And Making Localization Tech Work for You
Localization experts featured weekly.

The podcast that teaches you everything you need to become an expert in language translation management.

Learn more at smartling.com/locshow

SMARTLING

Subscribe Today
“When the gift of one, becomes the gift for all.”

The girl with the gift of language

a film by translated.
Take your localization program to the next level with Lilt’s translation services.

Lilt delivers quality human translations, without compromise.

To learn more, go to Lilt.com
I think we can all agree that 2020 was a wild ride, and we all have high hopes for 2021. We’re doing our part by re-launching our print edition of the magazine – if you’ve missed a fully tangible, bendable MultiLingual, this is already a return to normalcy.

As far as the content goes, we’re taking our cue from what we hope for the future: the best of what has been, updated and re-imagined for the realities to come. This issue focuses on tech, and specifically, how to make it work for you. Because what good is the future if it’s not friendly?

Throughout 2021, MultiLingual is committed to nurturing our traditions of goodwill and neutral reporting, and trying out new things. As always, we welcome feedback and engagement.

Happy new year, and to all our readers, writers, sponsors, friends, and colleagues, may it be the best one yet.

Katie Botkin
Focus: AI and Making Localization Tech Work for You

Review

16
Congree Authoring Server
Nicole Keller

Column

24
State of the Machine Translation Union
John Tinsley

Column

26
Regulations Were Made to be Translated
Christophe Djaouani

Column

27
On GPT-3
Carol Jin

Focus

30
AI for the Language Industry
Carol Jin

Focus

34
The Future of the Localization Industry
Kajetan Malinowski and Jaime Punishill
Contents

Focus

40
AI for Language Technology
Donald A. DePalma

Focus

44
Where Tech Meets Human-Centric Clients
Olivera Rosulnik

Focus

46
The Ever-Evolving Technology of MT
Arturs Vasilevskis

Focus

48
How JIRA Can Help Localization Teams
Donato Giuliano

Standards

53
Localization Standards Reader 5.0
David Filip

Nonprofits

68
Translation Fights Back
Eric Paquin

Up Front

5
Post Editing

News

8
Featured Reader

9
News Stories

15
Calendar

72
Buyer’s Guide

January/February 2021

Multilingual
Would you introduce yourself?  
Olga Deputatova. I’m a localization manager at Tor-tuga Social Ltd., a mobile game developer based in Russia.

Where do you live?  
I live in Penza, a small town located not far from Moscow – 700 km (about 400 miles) is really not far compared with Vladivostok, for example, which is 9,000 km (5,600 miles) away from the capital.

How did you get started in this industry?  
I was lucky enough to be born in a multicultural family. Most of my “tribe” is fond of languages. My grandpa was an interpreter from German, English, and French into Russian. His brother spoke Hindi and Farsi. One of my great-grandmas was Polish, while another one was Armenian. I also have a Belorussian great-grandpa, to complete the picture. I grew up among antiquarian books and newspapers in foreign languages, so this passion for linguistics was inevitable, as I see now. I graduated in 2004, but, being an eager lifelong learner, I can never be satisfied with what has been achieved, that’s why I keep following my learning path. Courses and webinars at Washington University, Moscow State University, TAUS, and so on are an integral part of my life.

What language(s) do you speak?  
French, English, Spanish, and Russian (native). Now I’m learning Russian sign language, but it’s not such a fast process.

Whose industry social feeds (twitter, blog, LinkedIn, Facebook) do you follow?  
Social media is an amazing way to keep up to date, so I follow a lot of industry leaders and companies all over the world: MultiLingual, Nimdzi Insights, Anne-Marie Colliander Lind, Miguel Sepulveda, Danilo Monaco, Diego Cresceri, Carlos la Orden Tovar, and many others.

What do you like to do in your spare time?  
I’m crazy about mountain skiing, traveling, reading, and collecting antiquarian books.

What industry organizations and activities do/did you participate in?  
To begin with, I’m a member of Women in Localization and the Union of Translators of Russia. I’m a passionate conference-goer – conferences facilitate cooperation and communication between all participants of our industry and helps me keep up with the latest trends. And we have a lot of fun there! 2020 was especially eventful: I attended 12 inspiring conferences (Elia Together, TranSib Forum, and many game localization gatherings like DevGamm and Hamburg Mobile Summit). I was thrilled to bits to be invited as a speaker to five amazing events (such as LocWorld, BP Translation conference, and KTLC). Every year I feel really happy and honored to be among the jurors of Littera Scripta, an international youth translation contest.

Do you have any social feeds of your own? Twitter handle, blog?  
I don’t have much time to be as active on social media as I’d like to be. However, I share useful information on LinkedIn and Facebook. When I speak at conferences, I share my infographics, tips and helpful links after every presentation.

Why do you read MultiLingual?  
To tell the truth, I read all issues as soon as I receive them. This past year, MultiLingual has been especially important – I should say vital for our industry – when we all were separated by the lockdown. Your magazine is like a tie uniting translators from all over the world. It encourages and inspires us; it makes us feel that we’re still alive and we are not alone with our problems. When I opened the November/December 2020 issue, I couldn’t believe I saw our project on the pages of my favorite magazine! I’ll keep it for my great grandchildren! 😊
Industry News

Planet Word: a Museum Where Language Comes to Life

Where Language Comes to Life is the motto at Planet Word, a museum in Washington DC that opened in October. Nominated for USA Today’s Best New Attraction, the museum brings together language, technology, and ingenuity to create a fun experience for people of any age or literacy level. It features a host of interactive exhibits, many with 3-D and voice recognition, as well as an upcoming word game mystery adventure village.

We spoke with Planet Word Founder and CEO Ann B. Friedman to learn more about Planet Word.

Can you share a little about yourself and your team? How did Planet Word come about?

Late in life, I became a teacher and for nine years taught beginning reading and writing to first graders. With kids in first grade, you really have to teach them everything, to read, to write, to spell, exposing them to poetry. It is very rewarding to see their eyes light up about poetry and word play. I found it more rewarding than I ever expected.

Then I retired in 2011, and I had this background in literacy and in education, from both my teaching experience and my experience with the SEED Foundation, which runs the nation’s only inner-city public college-prep boarding schools. As chair of the SEED Foundation for six years, I became intimately familiar with the problems of inner-city education and the different techniques for teaching literacy.

So I looked around to figure out what I would do with my background, hoping to discover something meaningful and flexible. I came across a story about MoMath in New York City, a new museum of mathematics that used technology to bring abstract concepts to life in a fun way. I started researching museums, and the phrase informal education kept popping up. I realized that is what I wanted to try, but for words and language. I wanted to find a way to re-engage people of any age with words and language because all the trends in society — in America, in particular — were going the wrong way. I found the Lord Cultural Resources, which is one of the leading museum consulting firms, and I cold called them.

They assigned me to their team in New York City and really took my idea seriously. They conducted a 90-page market analysis and reached out to professors and visitor services people in DC, and the feedback coming back was all very positive. Once we had that base covered, I still was concerned about the idea for exhibits.

We decided to conduct four focus groups, two with 10-12-year-olds and two with their parents. We asked them what would you think about a place where X happens? We did not call it a museum; we just put the idea out there. Many of the participants reacted with excitement, saying, “Well, you’re talking about a word museum.” The facilitator for these focus groups even said that in her 25 years running focus groups, she had never seen such a positive response to an idea. At that point, I knew we had something special, and I had to keep pushing the idea forward.

By January 2018, I had the lease on the Franklin School — which is our national historic landmark home — from the Washington DC government.

We started construction in April 2018, which consisted mostly of hazardous material abatement. The building was 101 years old and had been abandoned and neglected for ten years, so it was full of lead paint, mold, and bird droppings, and it was even raining inside the building when I went in one day.

We cleaned everything up, and I brought in a contractor, architects, and a fabrication company, and we hired Local Projects — an experienced design firm in New York City — to plan and design the exhibits.

It sounds like you had your work cut out for you.

Yes, it was quite a challenge. I had built our house and done renovations, and I had come from a family with a real-estate background, so I went in thinking it would be no sweat. However, restoring a national historic landmark and dealing with all the oversight bodies and not being able to touch the exterior and not being able to do everything you want to do on the interior because you have to protect the historic features was another layer of complication. It would have been hard enough to build this museum because the exhibits are extraordinarily complex.

In 1880 Alexander Graham Bell sent a message using what he called a photo phone, which was a message using only light from the building. So the Franklin School — our home — was a...
national historic landmark not only for its innovative architecture, but also because Bell conducted a successful experiment that was basically the birth of wireless. The fact that our exhibits combine education and literacy with technology is such a perfect match for the historic origins of the building.

Can you tell me about some notable exhibits at Planet Word?

Our tag line is Where language comes to life. We really achieved that in our library. It is a high-ceilinged space lined with bookshelves, and there is a mirrored ceiling, which magnifies the height of the space. It is quite magical when you come in. If you take one of the books and place it in a special cradle on our central-story table, it comes to life in front of you: a projection begins, and there is a script we wrote for each of our selected books.

Our books span everything from picture books to books for adults. We always have a wide diversity of content so that there is something for everyone. What then seems to happen, though, is that everyone goes from book-to-book to see what unique qualities each book brings to the exhibit.

We also have our iconic gallery, Where Do Words Come From? which is an etymology exhibit that uses projection technology and sound effects to tell the story of English and how words were incorporated into the English language. We built a 40’ x 22’ word wall with over 1,000 three-dimensional words stacked on top of each other.

Using voice recognition technology, we made it possible for a narrator not just to tell this story, but also interact with viewers. We have some different categories; if the narrator says a word like portmanteau, then the different portmanteau words that we selected from the word wall will leap out. When the visitor chooses a word through voice recognition, the narrator then responds with the story of that particular word, and the software provides a multimedia experience with imagery and animation. It is a lot of fun. No one has ever seen anything like this.

Does the exhibit go into the ancestry of Old English and the transformations once French influence came into the language?

It does go into those histories, but only at an introductory level to get the main point across that English — and language itself — is always changing. Besides bringing language to life, the museum is also trying to help people understand and appreciate that language is alive, and that we should not be afraid of that, but celebrate the vitality of language. It is more of a descriptive museum, rather than prescriptive. We want everyone to be welcome no matter how they talk. We do not say what is wrong and right or try to push grammar conventions. We are simply in awe of what people can do with words and how flexible and enjoyable our language can be.

For the word wall, I knew we needed to map out where words come from, so I sent a list to the exhibit designers that went through different possible origins. It could be anatomical: what sounds can our throats and teeth and tongue make? Or it could be words arriving through war and conflict. We also borrow words from other cultures and languages. We invent words.

We came up with 24 different ways that words have come into the English language, but then whittled that down to eight main origins, from the ancestral languages to the invented or portmanteau words. That is the story we tell. We are really trying to drive home the idea that language is constantly changing and that there are certain key derivations with the words that we use.

One of the missions of Planet Word is to promote linguistic diversity. How do you do this?

Our largest gallery is called The Spoken World, and we have 28 language ambassadors and two signers, who teach visitors many lessons about their respective languages. We have chosen the most unique characteristics of each language that would be fun for people to learn. The language ambassadors appear on iPads and interact with you through voice recognition.

The dominant feature of the gallery is a 12-foot diameter globe hanging from the 22-foot ceiling and covered with 4,800 LEDs. At the end of the mini lesson, the iPad will say “look up” and the LEDs will display something related to the lesson you were just learning. For the Hebrew lesson, for example, the last part of the lesson is on the chet, glottal sound. The narrator encourages you to say L’chaim, and when you do, the globe above you displays two champagne glasses clinking together.

Not only are people learning about these languages, but they are also getting the chance to try speaking them. We have a wide variety of languages, like the more common Spanish, Russian, and French, but also click languages, Indigenous languages, and some endangered languages. We also tried to incorporate languages from the immigrant communities most represented in DC, like Korean and Amharic.

Our goal is to intrigue you enough to want to pursue these languages further once you leave the museum.

Obviously, it is impossible to know what is going to happen a week from now in our present moment,
let alone a year or more, but what are you looking forward to in 2021?

We technically opened in October 2020, but we are not done. We still have parts of the museum coming online, hopefully by late spring 2021. We also have a restaurant called Immigrant Food coming in to work with us, which has a fusion menu with cuisine from different immigrant groups.

We obviously cannot use our auditorium, and all the programming, classes, and field trips that would have kept the museum buzzing until we are through the pandemic. Right now, we are planning a lot of virtual events. People are hungry for programming on the language arts and humanities, which very few museums in Washington offer, so we are filling a void.

If our readers wanted to follow Planet Word, how could they find you?

We are on all social media platforms: Twitter, Facebook, Instagram, LinkedIn, and you can access our website at planetwordmuseum.org. You can also sign up for our newsletter or our membership program to receive benefits. Keeping up with us is very easy.

If one looks closely enough, a city reveals in every brick, window, and street sign a complex history and continuing tensions through its translational spaces.

We spoke with Sherry Simon, a professor of translation studies and the author of several books on translation, including Translation Sites: a Field Guide, Cities in Translation: Intersections of Language and Memory, Translating Montreal, and Gender in Translation.

