Where Does Language Fit in with Big Data?

For the diverse universe of digital content generated by big data to be useful, it requires transformation for different channels (such as web, mobile, and print), conversion for various applications, and localization for other markets. This is an area of opportunity for translators and interpreters.

Go to any conference and you'll find a few new additions to the usual buzzword bingo of industry jargon—"big data" and numbers with lots of zeroes. You'll hear about the massive growth in digitized data, how often a given sector's knowledge base doubles, and what companies are doing to manage and interpret that flood of data. This burgeoning trove of bytes includes structured databases, application code, images, videos, and text. You'll also hear about machine learning and how big data contributes to making software more responsive and useful to customers' needs.

Just how much data are we talking about? Already huge, the digital universe of content, code, and structured data grows by a mind-blowing amount every 24 hours. Each day the world creates another 2.5 quintillion bytes of data.[1] This data comes from many sources, including documents, social media posts, electronic purchase transaction records, and cellphone GPS signals. That daily infusion is estimated to pump the global repository of information from the 7.9 zettabytes (7.9×10^{21} bytes) available in 2015 to 176 zettabytes by 2025.[2] Keep in mind that 1 zettabyte equals 1,000,000,000,000,000,000,000 bytes—an incomprehensible number.[3] And that total doesn't include the inestimable amount of content that is spoken every day.

Whatever the content being created, this truly immense volume includes massive, unrealized potential for translation or localization. So, what does this mean for the language industry, both humans and machines?

WHAT IS BIG DATA AND WHY DOES IT MATTER?

When we talk about big data, we refer to new ways of taking large amounts of data and using software tools to identify previously undiscovered patterns, trends, correlations, and associations. If you've ever bought a book because an online retailer told you that customers with viewing histories like yours have enjoyed it, you've been the beneficiary of big-data analytics.

This practice became possible because of the digitization of business, government, and everyday life over the past few decades. This information is stored in massive databases of structured data and repositories of documents large and small. We feed this growing beast with more bits and bytes every day. While all organizations rely on data to run their operations, a small but growing number use it to better understand behaviors, preferences, and trends in their world. Then, using those insights, organizations can make better decisions about how they market their wares, help their customers, improve operational efficiency, or build the next great thing.[4]

How do they do it? It's not easy given the diversity of structured data and text. For highly structured data, software specialized to deal with big data draws from very large databases, often distributed around a network. Then, analysts employ a new generation of business intelligence and textual analysis tools to turn this raw data into usable information and actionable insights.[5] They may combine transaction data with server logs, clickstream data, social media content, and customer e-mail texts, sensor data, and phone records to extract insights. They also extract insights using advanced analytical tools, including statistical analysis, data and content mining, predictive analysis, and text analytics. Traditional business intelligence and modern data visualization software help analysts present their findings in human-readable formats.

The language industry was actually one of the first areas of interest for big data applications. One of the early mainstream

applications was in the statistical machine translation (SMT) efforts of Google and Microsoft. A 2011 Common Sense Advisory (CSA) report on MT trends characterized these statistics-based approaches to MT as big-data applications because they leverage large repositories of bilingual content. For example, they compare source documents in English to their human-translated Russian variants.[6]

In simplistic terms, SMT translates by comparing the zeroes and ones of the source file with the translation to find correlations and patterns. In other words, massive processing power allows computers to disassemble texts and their translations, analyze the patterns, and predict translations for texts they have never seen before. Such analytics has increased the speed of language support over earlier MT solutions that relied on teams of linguists to create grammars, code them as rules, create dictionaries of bilingual translations, and then constantly modify or add to the rules as they found exceptions.

The 2011 CSA report predicted that experts would apply these mathematics-based big-data algorithms to crack inter-language communication and marketing issues as they processed more languages and a huge volume of multilingual content. And that, in fact, is what has happened.

Over the past several years, MT based on big-data analysis has drawn far more usage than the first-generation rule-based solutions. Google Translate draws massive numbers of users, which is a testament to its easy access and perceived, if not actual, improvement of the quality of MT output. Although academic research shows improvements using popular quality assessment systems such as BLEU[7] (bilingual evaluation understudy), these changes are not cumulative and results vary widely between languages and translatable content types (e.g., regular text, audio, video, and social media). Thus, data on quality improvement is anecdotal and may be balanced by lowered user expectations for quality.

The availability of cloud-based computing with unlimited horsepower from the likes of Amazon Web Services and Microsoft Azure supports these big-data practices. This kind of harvest and analysis will continue to grow into the "Internet of Things" as many billions of devices come online (e.g., sensors, embedded controllers, wearables, health checkers, and widgets not yet invented).

