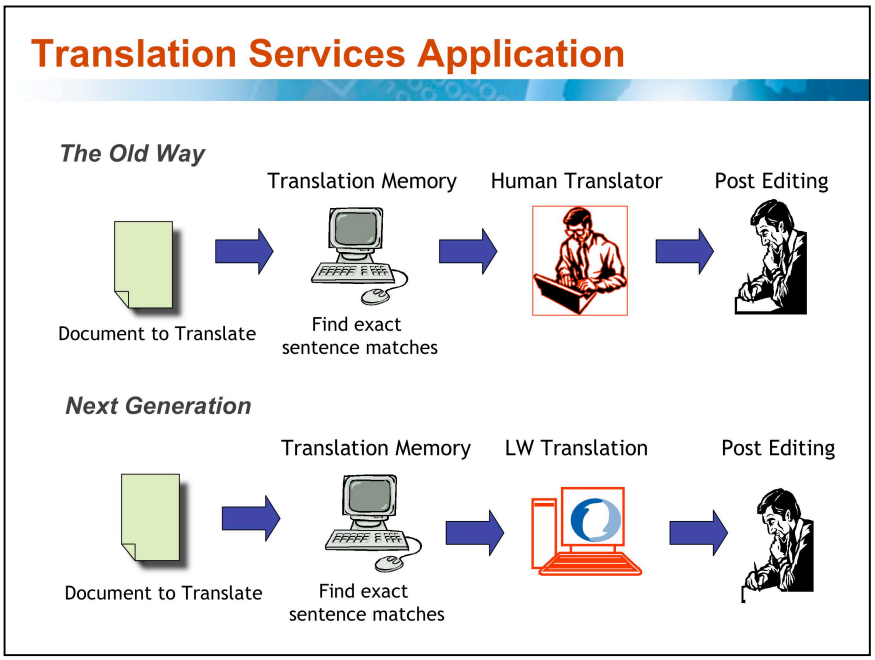
Source: Language Weaver

How Companies Use Language Weaver

SMT software is not designed to replace existing translation memory tools; in fact, it depends quite heavily in translation memory for customization and for ongoing quality improvement.

Below is an illustration of how Language Weaver would fit into an existing translation environment:

How SMT Integrates Into Translation Workflow



Source: Language Weaver

Once a company has SMT software, its use is not limited to just providing content for prospects and customers. There are other uses that may also justify the investment. For example, the software is currently used for business intelligence, through the monitoring of transcribed broadcasts and multilingual search.

Business Intelligence

The technology can be applied to broadcast monitoring. By partnering with a company that transcribes broadcast text, Language Weaver is able to provide real-time translations, as shown here:

Real-Time Monitoring of an Arabic-language Broadcast

Source: Language Weaver; Virage

Multilingual Search

Language Weaver has worked with enterprise search companies such as Convera and FAST to integrate this functionality within their products. This allows an enterprise to make its content available to all employees, across multiple languages. Shown below is a search application developed by MITRE to provide English-language search results from Arabic-language queries.

MITRE "Clipper" Search Engine

Source: Language Weaver; MITRE

Making SMT Pay for Itself

So now for the moment of truth: *What does Language Weaver's software cost?*

European language pairs range between $7,500 and $25,000 and "premium languages," such as Asian languages, range between $25,000 and $125,000. The pricing variation depends upon the language pair, the directionality (unidirectional vs. bidirectional) and the licensing terms.

Okay, so how quickly does the software pay for itself? Language Weaver cites numbers provided by one of its translation agency clients which shows a productivity improvement of between 200% and 300%. If we assume that this translates to a savings of two to three cents per word, getting to payback is not very difficult for companies that translate a few million words per year, per language. But cost savings is just one way a company can justify the investment.

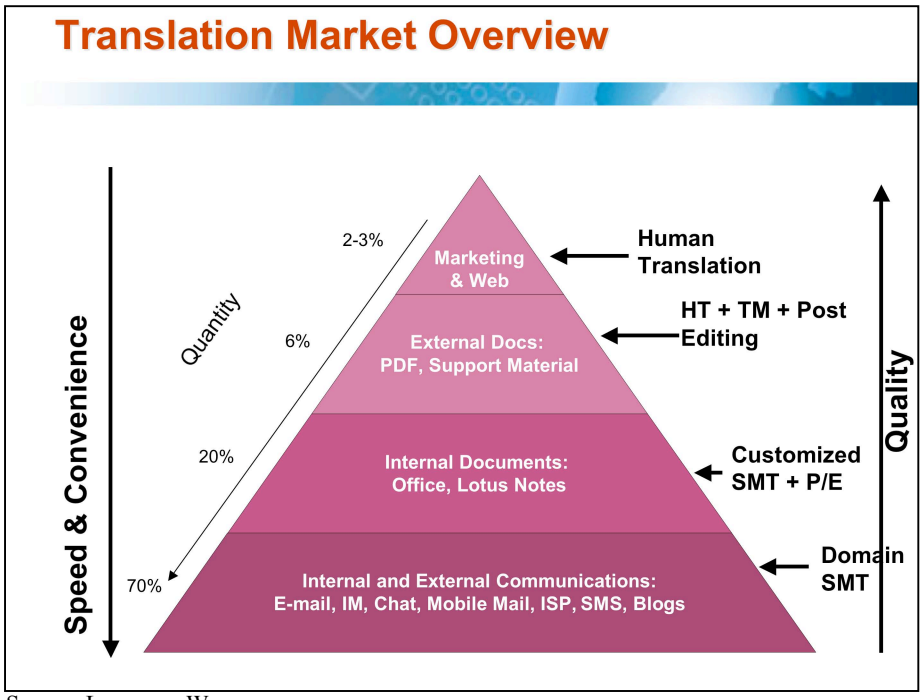
Unlocking "Hidden" Content

Potential revenue gains from additional translated content is another, potentially more significant metric that companies can apply to their SMT investment.