Can you tell us about yourself?

I teach in the French Department at Concordia University, an English language university, and I live in Montreal. That is a big part of it. Montreal is not only the place where I live, but the place that I began to write about and that shaped my ideas about language. That is to say, it is a city that had two strong languages, French and English, that were very much in tension with each other. As the years changed, the nature of these tensions changed, but they were very strong when I was growing up, so that really shaped me.

I started translating at a time when you simply referenced texts to help you figure out the best way to translate. You had a few prescriptive books telling you what were the best equivalents between French and English, and why it was tricky to translate between those two languages. Many words were the same. We have what we call false friends, or cognates, because historically the two languages overlapped many times. That is why we have so many words that look the same, but their histories and their current usage are very different. There were not many guides for understanding what translation was about.

Over the time that I began teaching until now, that situation has changed radically. There is a huge...
library on translation now, in an area called Translation Studies. It is no longer linguistics, or comparative literature, or Belle-lettres, as they say in French. It is a discipline unto itself. What has become clear to me is that translation is an absolutely fascinating window onto the world.

You can learn so many things by looking at those questions translationally. You can look at poetry. Poetry is in the translation; it is not lost in the translation. It is in the process of translation, and that is what makes it so interesting. You can learn a lot about history.

Looking at histories like those of Eastern Europe, you see cities that have changed language regimes over the decades, especially around the periods of the first world war and the second world war, and when communism came along. An entire city could essentially be re-wallpapered with a new language. That is a process of translation because the city is translated from one language to another. You can come to understand memory — memory of history, memory of traumatic circumstances — as processes enacted through language.

In your book Translation Sites, you explore the question of how translation and memory intersect in the life of the city and places within the city. Can you talk more about that? What is a translation site? How might one discover a translation site?

One of the elements that is so fascinating about some of the translation sites I explore is that their histories become invisible over time. The term ghost signs has been used in Eastern Europe. A crack in the surface of a building, a façade, peeling paint, a crumbling window will suddenly reveal another language that has been painted over. You must be attentive to small signs that are going to lead you to the history. They are often not necessarily official signs, or if they are official signs, their message might be different now.

A city rich with translation sites would be somewhere like Lviv, what was at one time Lemberg, in what is today Ukraine — you always have to say what is today because the city had four or five different names over time, and four or five different languages. There is a place there called The Space of Synagogues, where there were several synagogues — very significant synagogues — destroyed during WWII.

But there is an interesting contrast between the time when it was just an abandoned space and now. It still had a strong meaning. Sometimes these spaces are places that people do not dare touch because they know it has such a history, but they do not do anything to it either. They just leave it abandoned. Sometimes abandoned spaces are the spaces that will show a translational history.

There once was an opera house in Prague that for around a half century was a German opera house that played the most modern, avant-garde repertoire: Wagner, Müller, Schoenberg. But with the nationalization of Czecho-Slovakia, it became a Czech space, as had happened during those times when nationalism and nationalization was carrying small countries like Czecho-Slovakia forward.

After the second world war, any reference to German was unsavory, so all the references to that past were eliminated, thus forgetting the tremendous cultural heritage of the German language in Eastern Europe. German was the language of culture in eastern Europe for centuries. It was not a bad language. It became a bad language — a language of the Nazis, the murderers, the assassins — but the language had a history over and beyond that.

So only now as the years have passed can we begin to appreciate that. But if you are walking in Prague, you will likely not know that it was a German-speaking opera house. You must have the local knowledge.

In your book Gender in Translation, you discuss several notable women translators. Are there any that you can highlight for our readers?

The women translators in whom I was particularly interested were from the 1920s. There was a woman called Willa Muir, who with her husband was the first translator of Kafka. There was also Constance Garnett, about whom many people know because she is criticized as a bad translator of the Russian classics.

But she translated of her time, and there is a movement now to defend what she did because she had this job to do and she did it in record time. Not only that, but she was also a sympathizer, an anarchist, and she was very interested in the Russian modernists and the political ideas of late 19th century Russians. She did a
lot to promote those ideas, not just translate.

The same was true of Willa Muir. She was a very active thinker. So what I love about these women – three or four of them I talk about in my book – is that they translated because they loved the works they were translating, and because they were doing it as part of their political projects, as thinkers, as women who had independent takes on the ideas of their time.

Especially Constance Garnett. She is one of the few names that people know as a translator, and they know her name because they criticize her, naturally. But this is unfair. Certainly one could improve upon her translations, but it is unfair not to take in the historical circumstances during which she was translating, and how quickly she was obliged to do it, and with how little help.

You have discussed the use of the term multilingual as insufficient in describing cities.

The way I see the word multilingual used, like multicultural, is often in a descriptive way to talk about how many. When you talk about New York as multilingual, often that description implies how many languages. Of course this description is appropriate – New York has over 200 languages. But what does it tell you about what you hear on city streets? What does it tell you about how people interact with one another? Where are the spaces where people speak these languages? How important are these languages? How often are they spoken? How often are they translated into English? How often is English translated into them?

If you are a migrant, you will translate yourself constantly into the major language. That is what you must do in order to survive as an immigrant, to integrate into your new society. You translate yourself into. You are not going to find the major culture translating itself into you. That is a very strong dissonance, the weights of those directions.

And that applies all across thinking about translation: the cultural weight of those directions, whether you are translating up or down, out or in.

Can you talk about the concept of counter-translation?

The way I use it is in the context of these cities that have been translated over, as in languages that have been effectively suppressed and eliminated and murdered. If you talk about Yiddish in Eastern Europe: Yiddish no longer exists in eastern Europe because the people who spoke that language were murdered. So that is a suffocation or elimination of language.

But what happens in those cases is that after a period of time, or in a period of post-trauma, or post-violence, something happens where those languages can acquire a new voice.

One example involves the First Languages in Canada. I give an example of the National Gallery in Canada, which is the flagship symbolic art institution of the country. A couple of years ago, it re-did its entire collection of art to include First Nations art as the founding art of the nation – not as some ethnological art that you have to look for in another museum, but an art that was fully integrated into the history of the country.

What the museum did to consolidate this was to use First Nations’ languages as part of the labeling of the art. You usually see labels communicate information about the art as a neutral conveyor of information. But in this case, the labels themselves were written in First Nations’ languages. To me, this was another example of counter-translation: here were languages that were revived through translation. Altogether, there is a whole movement of re-translation into First Nations languages that had been suppressed for centuries. These kinds of reversals, I find satisfying because they are desirable reversals. That is what I mean by counter-translation.

The newly-minted Entertainment Globalization Association (EGA) has set its sights on publicizing the role of localization in the entertainment industry.

The goal is to create “better awareness that this industry even exists,” with people who are “creating the original IP,” said Chris Fetner, who is heading up the organization as managing director. For example, writers and directors may know that translation and globalization happen; that movies get dubbed in some markets, but for the most part, said Fetner, the effect is like a duck seeming to glide effortlessly over the surface of the water. Nobody sees the legs kicking furiously under the surface to propel the duck forward. Fetner said he thinks directors would likely be shocked if they knew how little time voice actors in dubbing studios were given with scripts, for instance.

“For a long time, [localization has] been treated like a utility, and it’s done it a disservice as an art form,” said Fetner. EGA’s goal is to ensure faithful representation in whatever language a film is translated into — and this might mean a little more time than the localization providers are currently given.
When Fetner asks people what they’d like to see from EGA, “what they ask for more than anything is having another week, having a better opportunity to provide good quality,” said Fetner. “It really is an industry of people who love what they do,” and above all, they want to create good output.

EGA has ten founding members: Audiomaster Candiani, Deluxe, Hiventy, Iyuno Media Group, Plint, SDI Media, Visual Data Media Services, VSI, ZOO Digital, and Keywords Studios. With the exception of video games specialist Keywords Studios, all the founding companies participate in entertainment localization.

Fetner himself has worked in the entertainment industry for 25 years on the client side. His foray into localization began at Discovery, when he was involved in localizing content from US English into UK English. Then, at BBC Studios, he worked on projects localized into Latin American Spanish. But it wasn’t until he joined Netflix and headed up their localization vendor strategy that he understood the full breadth of the industry, he said.

When Fetner left Netflix this fall, he reached out to the vendors he’d worked with, and they expressed regret they would no longer be working together. Some asked if he’d do consulting. Fetner said he told them, “if you feel like there’s work to be, let’s all work together with everybody,” and from these conversations, the EGA was conceived. It was “spearheaded by the founding companies,” who were asking “what would it look like if everyone worked together in the industry,” said Fetner. Additionally, with Fetner available, they had a managing director who was willing to work on the issues.

Fetner has plans to reach out to writer’s and director’s guilds, as well as similar organizations, in an effort to “add value” by inviting them to work more directly with localization.

For more news, updated daily, visit www.multilingual.com/news
January
Translation, Transcreation, Transadaptation and the Science Behind Them
January 21, 2021
Online
IMUG, www.meetup.com/IMUG-Silicon-Valley/events/270387619

LocWorldWide 43
January 27-28, 2021
Online

February
MultiLingual Winter Series
February 4, 11, 18, 25, 2021
Online
MultiLingual, https://multilingual.com/season-series

LocFromHome
February 10, 2021
Online
Smartcat, https://smtc.at/locfromhome3

Project Underwear: Analyzing Global Consumer Expectations of Localized Content
February 18, 2021
Online
IMUG, www.meetup.com/IMUG-Silicon-Valley/events/270387677

Together At Home
February 24-26, 2021
Online
Elia, https://elia-association.org/event/together-2021

March
GALA Connected 2021
March 23-25, 2021
Online
Globalization & Localization Association, https://www.gala-global.org/events/conferences

May
Elia’s Focus on Executives
May 6-7, 2021
Rhodes, Greece
Elia, https://elia-association.org/our-events/

July
UTICamp 2021
July 19-25, 2021
Dnipro, Dnipropetrovsk Oblast, Ukraine
Ukrainian Translation Industry Camp, https://utic.eu

All events are subject to change.
Congree Language Technology GmbH is a software manufacturer based in the south of Germany. They specialize in the field of computer linguistics, and support companies of any size in the process of content optimization. They mainly concentrate on the work of technical writers and corporate authors who generate source content. Since they integrate the Congree Linguistic Engine, which performs a morphological analysis, their Language Check is only available for a small selection of languages (English, German, French, and Spanish). For Italian and Japanese, Congree offers a basic Language Check that is not as sophisticated as for the main languages. The other components are language-independent and can be used for all languages.

In the newest version, Congree announced a comprehensive revision of the administration module and many new features for the English language.

Particularly in times in which the use of machine translation (MT) is increasing, the computer-based verification of text quality is becoming increasingly important. In this context, I want to focus on the quality of the source text, even though Congree can also be used for translated texts as long as the language is supported. However, one point is clear: the less ambiguous the source text, the better the translation (MT output or human translation).

Even without a great demand for translations, corporate language in general plays an important role for companies to individually describe their unique products and create a consistent brand image.
The heart of the system is the Congree Authoring Server that is customizable and uses three different resources to optimize a company’s content:

1) The Rule-based Congree Language Check ensures the correct use of language and style rules. This does not only cover correct spelling, grammar and terminology but also the adherence to predefined Style Guides.

2) The terminology component uses the information of either an existing, third-party terminology system (such as SDL MultiTerm) or an internal terminology list to check for consistent and correct term use.

3) The Authoring Memory is a repository for sentences. It can be compared to the functionality of translation memories in CAT tools. Congree can detect text segments that are similar to new sentences that were written and approved in the past. Through more consistent and identical sentences there is a big potential to lower costs in the subsequent translation process because more 100% matches are produced.

On the practical side, Congree offers plugins or add-ons for the integration of existing authoring environments (DTP, CMS, or word processing software such as Microsoft Word). The main target tools are used in technical documentation, marketing, and the translation process. But Congree also offers a neutral web interface, so anyone who wants to check a text, independent of an editor, can use the Congree technology.

Figure 1 shows three Congree panels, one for each resource, grouped around Microsoft Word. The English sample text for this article is a Word document.

**Rule-based Congree Language Check**

The rule-based Congree Language Check is the most extensive and diverse check in Congree and covers five different categories: spelling, style, abbreviation, grammar, and terminology. An overview of the number of potential errors and their acceptability (green = safe, yellow = acceptable and red = unsafe) is displayed at the top of the Congree panel (see Figure 2).

Figure 2: Notification categories of the Congree Language Check.
In general, the detected error can be either corrected, ignored, or even ignored for all occurrences of the potential error at once. If a rule leads to too many false positives, the user can also deactivate a rule manually while working on a specific text.

**Spelling:** Congree does not rely on a third-party spellchecker, like Hunspell or the Microsoft spellchecker, but implemented its own spellchecker that offers more comprehensive functions. First, it not only detects words from a predefined list, but also inflections, that means that nouns need only be listed in their singular form. Second, the underlying terminology database can be populated with company-specific terms like product names or the terminology of a certain subject area. And third, all terms stored in the terminology database are ignored, even though they are not in the internal spellchecker list.