To be useful, much of this content requires transformation for different channels (such as web, mobile, and print), conversion for various applications, and localization for other markets.[8] Corporate and government planners already know it's not enough to have all that digitized information available in just a single language. Their mission is to use as much data as possible to support customer experiences for the populations that really matter to them. Otherwise, it will be impossible to engage and retain international or domestic multicultural audiences.

Just consider the requirements necessary to translate that information into other languages to make it available to a broader audience. It's estimated that it takes 14 languages to reach 90% of the world's most economically active populations, but most websites max out with support for just six languages or locales.[9] Product and document localization at many companies lags even further behind. Spoken-language interpreting is even more limited.

As the volume of data organizations produce grows, so too will ambitions to reach a greater audience for goods and services. Client-side respondents to a recent CSA survey reported that they plan to increase translation volume by 67% over the next three years, from an average of 590 to 990 million words per year.[10] This increase is one that the language industry cannot meet with current methods, and that buyers in the CSA survey sample expect to address with a combination of post-edited content from their suppliers and raw MT.

WHERE BIG DATA FITS TODAY—AND IN THE FUTURE

Organizations are stating to realize that their plans for more translation could very well exhaust the capacity of all current translators, as well as those who will enter the field in the foreseeable future.[11]

To help keep up with the demand, many organizations are employing both productivity enhancements for human translators and MT to overcome the challenges associated with volume, turnaround time, the need to deal with more target languages, and flat budgets. Companies invest in human translation and post-edited MT for essential business content, such as product and marketing materials that are reasonably stable. For example, translation buyers rely on a large and growing cohort of providers that employ MT to pre-process the source material and then edit the output with human linguists. A small percentage of client-side organizations also use unedited MT output for business content, such as FAQs and knowledge bases.

Besides translating a limited set of business-oriented text, some buyers have increased the use of MT to process user-generated content, such as product evaluations, hotel reviews, and forum discussions that few organizations have bothered to translate in the past. But as research conducted by CSA indicates, online consumers and business buyers alike would prefer to have user reviews translated, even if these reviews are all that gets translated.[12]

WHY THE VOLUME OF BIG DATA CONCERNS TRANSLATION BUYERS AND SUPPLIERS

Big data represents enormous numbers, and it turns out that one day in the translation industry barely puts a dent in its volume. Let's focus on just the written word and how it relates to that 2.5 quintillion bytes of data being generated every day.

Despite today's objective of making humans more productive to save time and money, the world is far from the nirvana of having enough online content available in all languages. From years of research and consulting, we know that any discussion about whether or not to invest in translation, localization, and interpreting has to begin with a review of available data.

CSA decided to investigate the enormous challenges facing the localization industry in terms of translating what should be translated from the totality of all data that could be translated. We decided to start with a given day's output of digital content and determine what could actually be

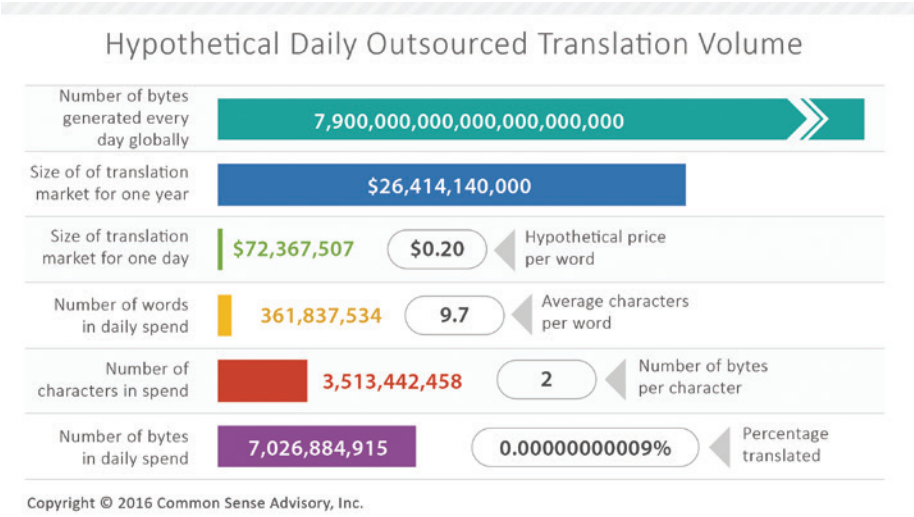

Figure 1: Hypothetical Correlation of Translation to Daily Content Creation (note: "daily spend" = daily spending on language services)

Source: Common Sense Advisory, Inc.

translated if we had the entire language industry working on just that content and none of the backlog of existing data.