Language Weaver estimates that companies translate only about 3% of their content, leaving the rest untranslated due to lack of budget. The other 97% of content is effectively locked in the source language.

Below is a Language Weaver slide that illustrates what content SMT is best-positioned to translate: Internal documents and communications as well as rapid-turnaround external content, such as blogs or chat.

Translated Content: SMT vs. Humans



Source: Language Weaver

One of the great misconceptions about machine translation is that it will replace human translators. More often than not, MT software is used to translate content that would never have been processed by human translators in the first place. Translators and their agencies are actually well positioned to leverage SMT to help them move up the "value chain" away from *first-pass* translators to proofreaders, editors, and even consultants.

A More Affordable Business Model?
Nonetheless, whether a company uses cost savings or revenue increases to justify the investment, the investment is not trivial given that this technology does not replace existing translation tools. And $50,000 is simply too tough an internal sell for many executives, particularly when they must overcome the negative perceptions of machine translation.

So Language Weaver says it is close to launching a new business model that allows companies to purchase translation in 100,000-word increments. Language Weaver is assembling bundles of language pairs – such as a European bundle and Asian bundle – and pricing will average between two and five cents per word. This model gets around the large upfront costs a company faces and also takes Language Weaver a big step closer toward a subscription business model, which might resonate widely.

Looking Ahead: More Languages; More "Confidence"
Also on the horizon, Language Weaver will continue adding language pairs. It will also begin adding in a layer of linguistics intelligence to its language modules to further improve quality.

In terms of software features, one feature we would like to see added is a "confidence indicator" that lets the user know just how confident the software is in the translated text. This may be as simple as a percentage placed next to the text or the changing of the text color. The more confident the software is in the translation, the more confident readers can be in making decisions based on this information. Language Weaver says this functionality is coming.

Any Questions?

Because SMT is such a new technology and can be used in so many ways within a company or agency, expect to ask lots of questions as you begin to evaluate the software.

Here are a few questions that we posed during our meeting with Language Weaver:

> **What if we outsource everything to a translation agency? Shouldn't the agency purchase SMT software?**
> Clearly, Language Weaver would love all translation agencies to become SMT customers. But the software requires a significant investment and agencies generally make purchasing decisions based on what their clients ask them to do. So talk to your agency about SMT and get them involved in the decision-making process. It's important that both you and your agency are in agreement about the potential benefits and risks of implementing SMT.
>
> We expect to see agencies gradually adopting SMT software, but slowly. The evolution of translation memory software was driven by clients mandating that their agency support the software. It could be that SMT follows a similar pattern. But if you want to be an early adopter of SMT, keep your eyes on *Global by Design* in 2007 – we

expect to see more announcements from early adopter translation agencies.

Can my company purchase SMT software and still work with external translation agencies who don't have SMT software?
Absolutely. In fact, companies are likely to use SMT in many different ways; translation for an external audience is just one such application. However, the key when integrating SMT is to have a clear focus on how your company will use the software now and in the future. Develop a plan that rolls out SMT in stages, to ensure that you get a success story or two as quickly as possible. For example, you may first use SMT to provide your global marketing teams with business intelligence internally. This limits any potential external issues and allows you to get the marketing team excited about what SMT can do. Then you can begin to integrate SMT into public-facing areas of your company.

Do we have a large enough corpora?
If your company is relatively new to translation and your translation memories consist of less than a half a million words per language pair, SMT might not be a good fit, at least not yet. For this software to excel, it needs to digest a lot of previously translated text as part of its customization phase.

We can't just provide machine-translated text directly to customers without any proofreader, can we?
Actually, you can, but you need to control the environment and manage expectations. Microsoft uses machine translation to supply a vast amount of text in its knowledgebase in Japanese and Spanish. There is a large disclaimer at the top of these pages warning the user that the content was not viewed by a human translator. Microsoft reports that customers have been very happy with the results – and would rather have translated text of average quality than no text at all.

That said, we do not see SMT being used to process mission-critical text without any human supervision. For example, medical device instructions or press releases will continue to require a human editor.

If SMT is such a hot technology, why aren't there dozens of competitors to Language Weaver?
Currently, there are two major players in SMT – Language Weaver and Google (and the person managing Google's SMT was formerly affiliated with Language Weaver). We do expect to see competitors coming to market over the next two years because this market does show great promise, but the fact is that this technology requires people with very specialized skills. The area where we see some interesting opportunities is for companies that play the role of "translation technology integrators." There are some small firms currently specializing in this emerging field, but we expect to see a lot more activity as the clients look for specialists to help them make the most of all these tools.