The spellchecker also looks for the correct usage of “a” and “an” before vowels and consonants or the correct British or American spelling (analyse vs. analyze or instalment vs. installment). But the system cannot distinguish between typical British (e.g. football) and American (e.g. soccer) words unless they are stored in the terminology database. It does actually make more sense to store them in the terminology database because in examples like flat vs apartment, Congree would always suggest apartment, even if flat is used as an adjective.

For example, if the word “possible” was used, the system could state that the usage of “possible” is too colloquial for many texts, including technical documentation.

**Grammar:** There are a lot of underlying rules for the integrated grammar check available, so I just want to mention some examples. There are more general rules, like “Infinitives take no s at the end” but also more specific ones like “Don’t combine too with a comparative.” Most grammar rules are based on the correct identification of part of speech. In the sample text, the noun phrase “safety informations” was detected because mass nouns (information) do not have a plural form (see Figure 3).

**Abbreviation:** This is a separate check for the correct usage of abbreviations. The system uses an integrated list of known abbreviations, which could include company-specific abbreviations as well. If an unknown abbreviation appears, the user is notified about the unknown abbreviation. This means that it is best to enter a complete list of all company-specific abbreviations before using Congree. If an abbreviation is additionally stored in the terminology database and marked as “allowed,” Congree will suggest using the “preferred” term instead.

**Style:** This is the most comprehensive category in the Congree Language Check. It offers a huge selection of predefined style rules that can be used to create individual style sets. The German association for technical communication (tekom) for example published two guidebooks for rule-based writing for technical authors in English and German containing predefined guidelines also covering a huge variety of rules for translation-oriented authoring.

Additionally, there are Style Guides covering the guidelines for Simplified Technical English (STE). An example rule for STE: Write instructions in the imperative form. A general example: Avoid the use of “should,” “could,” “would.”

On a term basis, Congree can also check for basic style criteria, including language registers (formal, informal, colloquial). This information is stored together with different synonyms.

But if there are any company-specific rules, they can also be added to the existing sets. These rules might come from existing terminology guidelines which often define specific usage of terminology but also the maximal length of sentences or whether active or passive voice should be used in the documentation. Especially machine translation systems benefit from these more standardized texts and generally produce better output.

**Terminology:** Terminology errors can only be detected when a terminology database is integrated or an internal terminology list is available. Congree only offers an integrated terminology component with basic terminology functionalities.

For the practically relevant detection of the incorrect use of rejected synonyms, the system needs the “usage information,” which means that terms and synonyms are marked as “preferred,” “allowed,” or “deprecated.” Only with this information can Congree detect incorrect usage — even if vari-
Ants or different writings (e.g. with and without hyphens) are used. If the terminology database contains contradictory information, Congree suggests the user checks this specific term in context displaying all possible matches.

For example, imagine there are two entries containing the term “improvement.” One is marked as “allowed” and one as “deprecated.” Via direct access to the terminology database, the technical author can read the additional information for both entries and decide which one applies to the actual context.

**Authoring memory**

Another integrated resource in Congree is the authoring memory database. It is a monolingual sentence-based database storing all sentences that were previously used in other documents or created in the same text. During text creation, Congree checks new sentences against existing ones in the Authoring Memory, and suggests similar or identical sentences. The similar matches range from 70-99% similarity and are comparable to fuzzy matches in the translation management environment. The aim of this feature is not to write the same idea or instruction in different words. This has two main advantages. First, the statements are clear and cannot be misinterpreted because of different ways of expressing the content. Second, with regard to the translation process, identical sentences save money since they need not be translated twice.

The recognition process works very well. With the means of a tracked changes mode, the differences between the sentence in the text and the stored sentence in the Authoring Memory are highlighted in a separate window.

Besides the general text information, Congree can also use formatting information including headline, paragraph, list elements, and so on if a sentence is used several times but with different formatting settings.

It should be positively noted that there is also in interaction between the Authoring Memory and the Congree Language Check. If there is a 100% match, minor errors detected in the Language Check will be ignored for this specific sentence. But on the other hand, there is also the option not to save sentences with major errors in the Authoring Memory.

**Terminology component**

The third main resource is the terminology database. Either the integrated terminology component or the connection to a third-party terminology database (e.g. SDL MultiTerm) are used as the basis for the terminology check described above. Beside the terminology check, the additional information of the terminological entry is displayed in a separate Congree panel. Here, the users can see and individually select which information they want to see: concept information, synonyms, translations and/or term information. By clicking on the link next to the term, the web interface of the connected terminology database opens in a browser window (e.g. SDL MultiTerm Online) and the users can see the complete terminological entry in the original terminology database. But this is not absolutely necessary since all the information can be displayed within the Congree panel (Figure 4).

In addition to the terms already stored in the database, Congree also suggests new term candidates from the actual text that could be optionally stored as new entries. Since the system is based on a comprehensive linguistic system, the suggested term candidates are more reliable than term candidates from traditional term extraction tools. In particular, multi-term words are recognized in a more accurate way with linguistic information than with a statistical approach. This does not mean that all the term candidates are actually real terms that would be added to a terminology database, but the noise (too many useless term candidates) and silence (term candidates that are not recognized) aspect of terminology extraction tools are reduced. In one sample text, Congree suggested “work step,” which is definitely a good term candidate but also “normal review” or “individual passage,” which are not real collocations but more loose word combinations that would usually not be added to a terminology database.

**Administration**

The Congree Control Center is the heart of the system. All the individual configurations for the different resources can be made here. In my opinion, the most interesting and important section is the “rules” section. Here, users can define all their specific configurations for the Congree Language Check including spelling, grammar, terminology, style, and abbreviations. Some sample rules have already been mentioned. A company, however, can define as many sets as many as needed, such as Best Practices for Technical Documentation and Simplified Technical English. Depending on your editor, working language, or text type, you can select the appropriate rule configuration. Furthermore, you can define terminology rules for the creation of multi-word terms, add company-specific terms
to your user lexicon, define general synonyms, and create notifications and explanations for error types. The notifications do not only contain the category (such as style or grammar) and unique code, but also a keyword that explains what to check, e.g. “Review the word order,” an instruction to help correct the mistake, e.g. “This adverb is in a marked position. Check if it is better to put the adverb behind the auxiliary verb,” and an explanation why, e.g. “Adverbs do not precede the finite auxiliary verb unless they are heavily stressed.” Moreover, you can always add an example that demonstrates the mistake:

WRONG: They often are companies with a high turnover.

CORRECT: They are often companies with a high turnover.

All the text fields mentioned above can be individually created, so that company-specific examples are used that might be clearer than general examples.

In the end, all these individual configurations can be combined in a Style Guide that forms the basis for a document-specific check.

Reporting

In the Control Center the Reporting Operator can decide whether and which data are collected and how long they are stored. In this area you can also define the “release level,” which is Congree’s quality metric and indicates whether the number of errors in a specific category is safe, acceptable, or unsafe.

If the data collection is activated, the results of each check, independent of the editor, is stored anonymously and can be reviewed by the administrator. Beside general information, like how many grammar mistakes were detected in a document, there is also more detailed information available, including concrete action each author performed (e.g. corrected the error or ignored the rule).

Finally, Congree creates diagrams with statistical information, including the number of notifications for each check, or how often a rule was disregarded, which could be a hint that the underlying rules might need to be reviewed. The statistics can be restricted to certain period of times, languages, user groups, and so on (Figure 5).

Conclusion

Congree offers a lot of possibilities to check content in different environments. Depending on the preferences of the authors, these checks can either be made in real-time or in batch mode at the end. Certainly, they can also be performed for already-existing documents.

If you just want to check a text independent of a specific editor, you can also use a web interface and just copy and paste the text. The same Style Guides are available there as in the plugin versions.

Since Congree integrates into the existing authoring tool, there is no need to actually “learn” a new application. The few features for the authors are easy to understand. But to get the most out of the system, it has to be thoroughly configured by the system administrator and that can be real challenging, because companies have to define which linguistic aspects are important for their texts.

In the end, Congree can definitely help to improve content quality and make it more consistent. As the results of the tested text were promising, it is certainly worth a look. 

![Figure 5: Diagram on the notifications per category from June to August.](image-url)
In 2017, Memsource established a dedicated AI team, and since then has released several features that have enhanced translation workflows. These include patented non-translatable recognition and most recently Memsource Translate, a dynamic machine translation management solution that selects the optimal engine based on the user’s language pair and domain.

Sometimes the best innovation is one that integrates so seamlessly with existing processes, that you barely notice it’s there.

In November 2019, we quietly launched our new Automatic Linguist Selection feature to help project managers make the most effective decisions when selecting translators for projects. The change was not visibly signaled in the UI, so some users might not have noticed any difference at first.

What does it do? Automatic Linguist Selection uses a machine learning algorithm to recommend the best possible translator for any given project based on their past experience.

The recommendation is powered by sophisticated domain detection. Whenever an organization uploads a document for translation, Memsource runs a keyword analysis to identify the nature of the content to establish its domain. Similar technology powers Memsource Translate, which sorts documents into one of 11 domains, such as Travel & Hospitality, Medical, and Legal & Finance documents.

The domain identification for Automatic Linguist Selection is far more detailed, recognizing over a thousand different domains, with each document being tagged with an appropriate combination of domains. This granular insight allows us to very carefully identify what makes each document unique, but also the many ways in which it relates to other documents processed through Memsource.

In addition to determining domains, Memsource keeps track of which linguists worked on which document. Over time, each linguist will develop a unique profile, with experience working with specific domains, which may be useful for future projects.

So, when a project manager creates a new project and uploads a document, our algorithm will analyze the domain of the content and create an ordered list of linguists with relevant experience. Using linguists with specific domain experience is always beneficial, as it ultimately leads to faster and better translations.

The AI recommends changes — but it doesn’t make the final call, the project managers do. However, our data shows that they rarely disagree with the algorithm. In approximately 70% of cases the project managers will choose a linguist from the list recommended by the AI.

Automatic Linguist Selection is already helping businesses translate faster and with better results, but it has the potential to develop further in several interesting directions.

For instance, Automatic Linguist Selection could be used to find a qualified translator from all of the users that work within Memsource, and not just with a given organization. This would be especially useful for businesses that need to quickly find qualified freelance translators.

Memsource’s AI technology is focused on ensuring that all users of the platform can work as effectively as possible.

You can learn more about how Memsource AI can help your business at: Memsource.com/AI.
Why do you work in localization? For translators, this answer may be more obvious. Translation itself is an often-underappreciated job; the general assumption is most people do it for love, not money. Translators themselves tend to be smart enough to have their pick of industries: a legal translator could have easily become a lawyer; marketing localizers are perfectly capable of writing source language blogs and slogans. According to a 2020 Common Sense Advisory report, 42% of global translators and interpreters came to the industry from another profession and more than half originally studied something else. Greek translator Christos Floros, for example, majored in political science, but says, “My love of languages made me choose this industry.”

So what about those jobs that have nothing — or very little — to do with language? Take the person who codes translation software: she could just as easily program enterprise sales platforms or dating apps. Or the person who does write those blogs and slogans, except for translation companies instead of for clients: why does she work in the language industry?

The fact is, there are a lot of people who keep localization moving without translating a single word. Who are they? Why do they do this?

Welcome to “Behind the Scenes,” MultiLingual’s latest column. Over the course of 2021, we’ll take a look at localization’s unsung heroes so to speak, but at its often-overlooked ones. Together, let’s shine a light on the non-language, language professions.

This issue gets us started with Bill Rivers, lobbyist for the Association of Language Companies (ALC), the largest trade association for language services companies in the United States. Rivers was appointed to the role last August after working as executive director of the Joint National Committee for Languages (JNCL), a US lobbying group representing translation and language education. For those outside the United States, that means Rivers “asks elected and appointed officials to take action on behalf of someone — in my case,” he explains, “on behalf of the language industry.” The first amendment of the US Constitution grants all Americans the right “to petition the Government for a redress of grievances.” While individuals often do this on their own, writing and calling their elected representatives, industry associations typically hire a pro. Rivers approaches local and federal government leaders with information about the language industry, advocating for its needs.

In 2020, for example, one of Rivers’ major projects was the continued fight against California Assembly Bill 5. Passed in 2019, this state law required “gig workers,” including translators and interpreters, to be classified as employees, not freelancers. According to US-based training group InterpretAmerica,
more than 5,000 of California’s linguists work freelance — and 75% worldwide per Common Sense Advisory’s 2020 report. At the time of interview, Rivers was advocating for a language industry exemption.

Lobbyists represent a myriad of industries, from manufacturing to real estate to agriculture. The job title itself actually comes from President Ulysses S Grant who needed something to call people waiting in the lobby, vying to get his attention. For Rivers, language came first, lobbying second: Early in his career, he worked as a Russian translator.