What is this data? It's everything that is digitally created every day, from documents to SQL data and telemetry to digital multimedia. For this hypothetical exercise, we began with the expenditure for outsourced services. It's estimated that translation in various forms—human, post-edited, transcreation, plus website globalization and text-centric localization—accounts for US$26.4 billion of the $38.1 billion market for language services and technology.[13]

We then calculated the daily amount spent by the word. We divided US$26.4 billion by 365 days and estimated that the translation sector is worth US$72 million per day. At a hypothetical rate of 20 cents per word, we estimated that professional translators would process nearly 362 million words every day. We then converted that to bytes at the rate of 9.71 characters per word, which equates to seven billion bytes of double-byte characters. (Note that some languages have fewer characters per word on average and others have more).[14]

Finally, we compared it with the daily volume of content creation. When we divide the 2.5 quintillion bytes by the amount of target-language content produced by language services providers, we estimate that translation firms could potentially process just 0.00000000009% of the content created every day. However, we can safely assume that much of that data will never be translated—either the material isn't translatable or translating it doesn't make sense.

But some of what isn't translated today (e.g., user reviews and social media posts) is on the future agenda of enterprise translation buyers as they strive to improve the customer experience. Even if we exclude all but an infinitesimal percentage of those daily bytes, the amount of content outsourced for translation is far less than 1% of what's created every day. And remember that we are talking about the shortfall in translation for just one day. That number doesn't address the backlog of content not yet translated.

As the results from this hypothetical exercise indicate, if the content is translated at all, it's typically into just six languages online (and often fewer elsewhere). This is far short of the total number of online languages that really matter for both global and domestic communication and commerce.

Of course, there are many other variables and mitigating factors that affect these calculations. For example, consider in-house translation, languages for which you should translate but don't, and the many zettabytes of existing content. The bottom line is that there's an enormous amount of content that will never be translated or localized. That means opportunity for the language sector, and not just the technology companies.

WHAT BIG DATA MEANS TO THE LANGUAGE SECTOR

The big data and translation needs we discussed represent an opportunity for the language sector, but many translators look at the situation and worry that widespread deployment of MT will take work away from them. Our research estimates that translators will, in fact, lose some lower value jobs to MT, but that the total amount of work they have will increase at a steady rate for the foreseeable future.

If we also consider the expansion in post-editing—a contentious topic to be sure—we see that reliance on human professionals will grow faster than the current pipeline of future translators can add capacity. As a result, translators and interpreters will require productivity benefits from big data if they are to keep pace with demand. A few will take a much bigger step and become specialists who can build, train, and improve MT engines.

On the productivity front, we see that big data today trains statistics-based MT engines and could be used to supplement the post-editing processes of other MT models. Connections to MT are available in CAT tools such as Kilgray memoQ, Memsource Cloud, and SDL Trados Studio. Meanwhile, startups like Lilt use MT output in a CAT-like tool to accelerate human translation. We have also been briefed by software developers who are evaluating big-data machine learning techniques to improve terminology, translation memory, disambiguation, and a variety of other content creation, localization, and reviewing tasks. In short, big data will underpin most of the software tools translators use. Interpreters will also benefit as MT technology evolves for spoken languages.

What does big data mean for professional linguists? Just as they saw with translation memory and terminology management, linguists will have another tool at their disposal. Employers on both the end-buyer and agency sides will expect them to use this software to speed up their work and improve the usefulness

of the output because of improved analysis of the source content.

Our 2016 survey of language services providers found that 49% of respondents have already committed to post-editing MT as a service.[15] As early as 2012, our research showed that 21% of freelancers had experience using the technology.[16]

Some will move away from the classic translation agency structure to become big-data specialists. They will create clusters of industry- and domain-specific memories and harvest, analyze, and translate content. Content curation positions in which language professionals work with data applications to integrate relevant results to "enrich" them with useful metadata (e.g., topic categorization, classification of names and entities) are just now emerging.[17] These positions will allow localizers to add market-specific value to content. Some will take the next step into the global marketing mainstream, adding to their portfolio services such as transnational business intelligence to help companies better understand their markets, or cross-language semantic and sentiment analysis to cull the opinions of consumers and business buyers out of multilingual content.