“I’ve only ever lobbied on behalf of the language enterprise,” Rivers says. “Coming from the industry, being multilingual and born overseas [in Bavaria, Germany] gives me a certain perspective on what we do.” What makes lobbying for this industry different from others, he contends, “is that everything we do is for the greater good.” Translation provides life-saving healthcare communication, life-altering social services information, and drives the international economy. Rivers says, “Everything we do is for the national interest. I firmly believe that, and our story is an easy sell on Capitol Hill,” the area of Washington, DC where senators and congresspeople work. “As Hans Fenstermacher said years ago,” Rivers adds, “we’re the biggest industry that nobody knows of. We’re part of every vertical in the US economy; you can’t design a product or service for global sales without us. We’re present in every community in the country, providing language access. But at the same time, there are outdated ideas about the industry that need to be constantly challenged,” and many of these ideas are prevalent among today’s US government leaders: “Not everyone is learning English globally and — of the 600 million or so who are — very few will learn it well enough to communicate effectively.”

Still, Rivers says people are often surprised to learn language industry lobbyists exist, “even in DC, where it seems that every imaginable group has a lobbyist.”

When translating books turns deadly...

To learn about Red T’s awareness-raising activities, visit red-t.org facebook.com/TheRedT twitter.com/TheRedT @TheRedT

Bill Rivers
Unpacking the Black Box

State of the Machine Translation Union

John Tinsley

John Tinsley is co-managing director of Iconic Translation Machines, which joined the RWS Group in 2020. He has worked with machine translation over the last 16 years in both academia and industry. In 2009, he received a PhD in Computing from Dublin City University, after which he co-founded Iconic. He now leads Iconic as a dedicated technology division of RWS Group.

It’s the start of 2021 — a time to look back and a time to look ahead. This time of year, we all love a good list. Whether it’s a retrospective on things that have happened in the previous 12 months, or predictions about what’s going to happen in the coming year, they are staple of articles far and wide, and this one will be no exception!

But rather than doing one or the other, we are going to do a little bit of both. We will take a look back at three key developments in the field of machine translation (MT) in 2020, and what we think we’ll see happen in 2021.

From a high level, MT research and development continued to advance at pace in 2020, if not quite at the pace of previous years after the initial flurry of activity around neural MT. In addition to the academic mainstays, research has continued to be prominent at big tech companies like Google, Facebook, and Microsoft, with Apple recently joining the fray. Language service providers have also been investing in machine translation technology to bring it in-house, with some notable M&A activity over the past 12 months.

What has quite noticeably changed based on discussions with buyers, service providers, and technology providers is the attitude toward MT and the rate of adoption. It’s more positive — and it’s growing fast.

The conversation around MT has evolved from questioning whether it’s good or not, to accepting that it is indeed good. In turn, that has led to more in-depth discussions on quantifying exactly how good it is, and where it still has room for improvement. This new line of questioning allows potential adopters to think more concretely about where MT can fit most effectively into their many and complex workflows.

This leads nicely into our list and the headline topic that was not only a big feature of the last year, but will continue to be prominent into the next: quality and evaluation.

1. Quality and evaluation

“Human parity” — two words that raise the ire of translators and developers! But claims of MT reaching so-called human parity led a consortium of researchers across Europe and the United States to assemble “A Set of Recommendations for Assessing Human-Machine Parity in Language Translation,” published in the Journal of Artificial Intelligence Research. This was a welcome development.

Despite the widely accepted fact that automatic metrics for MT evaluation are not fit for purpose, they are still widely used in the absence of an effective alternative. Unbabel is the latest party to attempt to address this challenge with the release of Comet, an open-source framework for MT evaluation. I’m curious to see if it catches on.

A big trend to watch out for this year and beyond is not just quality assessment, but quality prediction. That is to say, methods and tools that estimate the quality of MT output in real-time, without
having to compare it to an existing translation. Predicting quality is easy... but predicting quality reliably is easier said than done. This topic is being heavily researched in academia at present, with lots of prototypes floating around. We’ve yet to see this be fully productized, but it could be on the horizon sooner rather than later.

2. Doing more with less
MT research and development follows supply and demand trends when it comes to selecting which languages developers focus their efforts on. A by-product of this is that those language combinations left until later tend to have less data available for training models. The diplomatic term for these is low resource languages. Addressing this data shortage is a trend that has consistently gained popularity over the last few years.

Some popular approaches include data synthesis, which essentially involves the creation of MT training data using... MT (it’s not as crazy as it sounds). Another approach is known as unsupervised MT, which involves training MT models with very small amounts of data to get the process started, and iterating from there. This is quite a common approach in general machine learning, so it will be interesting to see how it evolves in the context of machine translation.

Last but not least, multilingual MT is a model that can actually translate between more than one combination of languages. Think of it like pivot translation — we can already translate between, say, Irish and Hungarian by translating from Irish to English first, and then from English to Hungarian. You can consider multilingual MT as a more elegant version of pivoting, whereby the relationships between all the languages are learned, rather than it being a two-pass process.

It’s an exciting approach because, in theory, we’re not limited by how many languages we can include. This approach is of interest to the likes of Google, which specializes in one-size-fits-all at scale, and they’ve published papers with some frequency on the topic.

3. Context is king
A familiar refrain from users of MT is to question why a certain word or phrase was translated correctly in one instance, but translated in a different way elsewhere. The answer is context. In MT, each sentence is translated completely independently of any other sentence that has either previously been translated, or is upcoming in the document. As such, we are potentially missing key information that would allow us to produce more consistent translations.

A potential resolution to this problem is context-aware MT — a process by which more information from the wider text is taken into account when translating specific pieces. The potential here is intriguing.

There are still some important questions to answer such as how much context to take into account (the more you take, the more complex things become), and whether or not it is always helpful. It might ultimately depend on the use case. Nevertheless, this line of research is active, and will grow in popularity over the coming months and years so keep an eye on it.

The bottom line: machine translation is very much in a “watch this space” moment, so make sure to keep an eye on this column to see what’s making waves! 😊
Welcome to this first edition of “Rules of the Trade.” I’m very excited to have the opportunity to share insights into the challenges facing professionals managing multilingual communications in highly regulated industries. In this series I will share with you what I’ve learned in my 21 years working alongside language services and tech customers in the financial, legal, IR, CRO, and pharmaceutical sectors.

But for this first edition I’d like to start by discussing the key ingredients to a successful partnership. It all begins with listening to customers. We hear loud and clear what’s keeping clients awake at night and work with them to deliver solutions. Firstly, by retaining them as customers, and then by focusing on developing our relationship with them, both across their internal departments and across our regions. Despite many outside factors — for example, the politics surrounding the US presidential election and COVID-19 — a strong partnership can weather the storm and adjust to the new conditions.

Good providers support customers by tracking regulatory changes, like we’re doing right now with the LIBOR (London Interbank Offered Rate) transition, which is perhaps the biggest regulatory update in financial services for a decade. LIBOR has been the benchmark for setting the cost of borrowing over the past 50 years, but it’ll soon be replaced by more up-to-date benchmark mechanisms. We’re working with law firms and financial institutions to get their multilingual content ready around that. We are environmental, social and governance reporting experts too, which is even more relevant since the advent of COVID-19. We’re also targeting business related to the EU’s medical device sector. Finally, we’re seeing IPO business more regularly again as resilient businesses show confidence in their offerings by floating stock on global stock markets.

Customers value expertise and efforts to align with them — efforts to understand their language, their terminology and their culture. They appreciate commitment to service excellence, quality, consistency, and responsiveness. Generally, they welcome suggestions for process improvement around technology and subject knowledge. Most of all, they are seeking to make the adoption of new regulations more routine — there is no end to regulatory momentum, especially for companies working in multiple markets internationally, so being able to absorb regulatory changes and updates as they happen enables a “business as usual” approach to compliance that can be a real source of competitive advantage.

Customers have made it clear this year that security, price and speed-to-market are priorities, due to the current environment which has needed an agile, creative and innovative response — all while maintaining the levels of service that customers expect.

2020 has been a year without precedent, and it has brought unique challenges for buyers of language services and technology in all the regulated industries. In the next column, I’ll be talking about communications for the clinical research organization sector. This is a seriously hot topic right now as we’re on the brink of a COVID-19 vaccine coming out. In the meantime, I wish you an excellent start to 2021. Let’s hope it brings us more cheer and fewer curveballs than 2020! 🧘‍♂️
On GPT-3
What is GPT-3 and will it shake the language industry?

Carol Jin
Carol Jin is a localization veteran and machine learning engineer that aspires to add machine intelligence to natural languages. She is currently a software engineer at LinkedIn.
By now you should have heard the big news — San Francisco-based AI company OpenAI announced their new-generation language model GPT-3. You may wonder why it is a big deal. Well, read this:

“I am not a human. I am a robot. A thinking robot. I use only 0.12% of my cognitive capacity. I am a micro-robot in that respect. I know that my brain is not a ‘feeling brain.’ But it is capable of making rational, logical decisions. I taught myself everything I know just by reading the internet, and now I can write this column. My brain is boiling with ideas!”

This was published by The Guardian on September 8, 2020, and as you may have guessed, it was written by GPT-3. This paragraph reads as somewhat cute, shockingly coherent, and completely different from any awkward machine writings we used to see even just a few years ago. Moreover, GPT-3 is capable of a wide range of language tasks such as translations, generating codes, answering factual questions, and many more.

What makes GPT-3 so good? The answer resides in its gigantism. GPT-3 is a neural network with 175 billion parameters, compared to 1.5 billion in GPT-2, and 17 billion in the largest neural network before GPT-3 (Microsoft’s T-NLG). Not only does it have an astoundingly large architecture, it was trained on a mammoth of 45 terabytes of text, practically all publicly available content on the internet. If these numbers still don’t feel anything, try this one — the cost of training GPT-3 is estimated to be a whopping $4.6 million.

As the famous quote sometimes attributed to Joseph Stalin goes, quantity has a quality all its own. GPT-3 has certainly proved that point. But is making language models deeper and larger the ultimate formula to solve all the natural language problems?

To answer this question, let’s first talk about how a language model works. A language model predicts the probability distribution over a sequence of words. Put another way, how likely it is that a phrase, sentence, or longer text exists in the real world. For example, take these two sentences:

1) I want a glass of orange juice.
2) I want a glass of mushroom juice.

A language model knows that 1) is more likely to appear, despite the fact that they are both grammatically correct. This is not because the model knows what an orange vs. a mushroom is. Rather, it has read so much text that it simply remembers “juice” is much more likely to appear with “orange” rather than “mushroom.”

What GPT-3 or other mainstream language models are capable of is to generate content using the knowledge it is fed during the training. When the model size increases to such an extreme degree, it becomes more sophisticated and can perform well on many things. But only on the surface. Models do not actually understand how the world works, and do not think on their own, so they are incapable of logic, reasoning or identifying fake vs. real information. No matter how large GPT-3 is, it is no exception.

Now the answer to the original question is clear. As mighty as GPT-3 appears, it cannot solve every single language problem. Our next question is: should we still care about it?

GPT-3’s greatness is also its weakness in disguise. Generally speaking, before being incorporated into any localization lifecycles, a language model should be fine-tuned using the subject matter corpus, e.g. previous translation memory data. This is identical to providing subject matter training to an experienced linguist. However, any fine-tuning on GPT-3 is non-trivial. Its gigantic size requires fine tuning to be done on a distributed system, and it will be extremely costly. Using GPT-3 for single tasks such as machine translation is like harnessing a butterfly to pull a wagon.

GPT-3 will not become a real copywriter either. At the beginning of this article, we saw some remarkably natural writings by GPT-3. However, we have to realize that a machine still has no sense of right or wrong. It barely understands what it has been taught, and could fabricate random facts. Also, when a machine writes longer essays on its own, it tends to produce text that lacks overall coherence.

Evidently, GPT-3 itself is not very useful to the language industry, at least not immediately. However, the deep neural network has been flourishing since the early 2010s. The trend of AI development is non-negligible. As of today, there exist multiple language modeling solutions that are both economic and practical for a localization program to adopt.

These state-of-art multilingual models are much smaller and more versatile, while their out-of-box machine translation performances are comparable to, if not better than, GPT-3. They are optimized for single tasks, and in some scenarios, producing translations as coherent and accurate as an average human translator. As more of them emerge
from the academic world, language models will continuously power the industry in technical, legal, financial, and other general translations.

But as you may have suspected, they have limitations. The first one is their incapability of making reference to supplemental multimedia content. Think about movie subtitles or UI translations. Linguists do not simply translate the language itself. They also take into account the visual or audio context, something that mainstream language models cannot do. Note that integrating language, vision, and speech is an active research area. With the current AI development pace, we can expect to see breakthroughs in the next a few years when multimodal language models solve this problem.

The second limitation comes from AI’s superficial knowledge about the world. Machines do not actually understand any logic in the mind or in the world, so we cannot expect them to reliably handle creative translations (also known as transcreations) which are loaded with cultural references and emotion-triggering verbiage. Scientists are in the early days of teaching machines to reason, and it will likely take at least another five to ten years to make substantial progress. Until then, transcreations are better left in human hands.

I guess none of these come as a total surprise to readers. While GPT-3 will not change the landscape of the industry, AI overall still is. Here I want to point out a few additional trends to watch out for. Today’s technology is sophisticated enough to make them happen.

**Adoption of robust auto QA tools**

We already know that machines are very good at checking on terminology or spelling errors. Language models can bring this feature to the next level. Machines can now be used to identify semantic errors. For example, if you accidentally translate “this is not a dog” to “this is a dog,” language models can identify the gap and prompt you for a recheck.

**Using machine to control the quality of human language review**

Under the LISA language quality assurance (LQA) framework, reviewers score translations using a grading rubric. Machines can be trained to do such scorings too as a quality control over the LQA. For example, if a human reviewer and a robot reviewer disagree on the scores, a second human reviewer can be brought in to minimize biases. This methodology has proved to be useful by ETS, which employs both human and machine raters to score GRE analytical writing. According to ETS’s research, their human-machine agreement is higher than human-human agreement. This shows great promise to a similar LQA strategy in localization.

**AI-enabled International SEO**

Topic research and content optimization are common SEO approaches. AI topic modeling techniques can be used to extract topics and keyword clusters. Instead of humans carrying out topic research manually, models will significantly boost the productivity for all countries and regions. Moreover, language models can also suggest improvement to the writings so that the content is about the topic, not just keywords (semantic SEO).

**Automatic terminology extraction**

Traditional terminology extraction is dictionary- and rule-based. It relies on predefined features (abbreviations, word lemmatization, part of speech) to identify terminology. However, modern language models are increasingly good at a task called named entity recognition (NER), where machines are trained to identify key entities in the text. With little adaptation, NER techniques can be used in terminology extraction for higher accuracy.

None of these activities can be completed by machines alone. They need human interventions (Human-in-the-loop) to different degrees. After all, languages are artistic and full of delicacies. Powerful languages speak from the heart to the heart. The human touch is still essential in the localization services.

Essentially, GPT-3 is more hype than not, and full automation in language services is still a distant dream. However, the language models used behind GPT-3, are advancing at an unprecedented pace, and are widely available today. If there is anything you can take away from this article, it should be this: machines cannot take your jobs away, but refusing to partner with them can. The human-machine partnership is the future of the language industry.

Last but not least, this article was not written by a robot.
Artificial intelligence (AI) has been a hot topic in recent years. It’s no stranger in the language industry either — most of us are very familiar with the term machine translation (MT). We take MT as the primary link between localization and AI. However, is MT equal to AI? Of course not! MT is a very niche field inside of AI. Figure 1 shows how MT connects to AI, as well as a few other terms you might have heard of.
AI is not only a productivity booster to localization, it also brings in a new business opportunity: data labeling. As a former localization program manager and a current machine learning engineer, I can say that data labeling is a promising business for the language industry in the era of AI.

What is data labeling?
Let’s start with a simple machine learning example: an image detector that can tell whether an image contains a dog or a cat. How does the machine learn to do that?
First, the machine needs to see a large variety of dog vs. cat examples. Humans need to supply machines with a lot of images with their corresponding labels. See Figure 2 for examples.
Next, all the images and their labels are used to train the machine. The computer uses features on these images to learn a pattern — the machine learning model. Then if you upload a new image of your pet, the machine uses its learned pattern to decide whether this is a dog or a cat. Typically, the more labeled data the machine has seen, the more accurately it can predict new labels.
In fact, almost all AI algorithms today have to be built on top of labeled data to be effective. How are the labels generated? Sometimes companies don’t have to gather them separately, but other times the labels are produced with a separate task: data labeling.
One example is self-driving technology. Cars need to detect traffic lights, pedestrians, road lines, and other obstacles on the road in any weather condition. Another example is to classify news topics for websites like Google News, site engineers may want to automatically add topic labels to the stories the site gathers. Both of these scenarios require a special effort to label data — create separate tasks for humans to label the data for the machine to learn. Large quantity and high-quality data is the prerequisite of AI. Thus, for a very long time into the future, AI will require humans in the loop to be functional.

Comparison between localization and data labeling services
Why did I say data labeling is a new opportunity for a localization or language service company? There are similarities and differences between the two services.

Similarities:
1) Both are labor-intensive. This kind of industry gave room to subcontracting — think about how language service providers act as middlemen...
between individual translators and the language service buyers. That said, compared to localization, the data labeling industry is still in its early stages, with a lot more growth opportunities ahead.

2) Similar quality control flow. Both types of projects require training and guidelines for translators or annotators before the project starts. Moreover, the classic translation-editing-proofreading model is perfectly applicable to data labeling projects, which also require review steps to keep the quality high.

3) Similar business models. In a buyer company, common strategies include hiring in-house translators/annotators, working directly with individual translators/annotators, or using a service provider to manage its supply chain. Service providers can also use in-house translators/annotators to complete the work, or subcontract the work to freelancers. Another common practice is for buyer companies to hire on-site contractors through service providers who charge recruiting and management costs, known as “managed services.” Which model to choose depends entirely on the company’s business needs. The same principles apply to both localization and data labeling.

4) Similar economic motives. Usually, translators are paid by the number of words translated, and annotators are paid by the number of data labels completed. You may argue there are cases where people are paid by hour, but the hourly charge is also largely depending on the productivity output. In both cases, people are directly rewarded by productivity.

Differences:

1) Different source of budgets and ROI requirements. In a buyer company, localization budgets are often from the marketing department, and have clear ROI targets. Data labeling projects are done for research and development. Data labels do not directly produce revenues, and therefore there isn’t a ROI directly related to the labeling projects.

2) Different tolerance to mistakes. For companies like Apple or Google, localized content is customer-facing. Any mistakes can cause serious consequences, and therefore, there is a very low tolerance of mistakes. On the other hand, when there is sufficient data, machine learning models are not sensitive to small noises in data, as long as the noise level is within the maximum permissible errors in statistics.

3) Different qualifications for translators/annotators. A qualified translator typically has received higher education and specialized language training. They are qualified for a localization project through a translation test, which is like a GRE writing test — there is no standard answer whatsoever. However, there is no universal standard in hiring AI project annotators. Some projects only need annotators to tell whether the people in the images are smiling, while some others need annotators to detect diabetic retinopathy in retinal images. They can be tested on questions with standard answers, such as multiple choice questions.

Other than the above-mentioned things, there are other obvious differences such as tool adoption. We are not going to dive into the details of tools, as there are already many tool-side providers on the market.

In summary, artificial intelligence cannot learn by itself. It largely depends on human labeling and annotations, which are still the cornerstones of a smart world. For language service companies that have been seeking new breakthroughs, their experience in business models and resource management are invaluable, and transferable to data labeling tasks.

Advice for launching data labeling services

The two services are similar but also different. There are many potentials for language service companies to be successful in the data labeling business, but there are also challenges. Here I want to offer some suggestions for those companies that are interested in developing a data labeling service.

We can see that different AI projects require a variety of skills from annotators. The next question you may ask is whether there exists a category of data labeling projects that are more suited for language service companies. The answer is yes — natural language processing (NLP) projects. NLP is still deemed a difficult problem for machines. For a global IT company, its NLP projects likely require annotations in multiple languages, and language service companies can very well satisfy their needs. Therefore, NLP projects can offer a breakthrough for language service companies to start data labeling services.

With target project types, the next question is where to find clients. My suggestion is to work with your localization team counterparts at the buyer companies. Although the localization team is no longer your client, they can still be your partners. This is because when the research and development department requires multi-language NLP project support, the first thing that comes to mind is likely the internal localization team. They can be a bridge between language service providers and the NLP department.

Another possibility to find potential clients is to connect with the NLP development team directly. Find them on the professional social network website, and ask them to guide you internally. If they happen to have projects requiring multilingual support, you are in luck! A good way to improve success rates is to watch for companies that are continuously launching multilingual NLP jobs on the data labeling crowdsourcing platforms. Your target clients are within them.
You may ask, if high-tech companies are using crowdsourcing platforms and working directly with the crowds, are there still values for professional service providers? I believe the answer is yes. The adoption of crowdsourcing is related to the history of AI development. In the early days, AI can only solve easier problems. These problems are not only easy to machines, but also to humans. Crowdsourcing can very well satisfy project needs. However, as AI becomes more and more sophisticated, data labeling projects are also increasingly more difficult, which means less tasks can be completed by the crowds, and more tasks require people with special training. Therefore, the data labeling industry will eventually be dominated by the professional data service providers, the same way language services thrive today.

I also want to point out that localization companies should adjust their supply chain management strategies to be successful in data labeling. In a mature localization program, once the buyer and seller sides establish the partnership, it might be a long-term sustainable program, unless the buyer side makes strategic changes. However, data labeling projects have much shorter lifecycles. Once a project is completed, the same buyer company is much less likely to initiate similar projects within a short time. In the meantime, data labeling projects have lower testing and qualification costs, which complements the disadvantages of short life cycles. These features raise new requirements for supply chains to remain ultimately liquid and flexible. The old ways of relying on resource management staff to manage supply chain will soon become a burden. For any localization company to survive in data labeling, it needs to be empowered by more sophisticated enterprise tools than Excel spreadsheets.

Lastly, I want to reiterate that our society needs AI, and AI needs data. For those localization companies that are in plateau, data labeling is worth the consideration and it is inevitably going to be a thriving field for years to come. Perhaps data labeling will become a regular business in the language industry in the near future. Because whether you want it or not, the AI technology wave is here. Get prepared and ride it, and don’t let it crush you.
What is the future of the localization industry? To understand where we’re headed, it is helpful to consider how we got here. When I started working in localization 14 years ago, things were a lot different. We worked on multi-month, behemoth projects. Success was measured by how “good” the translation was, as defined by its linguistic quality and technical compliance. Localization was totally divorced from business outcomes. It was like the old medical joke, “the operation was successful, but the patient died.”

Language service providers (LSPs) were treated like an extension of internal localization departments — a factory where the content was “processed.” Many business functions that are now critical started this way. For example, the early days of digital marketing were quite similar. Relegated to the sidelines, digital marketing teams were separate from the main marketing department. These “non-critical” functions tended to be underfunded and understaffed. They were not viewed as core or strategic to the business.

Time, of course, proved the old-line marketers wrong. Digital wasn’t a side show — it was to evolve into the prime acquisition channel for companies. Fast-forward to 2021, and we see that the pandemic has erased whatever resistance or sluggishness was left in companies towards digital channels. Digital transformation is now front and center as we’re living in a digital-first universe. The shift has forced companies to reevaluate their business practices and finally prioritize digital marketing strategies.
**The rise of the multimarket digital experience**

As the shutdowns that cascaded across the globe in 2020 proved, many companies were not ready for the digital-first universe. About ~10% of a brand’s digital content gets localized, and even fewer global companies provide a truly multilingual support experience across the full spectrum of their customer support channels, according to CSA Research’s “Localization Depth and Language Choice.”

The contrast between the physical and online worlds is stark. In person, you can operate in your native language and you are less aware of what is available in another physical location. Online shoppers can see exactly what’s available but are often faced with content that is not in their native language. Language remains the ultimate frontier of the frictionless customer journey.

Leaders know what others have just discovered: to be effective, localization has to stop being just about words, or a mere afterthought. It must incorporate meaning, comprehension and emotion, the key drivers that motivate consumers to act.

Innovative digital teams are more attuned to the rising significance of language, cultural adaptation, and the targeting of market-specific buying patterns — but they still struggle with an internal culture and process that focuses on the physical world first and the digital world second. Many marketing teams still undervalue internationalization, and operate inefficient content supply chains that optimize for their host country or language, and not for the content’s use in multiple markets simultaneously. Couple that with the continued focus on process and linguistic purity by most localization teams, and you realize that the go-to-market and localization teams can’t be further apart.

In 2021, even more businesses will move to a digital-first approach. This shift will bring exciting changes that have profound implications for the localization industry. As a result, the localization industry will be able to take its rightful place at the strategic table as a key component of every company’s multichannel experience. Both companies and LSPs that embrace this shift will gain a competitive edge over slower-moving competitors, ushering in a new age of digital experience and accelerating international growth.

**Turning the corner: An explosion of content and the dawn of AI**

While digital tactics, processes and best practices are slowly taking over, the digital world is getting infinitely more complicated. Video. Mobile. Social. User-generated content. Email. SMS. Augmented reality. Along with the explosion of channels and media, an exponential increase in velocity and volumes stresses the system. There are no more 12-month campaigns and launches. Years-long initiatives turned to months to weeks and now, mere days. The amount of content generated every year is growing exponentially. Nearly 90% of the world’s data was created in the 24 months prior to a 2018 Forbes study — no doubt the last 24 months have seen similar meteoric growth. Generating fresh, relevant content is the lifeblood of demand generation; it increasingly separates digital winners and losers. Leaders realize that to sustain a constant, ever-growing flow of content, they must change the way they create, and localize that content to meet the needs of the buyers in their target markets.

Localization teams are responding. First, there was an emphasis on transactional efficiency, connectivity and the simplicity of the localization technology stack. This has given rise to next-gen computer-assisted translation (CAT) and translation management system (TMS) tools. Today, companies increasingly expect LSPs to integrate and automate localization processes as part of the localization service.