Big data has increased the volume of content dramatically. At the same time, automated content enrichment and analytical tools based on big-data science will enable the training of more sophisticated tools to help humans translate the growing volume of content and enable machines to close the yawning gap between what's generated and what's actually translated. No doubt some linguists will view these big-data-based innovations as threats. Others will view such advances as opportunities that will help them enhance the meaning of the source content, increase the usefulness of the other tools they employ, and increase their productivity in the process.

Although it has not happened yet, we speculate that MT driven by these phenomena could remove the "cloak of invisibility" from translators, giving them greater recognition and status.[18] Even if machines generated the lion's share of translation and humans did a smaller percentage, the sheer absolute volume of human translation would increase for high-value sectors such as life sciences, other precise sectors, and belles lettres. In turn, the perceived value of human translation could increase. Why? Because when you bring in a live human, it means the transaction is very, very important. It's not so different from accounting. Software can handle routine tasks, but when problems arise or something is critical, you bring in a high-paid accountant to deal with it.

As interlingual communication becomes transparent, we predict that the number of situations where high-value transactions occur—i.e., those requiring human translators and interpreters—will go up, not down. If provider rates increase and companies use MT to address a larger percentage of their linguistic needs, human translators could benefit as they're paid well to render the most critical content supporting the customer experience and other high-value interactions.

NOTES

1 "Bringing Big Data to Enterprise" (IBM), http://bit.ly/big-data-enterprise.

2 Legendre, Chelsea. "Year 2025—An Age of Machine Learning and Data On-Demand" (U.S. Department of Defense, April 27, 2016), http://bit.ly/year-2025-data.

3 Foley, John, "Extreme Big Data: Beyond Zettabytes and Yottabytes," *Forbes* (October 9, 2013), http://bit.ly/beyond-zettabytes.

4 Allen, Lisa. "What Is Big Data?" *Forbes* (August 15, 2013), http://bit.ly/big-data-trends.

5 Buluswar, Murli. "How Companies Are Using Big Data and Analytics" (McKinsey&Company, April 2016), http://bit.ly/big-data-analytics-insights.

6 "Trends in Machine Translation" (Common Sense Advisory Research, October 2011), 5.

7 Papineni, Kishore, et al. "BLEU: A Method for Automatic Evaluation of Machine Translation," in the *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (July 2002), 311–318. BLEU, or bilingual evaluation understudy, is an algorithm for evaluating the quality of text that has been machine-translated from one natural language to another.

8 "Content Strategy for the Global Enterprise" (Common Sense Advisory Research, April 2011), 11–14.

9 "Global Website Assessment Index 2015" (Common Sense Advisory Research, July 2015), 2–8.

10 "MT's Journey to the Enterprise" (Common Sense Advisory Research, May 2016).

11 "Translation Future Shock" (Common Sense Advisory Research, April 2012), 16–18.

12 "Can't Read, Won't Buy" (Common Sense Advisory Research, February 2011), 46–47.

13 "The Language Services Market: 2015" (Common Sense Advisory Research, June 2015), 2–5.

14 This is a conservative estimate. The Unicode Consortium's UTF-8 character encoding representation, which accounts for 87% of all non-binary data on the Internet, requires one to four bytes per character. However, European languages Roman script uses mostly one-byte characters. For more details, see pages 12-14 of "Translation and Localization Pricing" (Common Sense Advisory Research, July 2010) and https://en.wikipedia.org/wiki/UTF-8#Description.

15 "Post-Editing Goes Mainstream" (Common Sense Advisory Research, June 2012), 6.

16 "Translation Future Shock" (Common Sense Advisory Research, April 2012), 12.

17 FREME Open Framework of E-services for Multilingual and Semantic Enrichment of Digital Content, www.freme-project.eu.

18 "How Google Translate Will Increase Demand for Human Translation" (Common Sense Advisory Research, March 2010).

Don DePalma has more than 30 years of experience in technology, language services, and market research. He is the founder of Common Sense Advisory (CSA), a market research and consulting firm serving the language services sector. He initiated CSA's coverage of localization maturity, enterprise language processing, business-driven globalization, practical machine translation, return on investment for localization, and multicultural domestic marketing.

Prior to founding CSA, he co-founded Interbase Software, served as vice-president of corporate strategy at Idiom Technologies, and was one of the first analysts at Forrester Research, where he consulted to senior management at Global 2000 companies. He lectures, writes, and is frequently quoted on topics concerning online marketing, content management, multicultural marketing, localization, return on investment, and website globalization. His book, *Business without Borders*, is widely used in university and business training courses. Contact: don@commonsenseadvisory.com.