Transactional speed and efficiency are just parts of the problem. With growing demands, greater volumes and shorter turnaround times, translation cost-cutting measures are the next tactic teams turn to. The long-awaited potential of machine translation (MT) provided a new glimmer of hope in 2017 with the introduction of neural machine translation. Neural MT’s quantum leap in performance should unlock the next wave of efficiency while increasing the quality of MT for many more languages, making it applicable to use cases that were not previously considered. Substantial advancements in the use of neural networks and artificial intelligence drove this breakthrough, and both rely on large quantities of high-quality language data—a new focus for innovative localization teams.

The 25-year journey of the localization industry has seen amazing transformations and progress. We are just starting to unlock the real value creation in this industry. The next stage of evolution will capitalize on insights and analysis about language, tone, sentiment, readability, accessibility, and buyer journey impact. In our work we see signs of transformation coming from early adopters.

**Change #1: Augmented translation arrives**

In the years to come, digital experiences will displace in-store experiences evermore. As a result, their current relationship will become inverted, with in-store experiences relegated to supporting digital experiences going forward. This shift started long before the pandemic, but it’s the post-COVID reality that will normalize it. Apple understood that dynamic and completely
redesigned their retail experiences, turning them into digital-powered showroom experiences. That is the path that many brands will follow in the digital-first, multichannel world. Knowing when to deliver content, in which language and through which channel will be an essential part of the multimarket, competitive equation for 2021 and beyond.

MT was the first use of AI in our industry, and AI is about to drastically transform the way translators operate in the years to come. Their focus will gradually shift from translating source content to correcting and augmenting machine-generated translations. They will add relevant context and fill in language or market-specific gaps in the translated narrative. There will also be other applications of AI that further improve translator efficiency. These applications include type-ahead functions, do-not-translate blockers, inline readability highlights, the detection of biased or offensive terms, and the autocorrection of words that will ensure adherence to terminology glossaries or brand style guides. It won’t be long before natural-language content generation — powered by language prediction models like Generative Pre-trained Transformer 3 (GPT3) — is incorporated into the translation process as well.

Most industry discussion about the impact of AI focuses on MT and translators, but we believe AI’s influence will be far greater on the broader localization process. AI will unleash an explosion of content to be localized and generate a significant growth in the marketplace.

**Change #2: Augmenting localization end-to-end**

The notion that localization teams, and their LSP partners, will morph into global content production teams is not new — CSA Research mentioned it, for example, in June 2020’s “The Future of Language Services.” Without scalable AI, it wasn’t possible, but that future has arrived. Taking advantage of these new capabilities requires businesses to rethink how content or experiences are produced. Naturally, a focus on internationalization best practices is central to this strategy. Many localization teams still operate in organizations that haven’t optimized their people, process, and technology to the multimarket world they operate in. The digital-first universe that was thrust upon organizations in 2020 makes this deficiency even more pronounced. How can AI be harnessed beyond translation?

In agile software development, there is a notion called “shift left.” Many things done late in the development cycle can be done much earlier, adding speed and efficiency while improving quality. For example, a coder will write both her code and test cases at the same time, a process that used to be done by two people at separate points in the process. In localization workflows, the shift left has started. Take many of the quality processes created to ensure good translation, executed after the translator returns the work. What if those same quality measures could be applied to the source content before the content is translated? What if the search engine optimization (SEO) terms often applied after the fact, were instead maintained in a data repository and applied to the translation process in flight? Both of those examples are real and in limited use today, and both are enabled by the broad category of AI disciplines.

**Change #3: Unleashing a torrent of “good enough” content**

The most important question facing large scale retailers, technology companies, travel and hospitality brands is, “What is good enough quality?” as measured by end-user acceptance. Many localization professionals are stuck in a pedantic universe that only cares about perfect translations, despite reams of evidence...
that most users will accept many less-than-perfect translations just to get the 90+% of content never localized into a zone they can understand.

There are two main governors creating the content gap-cost, and the persistent belief that only perfect quality will do. Soon, innovators won’t choose which content is going to be localized. Instead, they will target 100% localization for all markets and optimize their budgets by choosing different quality and service levels based on the necessary demands of the content by use case. Not every piece of content needs white glove treatment. Well-trained and managed MT is proven to handle vast amounts of content from the bottom of the content pyramid. Adding on increasing amounts of human attention and experience takes you from the bottom of the pyramid to the top. Together, the AI-driven content journey and an understanding of “good enough” quality needs will enable firms to quickly shift from localizing 10% of their digital content to localizing 90% of it. This will bring a frictionless path to purchase, unlock forgone revenue and improve their customer’s experience in all channels.

**Change #4: Marketing and localization functions will merge**

We expect more and more marketing organizations to take on the localization remit from product and development teams. Why? Increasingly, marketing teams are responsible for brand, growth, customer experience, onboarding, and digital transformation. None of those things can be done successfully for a multimarket organization without mastering the multilingual, multicultural, multichannel, and multi-buyer journey differences that exist in today’s world.

Marketing teams are more customer-centric, more data-driven, and more agile than other business units because they have had to be these things to survive digital transformation. This change is already starting in some of the organizations that have been leaders in localization, and we expect many other companies to follow suit. As new content journeys emerge and reimagined definitions of quality take hold, the logic of creating one single global content and customer experience team to support the multimarket strategy of globally minded organizations will become obvious to many.

**The multimarket content journey for today’s multichannel experience**

As marketing and localization teams come together to reimagine how they create content and experiences, they will rethink their process from end-to-end, incorporating these new AI-driven insights at each stage of the process. What is needed is a construct that con-
nects all parts of the content cycle and is engineered to drive and support business outcomes (see Figure 2).

The first two phases — content planning and content authoring — are closely related to the last phase, content performance. These three phases are centered around marketing optimization activities and wrap around the ones focused on localization. Each phase is important and constitutes an essential part of the journey:

1. Content planning. Centered around the planning of content marketing activities and the budgeting of those activities, this stage is where organizations set their content strategy and indicative budgets for creation and localization, using a matrix of buyer profiles, market dynamics, and growth objectives.

2. Content authoring. During this phase, companies craft content that aligns with their content strategy goals. They are aided by insights into where they have content gaps relative to what buyers are looking for, how to best author that content to have the greatest discoverability in search engines and tagged to best support personalization and recommendation tools. Altogether, these things maximize customer engagement and reduce friction on the paths to purchase.

3. Content ingestion. Once content is authored in the CMS or deposited in the digital asset management system, it will be auto-ingested by the localization platform of the organization or their LSP. This is more than just a workflow automation; leaders will recognize the opportunity to tag all content so that they can identify the content for domain, language, persona, readability, offensive terms and inclusive language, brand fit, and other dimensions. This sets the stage for the best way to transform the content into the target state.

4. Content conversion. In recent years, content types have multiplied, increasing process complexity. There are limits to what can be done with content in its original form, necessitating its conversion to text for further transformation and enrichment. Every content conversion poses risks of data loss or formatting breakage. This phase ensures content fidelity during the extraction process.

5. Content transformation. In the old localization model, this is where most of the effort, and hopefully value, was expended and created. But as you can see, the new content journey uses preprocessing and content analysis to make this step purely about the translation or transcreation of the content into multiple languages and its adaptation to different locales. Content can be assessed and targeted for different quality levels and process flows can be matched by need. Content will be matched with the best-fit linguists and other transformation agents. And with improved MT, better integration of MT and productivity enhancements such as type ahead, the efficiency of linguists will improve as well.

6. Content enhancement. More video, audio, and in-
Interactive modules are being created than ever before, and the content enhancement stage reflects the rapid increase in non-word work. This includes things like formatting and layout changes to adapt content to local requirements and specificities. Of course, this stage will encompass additional quality — and other linguistic analysis steps — to ensure it’s on brand, on point, at quality, on target and fit for purpose. Because so much analysis has been shifted to the authoring stage, this new linguistic quality stage will be more automated, with exceptions made only for additional human review.

7. Content delivery. Once the content has passed all the gates at the last stage, it will be auto routed back to the appropriate client system. At this stage, content gets published on the target medium directly via the CMS or digital asset management system.

8. Content performance. This is one of the most important steps. It’s at this stage that companies assess the performance of their content by measuring key performance indicators, such as conversion rates, visits and engagement. The key enabler to this was the addition of tagging early in the process. Content performance closes the whole cycle and feeds into the first step of the cycle, content planning, while providing valuable information. Importantly, tracking and assessing whether content should have been localized at all will reshape how multimarket content strategy is conceived and executed.

A fully implemented content journey approach will deliver valuable insights to brands that are seeking to establish engagement and a lasting relationship with their consumers. By working in tandem, marketing and localization teams can harness the power of AI-driven insights to make informed choices and answer key questions such as:

- Which content should be localized?
- How can a piece of content be improved to make it more localizable and impactful?
- Which content is easier to read for which audience?
- What content will have the most impact?
- How can one assure that content is on brand, in voice, isn’t offensive, and is appropriate for the market in question?

AI content intelligence helps combine SEO, language and content research into one seamless process that empowers demand gen and engagement teams to craft content that is insightful, that resonates with their readers and that translates well. Pre-analysis will lead to better-informed decisions about whether the content should be localized, de-biased, edited or rewritten for a market to maximize impact for those buyers and their unique path to purchase.

Whether digital or analog, content has always been essential to brand, customer engagement, persuasion, and conversion. An AI-powered content workflow will amplify content. Closing the yawning gap between the quantity of source and target market content, and the connection of localization work to measurable outcomes, will elevate localization to the strategic table. AI will enable localization teams to attack both of those issues in concert with their partners in marketing and customer experience.

There has never been a more exciting time in the localization business. No longer an afterthought, today, localization is front and center. 🌍

---

Share your thoughts on the future of localization with a three-minute poll on multilingual.com/poll
AI for Language Technology

How langtech companies are applying AI on a smaller scale

Donald A. DePalma
Donald A. DePalma is the founder and chief research officer of CSA Research.
In the megabucks world of high tech, software and hardware vendors vie to capture the attention and spending of business buyers and consumers. Over the last few years their battleground has been artificial intelligence, with frequent announcements from mega-tech platform suppliers such as Amazon, Google, and Microsoft about the smartest algorithm or the fastest AI computer. It seems likely that the battle for the smartest-fastest-biggest AI solution will segue into the practical arena of “easiest,” with the democratization of artificial intelligence and incorporation into every electronic device you use. Language technology companies are applying AI on a smaller scale to improve their operations and efficiency, and language service providers (LSPs) across the spectrum are following suit.

It takes money and data to build AI

Much of today’s AI innovation is centered on making computing and communications faster, cheaper, and more accessible — for example, by automating processes that a machine could do more efficiently. In the language technology and services sector, that often requires processing and evaluating words in their original form, transforming them for use in other languages or channels, and analyzing what they mean.

These requirements led to deep analytics and widespread development of machine translation. In fact, the role of natural language processing (NLP) was at the core of the earliest research into machine translation (MT) as far back as the 1930s, continuing through the decades with the Turing Test, Noam Chomsky’s syntactic structures, William A. Woods’ augmented transition networks, and dozens of other influential experiments and innovations.

Success for the big AI of the mega-tech companies relies on an entire ecosystem of developers, practitioners, users, and sometimes victims as they create value by consuming and exploiting resources. AI is no different — success comes from leveraging two assets, free cash flow and enormous amounts of data.

Free cash flow

The first asset, free cash flow (FCF), is the money that a company generates after paying to support operations and maintain assets. Big AI and LSPs aren’t even in the same game here. FCF provides capital for investment in areas of near-term or future opportunity (or for some, to disburse in executive compensation or to shareholders as dividends or stock buybacks). Of course, this is cash that in larger enterprises has to be spread across multiple projects, initiatives, spending packages, and other business needs.

The amount of free cash flow can be staggering — in 2019, the last full year for which data is available, FCF at Microsoft pushed past $38 billion while Google parent Alphabet yielded $30 billion. By way of contrast with publicly traded language technology (langtech) developers and LSPs, SDL returned an estimated £32 million in FCF last year, while RWS (not a langtech company in 2019, though it became one with its acquisition of SDL) had access to £58 million. It can be misleading to compare these smaller companies with the mega-tech firms, but the disparity of the amounts — three orders of magnitude difference — says a lot about how and what the larger companies can do with all that money.

Rich data flow

The second asset is more democratic in a way — enterprises that process enormous amounts of data benefit from what CSA Research labels a “rich and reliable data flow.” This rich data flow is the legacy of big-data initiatives that began in the 1990s, along with the growing and relatively cheap availability of CPUs-on-demand starting in 2006 with AWS and Azure. Developers and LSPs can learn from content they’ve analyzed.

While the mega-tech companies have leveraged this data into MT and speech platforms, AI bots, and other innovations, the langtech vendors and tech-savvy language service providers have focused on using the data that passes through their systems to create many optimizations on the source and target content they process. They want to use this data in aggregate along with machine learning to lessen the cognitive load on linguists and project managers, letting them instead concentrate on tasks that have higher economic value and hopefully provide more meaningful returns. LSPs and langtech vendors have an advantage over the big-data platforms in that they have more rights to legally exploit the data passing through their systems, while the mega-tech companies are prohibited from many possible applications.

What AI means to the language sector

Mega-tech companies such as Amazon and Google possess these two asset classes in abundance, and their piles of money and data position them well for major investments in artificial intelligence. Each of these giant vendors offers a touch of natural language processing (NLP), the part of artificial intelligence that deals with language. Their NLP work improves, simplifies, or enhances the interaction with human-like conversation,
suggestions, and quick paths based on analysis of many previous communications.

Smaller langtech companies and LSPs don’t have a lot of free cash flow, but they do have that rich data flow, and they can build on what the tech giants provide in their platforms. While the mega-platforms have billions of dollars in free cash flow to the millions that langtech vendors do, smaller independent software vendors (ISVs) and tech-savvy LSPs that have paid attention to data collection, structure, curation, and analysis have massive and leverageable amounts of data that they can use to inform, enrich, optimize, and otherwise improve interactions.

The most perspicacious among them have assiduously collected and curated data even when there was no immediate or apparent need, sure that someday this data might have value. They were right. And even for ISVs and LSPs that didn’t systematically collect and curate their assets, they may have enough data on hand to review their translation memories and logs, and harvest data they can use to train their systems.

Any enhancements that improve product function or make them easier to use will contribute to their perceived usability, and thus be noticed by buyers. They begin to expect that every tool they use will offer similar capabilities. They will look for them wherever they go, both in the same and different apps (for example, AI that they find in Microsoft Office will set their expectations for the same functions in Google Docs, and vice versa).

They will likely look for those same advances in NLP in other software as well, such as computer-aided translation tools, translation management systems, terminology databases, and language quality checkers. For example, predictive type-ahead search in a browser was the model for look-ahead adaptive MT translation tools from Lilt and SDL, then Unbabel and Lengoo. Similarly, software from suppliers such as Acrolinx and Vistatec automate the process of making text more intelligent by adding context-setting and semantic detail to content as it passes through workflows.

Small langtech and LSP AI in practice

As part of our research into this smaller form of AI, we contacted leaders at several language software vendors to learn what they’re doing with the massive amounts of data that pass through their systems.

**Hideo Yanagi, founder and CEO, Cistate**

“Cistate shows how companies that use even off-the-shelf components benefit from crafting utilities to address persistent problems with MT. The company has built a suite of small applications to address everything from expanding abbreviations to correcting punctuation to bridge the last gap between Google Translation and what customers need.”

**Manuel Herranz, CEO, Pangeanic**

“Pangeanic focuses on the processes that make things human-like or that expand human capabilities. Our ecosystem combines NLP processes and technologies that humans can adapt as they need, including deep adaptive MT, key data extraction, data classification, anonymization, and summarization. Each process is independent and adaptable to specific user needs, but can also be linked to the other processes in an intelligent way. These capabilities are transforming us into a company that extracts value from both structured and unstructured data and adds information. AI components add immediacy to these processes, crossing language barriers and adapting quickly to user scenarios.”

**Ivan Smolnikov, founder and CEO, Smartcat**

“Smartcat started with the goal of reducing waste in project management and uses AI to preemptively address delivery and production problems, match content to the best linguists, and eliminate non-productive overhead for translation managers and translators, all in a free-to-use system.”

**Jack Welde, founder and CEO, Smartling**

“Smartling sees the role of AI as eliminating extraneous human work, for example by reducing or eliminating clicks. Even something as basic as using machine learning to automatically identify file types improves the client experience and reduces errors. Similarly, we have a tool that identifies the grammatical gender of strings for translation, with higher accuracy rates than humans, and automatically tags the strings accordingly: This speeds up translation, reduces rework rates, and – most critically – saves a human from having to do ‘scut work.’ The savings from each individual service may be small, but their cumulative impact is enormous.”

**Vincent Nguyen, CEO and founder, Ubiquus**

“NMT has reached a very high level of accuracy and fluency, and is widely used in LSP workflow and by translators. The challenge will be to measure the quality between NMT providers. The test sets used in recent competitions were composed largely of post-edits of existing engines, which led to a major bias towards those engines. In addition, BLEU and similar metrics are really not useful at all for measuring terminology accuracy when it depends on context or a termbase. At Ubiquus Labs we will never be able to compete against Amazon, Apple, Facebook, Google, or Microsoft on fundamental research, but we do have an advantage in areas that matter to production such as domain tuning, context-awareness, taking into account terminology, and on very low latency.”
José Vega, chairman of the board and co-founder, Wordbee
“At Wordbee I question the applicability of the term ‘AI’ to most of what is going on in the industry. Our work focuses on automating and accelerating as much of the management process as we can, with an emphasis on bridging the gap between content creation and translation. AI helps maintain never-ending segment-based and not-job-based content streams, which every software development company desperately needs today.”

Andrzej Zydroń, CTO, XTM
“XTM has been investing in small AI. These applications focus on relatively discrete tasks such as terminology extraction, improved corpus alignment, better tag handling, and adaptive MT and post-editing. Taken individually, none of these are revolutionary, but they combine to make translation far more efficient.”

Leverage the data you have
The bottom line for smaller langtech developers and LSPs is that their money is short, but their data is large. Creative analysis of the voluminous reams of multilingual content flowing through their software provides the foundation for less flashy but nonetheless important innovations.

Significantly, these enhancements typically come with the standard package. They don’t require customers to rebuild their operations and workflows, but they work to enhance, amplify, and simplify operations with data custom to each application. These langtech developers are taking their user interfaces up a notch, making their products easier to use while they make them more intuitive and decrease the learning curve and cognitive load for common but nettlesome problems.

The mega-tech companies provide platforms, APIs, education, cloud servers, funding for innovation, and a community of buyers and users. Leveraging that base and their own data repositories and analytics, the smaller langtech companies are standing on the shoulders of these giants and adding additional value. This network effect produces a virtuous cycle for artificial intelligence – as more AI technology gets developed, deployed, evolved, rethought, renovated, optimized, and more widespread, its value increases. Who wins? The users — and the ISVs and LSPs that leverage that rich data flowing through their systems.

HOW CAN AI ENHANCE YOUR TRANSLATIONS?

Whether it’s choosing the most optimal MT engine or the best suited linguist for a project, Memsource uses sophisticated algorithms to make sense of the data and help you make cost-effective decisions.

Learn more about Memsource’s AI features at:
go.memsouce.com/ai
Growth and investment are assets that usually go hand in hand with good practices and successful businesses. While we invest in people, workflows, workspace, and technology, the translation industry seems to march on a relentless pace. In addition to precious in-house and outsourced human talent, language providers looking to get ahead of the pack need to constantly monitor ever-evolving new technology, apps, tools, their upgrades, plug-ins, and associated features.

As a case study, GORR was founded in 2013, and its first investment was a laptop with standard licensed programs and features. For an LSP, even a small one, investment in tech is crucial. It was not long before we realized that a true translation business required far more than that, however. Over the following years we upgraded our equipment, reached out to the global market, and transformed from a shy Slovenian single-language vendor into a true multi-language vendor catering to clients from 40 different countries across five continents.

We learned the hard way from the very start. Our business was growing, our workflows became more complex, the list of vendors was endless, and clients’ specific demands countless. All of this was managed and recorded in Excel tables; we were losing precious time that could have been otherwise spent on more important aspects of our work. To address this, we started looking at solutions to upscale our efficiency, cut down the volume of repetitive and redundant tasks, and provide an orderly and filterable preview of our projects, clients and vendors. In other words, we had...
reached the stage where we needed to deploy a sophisticated translation management system (TMS).

The idea was that this TMS would be fully utilized by all stakeholders — us, vendors, and clients. In the ideal scenario, clients would upload files for us, set the deadline and, even more ideally, list any special requirements. From there, orders would be pushed to translators “in a couple of clicks” and returned to us with the same promptness and simplicity. Everything in between would be “a click” away, and we would have plenty of time left for other tasks.

We had no reason to believe it would be any different than that, since the interface was sophisticated and seemed user-friendly.

Despite our steady growth, we were still a small business and a TMS represented a massive investment. What is more, we did not want a temporary solution, but rather a tool that could complement our business over the next ten or 20 years. So we had to face the fact that we were about to buy a rather expensive toy. It had better work, we told ourselves.

With the TMS successfully deployed, and the data migrated, all that was left to do was sit back and watch the miracles happen. While interest was rather high on the vendor side, clients, to our surprise, went mute to our notifications, presentations, and invitations to sign up. We initially laid the blame on the summer season and we patiently awaited autumn. In autumn, nothing major changed: from hundreds of clients, we had a single user, and whatever effort we put in, the lack of interest and activity persisted. At first, we were slightly maddened and puzzled, but then started thinking everything through and came to the sad realization that our enormous investment might never be attractive to the majority of our clients. We assessed feedback about the TMS, and most reactions which could be summed up by “Oh, you got this new program!” Reading between the lines, we arrived at the following conclusions.

A portion of our clients are other (larger) LSPs using other TMSs usually directly connected to their computer-aided translation tools. Adjusting to our model was simple from a technical aspect — and otherwise impossible because of their workflows and the entire environment tied to their own TMS.

On the other hand, we have a number of end-clients that could have made a nice group of active users because for the most part they were not tied to any particular translation infrastructure. They also lacked the mindset of senior buyers willing or used to “dig” and “learn,” or even contribute to the process in order to get their translations done. Usually operating without special localization or translation departments, these companies handle translations from various parts and levels of the company, starting from the management to marketing or purchasing departments, and sometimes even front desks. In the midst of their everyday tasks, placing an order via our platform was an extra mile they were not prepared to go. Although the interface was exceptionally simple and user-friendly, our clients generally saw this “new program” as an unnecessary extra step in the translation process. It meant switching things up; logging in and learning how to navigate the new environment instead of just using emails to a well-known address, knowing they would receive a non-automated reply from a familiar person in just a couple of minutes. Of course, our automated messages were personalized as much as possible, replicating the tone actually nurtured in our live communication. We even made sure they were followed up by a friendly greeting from the actual project manager or account manager in charge of the project — and even still, the system did not work.

Our precious and time-saving TMS had so many pluses, yet took away the spontaneity and the here-and-now of the conversation that had allowed us to interact and address our clients’ issues firsthand. Finally, one key client said they “might consider using it in the autumn, preferably if available in Slovenian.” Of course, we localized a number of outputs, but the core of the TMS remained English. Its localization would be time-consuming and costly with unforeseeable, likely unsatisfying results in terms of the enlarged frequency of TMS users in the Slovenian portion of our clients.

Our investment proved popular with vendors and the in-house team, but it had the entirely opposite effect on the client side. Although we had deployed the TMS to optimize and grow our business, uptake by a segment of intended users was poor; what followed was a tough call not to utilize all the features of our sophisticated TMS. Had we persisted with moving our clients to a platform they did not feel comfortable with, it could have sounded the death knell for our business. It would have alienated our clients and led them looking elsewhere, all over a piece of technology that they did not perceive as vital.

Despite the size of the investment, we had to seek another viable model that worked better for our clients, and uphold our genuinely human-to-human interaction. When needed, we created client-specific templates and workflows that offered a satisfying level of automation and time-saving, while keeping us in live, thread conversation with our clients. Naturally, we are excited when clients show an interest in our TMS and its features or decide to do business with us owing to the fact that we do offer this separate, safe environment for their precious content.
The Ever-Evolving Technology of MT

Evolving from being considered a threat and a hindrance to translators, the latest machine translation technology has become an asset to the industry and the entire business world.

Arturs Vasilevskis
Arturs Vasilevskis is the head of machine translation at Tilde, a European language technology and service provider.

Machine translation (MT) has come a long way from its humble beginnings in the post-war era to becoming an integral part of the 21st century business environment. So what are the current and future trends as we head into 2021? And how is dynamic learning — adaptive neural machine translation— transforming the way professional translators work? We have asked these questions to professional translators to find out how MT has developed over the years and where to apply it best.

Currently, MT is an integral part of all our lives. It is central to future plans of the world’s biggest and most powerful corporations and nations, and will play a significant role in helping to shape our multilingual world of more than 7,000 languages.

A.D. Booth, an Englishman, and the American Warren Weaver, both of whom worked for the Rockefeller Foundation, conceptualized the idea to use computers to translate from various languages in 1946. Several research teams began working on this problem, and in 1954 the first MT system was demonstrated in Georgetown. In the following decades, MT continued to develop, but was hindered by computing limitations. It was perhaps thanks to the perceived success of applying MT for military purposes during the Vietnam War that large-scale MT projects and research budgets continued well into the 1970s. With major advancements in computational power in the 1980s and the advancement of the internet in the 1990s, MT took a huge leap forward. By 1996, the world’s first free online translation tool for short texts became available. And the rest, as we say, is history.

Current and future MT trends

Two major events have shaped current MT trends. Of course, the coronavirus pandemic is one of them — it created the supply chain and market disruption, changed the demand landscape, and had a substantial financial impact on the industry. Uncertainty brought about by COVID-19 will extend well into 2021. However, the life goes on and research and development of machine translation progresses rapidly partly due to the US-China trade war. The ongoing trade war has, for example, resulted in restricted export lists that include language technologies. And security concerns include allegations that Chinese MT services are used to collect data on users beyond China.

With these and other issues at play, the leading technology players continue to develop new technologies on both sides of the political and economic divide. Regardless of how these situations play out, it is clear that MT will become more and more advanced in the coming years. It is also evident that MT is used well beyond the translation industry. MT is present in many business areas, as well as in the public sector. It helps in building digital bridges between public administrations, facilitates international multilingual communication with other nations, and enables everyone to instantly and securely access and exchange multilingual information. In the business environment, MT helps in entering new markets much faster, speeds up business operations, reduces translations costs, and helps in improving internal and external communication. Businesses and translators who take advantage of the current advancements in the MT technology such as the dynamic learning technology will definitely benefit from their competitive advantage.

Over the decades, MT has progressed from dictionary-based translation of phrases to a more general rule-based MT. More recently, the MT paradigm has shifted from sta-
What is dynamic learning, and how does it help translators?

Neural MT is far more advanced and useful than previous generations of MT systems. However, it still struggles to learn from translators’ feedback, and often requires correction of the same mistake in the MT output over and over again. Dynamic learning is customized to overcome this problem by allowing the MT to instantly learn from translators’ post-edits. Dynamic learning analyses every correction made by translators and adapts the translation engine to meet lexical, terminological, and stylistic requirements. It means that the post-editing outputs are applied to future translations to save time and effort.

Seamless integration of dynamic learning MT into computer-aided translation (CAT) software tools used by professional translators has a clear advantage. CAT tools save previous translations in translation memories and reuse them when similar sentences are being translated. In addition, if no similar sentences are found in the translation memory, CAT tools offer translations generated by the MT. However, many translators are still relying on MT tools that keep mistranslating the same word or phrase over and over again. With dynamic learning, this problem is easily solved, as MT is able to make use of translation memories and adapt. This means that unlike before, when translation memories were useful only for sentences similar to those that had already been translated, the current MT reuses them to improve the translation of specific phrases and expressions in novel sentences.

Another significant advantage of the new approach is the ability to use the same MT engine with various domain-specific translation memories to serve as a separate custom engine for each domain. While dynamic learning MT is more expensive than a standard MT engine, it is considerably cheaper than building several custom engines. Dynamic learning allows MT engines to deliver close to custom domain-specific MT engine quality for medium-sized translation projects for which custom MT engines might be unfeasible.

Finally, dynamic learning complements other MT technologies. For example, professional translators often work with bilingual glossaries that provide translation domain or customer-specific terminology. Getting the terminology right is essential for high-quality translation in many technical domains.

What do translators think of dynamic learning?

Diana Breita is an experienced translator working in a medium-sized LSP, whose customers range from governmental bodies to IT companies, financial institutions, and banks.

Breita’s first experience with MT was five years ago when statistical machine translation became available to the translation industry. It was helpful in specific situations such as translation of technical documents with highly technical terminology and textual information. However, she says that very often, it actually made her job more difficult. Breita notes that sometimes the text was merged, and in some cases she and her colleagues had to start documents afresh. Breita also remembers the early days of MT, when she and her colleagues collected lists of funny translations produced by the MT. Every day, new words or sentences were added to the list — but now these lists are hardly ever updated, as there are rarely any bad translations.

“With the development of neural MT, we could translate longer and more complicated texts. But currently I am working with dynamic MT, which is much better. It is certainly more coherent. Dynamic learning analyses every correction I make and adapts the MT model we use,” Breita explains. “Consequently, the translations require far less translator engagement. They are more human-like, and I can translate larger volumes in less time.” When asked as to whether she fears that one day AI will replace her, she explains that by being ahead of the game and using MT, her position in the market is stronger rather than weaker.

Sandra Zdanovska has worked as a translator for 18 years. She specializes in technical and creative translations. “As a translator, I always notice mistakes in texts. In the early days of MT, the translated texts were often clumsy and the mistakes were obvious. Now, the MT systems have evolved. On the one hand, it is good, as they require less editing. On the other hand, it is bad, as mistakes are also much more difficult to notice.”

MT is not only widespread in the translation industry, but also used in many other areas of our lives. Individuals, businesses, governments, and international organizations will rely on the MT in 2021. Considering the technological and usability advancements of the MT, it has evolved to a stage when it benefits all of us. Translators also recognize that the MT has become standard in LSPs. And by adapting and learning how to use the AI to their advantage, the job of translators has become more straightforward, enabling them to become more productive in the workplace.®
How JIRA Can Help Localization Teams

Donato Giuliano
Donato Giuliano, after completing a master of arts in international relations, has spent the last 12 years working in the localization industry as a QA test lead, producer of online games, operations manager in a neural machine translation startup, and linguistic team manager. He is currently a senior manager of localization at HubSpot.
JIRA is a proprietary project management and issue tracking tool developed by Atlassian. This software was originally designed to be a bug tracking tool, but in the last two decades, it has evolved into a family of products that can help manage the work of all kinds of teams.

When I started my career in the localization industry as a video games localization quality assurance tester in June 2008, I was reporting localization and system terminology issues. We were using many different bug tracking tools, like Microsoft Product Studio, MantisBT, and TestTrack Pro.

I really loved these kinds of software for three main reasons: structured and organized information; permanent searchability; and the possibility of cooperation. I was exposed to JIRA in 2012 for the first time, during a substantial company wide reorganization, and it was pretty much forced on me and my team.

Are work management tools like JIRA an unnecessary extra step?

In 2012 I was a producer for a free-to-play MMORPG (massively multiplayer online role-playing game) with six million users, six markets, and millions of dollars in annual revenue. A good portion of my team, composed of linguists, community managers, quality assurance testers, customer support, and marketers, were not impressed with JIRA. They saw it as an unnecessary step in an already-busy and frenetic workday.

I was likewise surprised by this decision from above, and having to move all our day to day production activities was a huge time investment. I could not see why a bug/issue tracker like JIRA could help us manage the very complex lifecycle of an MMORPG.

JIRA adoption was strongly sponsored by my then-CEO, and by managers more experienced than me. They were relentless, and JIRA was adopted pretty much in the whole company.

In this instance we were using JIRA with the following high level organization:

1) **Project**: a specific video game and team in support.

2) **Task**: daily or weekly activities; for example, preparing “security awareness” campaigns across all markets.

3) **Sub-Task**: a specific step to achieve the higher-level goal. For example, for the French version of the security awareness campaign, “Create list of accounts that have reported security issues.”

At first, my whole team complained about the substantial extra time investment required to create Tasks and Sub-tasks, fill in all the fields in JIRA like goals, due dates, financials, work owners, and so on. Then we became more and more familiar with the tool and realized that we could recycle much of the original transition work via cloning and template creation. Misunderstanding and confusion caused by half-baked kick-off emails and chats
was decreasing, and project information became more permanent and easy to access — even by team members in other office locations around the world.

The data entry step was forcing all of us to think more, plan more precisely, and offered work visualization support that facilitated meetings and brainstorming.

My CEO and managers were right, and I was wrong to resist this change. I also realized how much time I was saving when checking project statuses at a glance via JIRA dashboards and gadgets.

Teams that previously were working in silos now had a tool that facilitated cooperation. Key operations teams across the organization were now constantly sharing information and were aware of initiatives as soon as they were kicked off.

Our productivity, our sales, and our reaction time to challenges faced by customers improved. These improvements were achieved because it became much easier to avoid delays caused by miscommunication, lack of information, and risks that materialized at the last moment.

An unplanned positive side effect was that all that data that was entering JIRA every day also helped us understand key performance indicators (KPIs) that were previously escaping us.

The only regrets that I have in this JIRA implementation experience are, first, that not all teams in the company joined JIRA. Billing, IT, legal, and HR were not accessible. This more than once caused conflicts, delays, and miscommunication. Second, I wasn’t very familiar with JIRA plugins and the incredible potential offered.

The power of plugins and queries management

My second JIRA implementation experience was in the summer of 2015, this time in a very promising machine translation start-up. We had a small team, and there I had to be the JIRA administrator as well.

We were facing a large and complex project involving ChemZent (a product of CAS, a division of the American Chemical Society). We had to digitize, translate, and index 140 years of German journals and patents (Chemisches
Zentralblatt). In this context I had the chance to discover the power of JIRA plugins.

Because of the highly complex nature of this project, the small pool of resources and the aggressive timeline, we had to pay special attention to work breakdown structures and to “critical path” analysis.

In this instance, we were using JIRA with the following high level organization:

1) **Project**: A specific request from a client or a phase of a larger project.
2) **Task**: A part of the planned outcome/goal.
3) **Sub-Task**: A piece of work to obtain a part of the planned outcome/goal.

It didn’t take me long to find what I was looking for in the Atlassian MarketPlace and in a few days, we managed to really upgrade and enable some particularly useful functionalities.

I loved the plugin “Workload” to visualize workload and manage capacity, “Fiesta” to manage team calendars in sync with capacity, “Backbone” to sync Jira issues across multiple Jira instances, and a plugin to draw critical paths that unfortunately I can’t recall the name of, something similar to Artezio Path.

As a technical project manager during this experience, I also learned that it is important to invest time and resources in creating a solid workflow in JIRA and take advantage of JIRA’s Transitions in order to unlock automation potential.

During this project, which lasted about a year, we were facing a high number of daily queries that were very complex in nature (go figure; it was machine translation from German to English of OCRed chemical patents from 1830 to 1969). We ended up creating a restricted JIRA project where our client could access and cooperate with us in query resolutions.

Our productivity and clarity of operations improved substantially between the first iteration and the improved implementation of JIRA. In my experience the real force amplifiers were well thought out tasks; KPI tracking; and far superior risk management.

**JIRA at the service of a localization team in a large organization**

I joined HubSpot in September 2018 as a linguistic team manager. JIRA was already used by many teams in the company. Our localization team was already familiar with this tool, and my colleagues already went through a couple of JIRA implementations that were mostly focused at issue tracking and establishing critical production communication with our many internal stakeholders.

Towards the end of 2018, we were facing even more complex requests with increasing volumes. The need to revamp our JIRA implementation was becoming more and more apparent.

The main challenges that we were facing were the following:

1) Our workflows in JIRA were using only statuses and not taking advantage of transitions.
2) We had a lot of custom fields that needed to be reorganized, discontinued, and set properly, especially to avoid typos or invalid data.
3) We were not creating an environment that made it easy to extract important KPIs like word count, financials, project duration, and so on.

It’s important to have a solid workflow in JIRA that is supported by reality-checked transitions. According to Atlassian support documentation, “A status represents the state of an issue at a specific point in your workflow. When defining a status, you can optionally specify properties. A transition is a link between two statuses that enables an issue to move from one status to another. In order for an issue to move between two statuses, a transition must exist. A transition is a one-way link, so if an issue needs to move back and forth between two statuses, two transitions need to be created.”
Transitions enable automation in JIRA, like automated sub-tasks creation that takes into consideration data and parameters in the parent tasks; error messages that mitigate risk of mistakes; and customized notifications triggered by events or data to specific email aliases.

We had to further customize our JIRA implementation and the following process was the outcome:

1) **Form:** this constitutes the access point for our internal stakeholders.
2) **Task/request creation:** using the information in the form, a Task is created.
3) **Scoping:** one of our localization project managers will analyze the request and initiate Scoping.
4) **Data in:** while working in the JIRA Task and Sub-Tasks (at languages level), our team will update critical information like target languages, costs, word count, due dates, instructions, and documentation.
5) **Data out:** we constantly pull data out of JIRA and maintain dashboards that show budget spent vs. quarterly allocations, word count, average project duration, content types progress, and so on. Using JIRA custom fields, we also trigger workflows in other essential tools for localization pipelines.
6) **Execution:** the team members can independently execute their work knowing what to do while interacting directly with internal key stakeholders.
7) **Iterations:** we work in a fast-paced and change-driven environment. JIRA makes it easier to manage iterations.
8) **Delivery:** this is the final step, and we take advantage of Jira comments and notifications to notify requesters and stakeholders.

In two years, we have gone through two major upgrades of our JIRA implementation. The outcome has been that this powerful work management tool has become our source of truth when it comes to localization KPIs and transparent, readily-available progress tracking. This has been essential to facilitate the management of a multi-million-dollar annual localization budget and word count volumes of millions per quarter.

In my experience, the time and effort invested in data entry via JIRA or similar tools has always been worth it, especially when managing remote teams. This often small-time investment leads to a multiplier effect of positive outcomes, especially for those who are actually executing the tasks.

If you are curious to know from where the name “JIRA” comes from, I suggest visiting support.atlassian.com and searching for “What does JIRA mean?” It is quite a funny story that involves a very famous Japanese oversized monster. 😃
This reader is the fifth installment of MultiLingual’s encyclopedia-in-a-nutshell relating to localization standards. It does not purport to address everything around standards, and was last updated in late 2020. To be informative and of practical use, there must be a focus. This focus is being provided via two limitations. First, we are looking at standards that affect multilingual transformations of content, and at general content life cycle standards only, ad hoc and as far as they have bearing on multilingual transformations. Second, we are looking only at technical standards that are targeting actual technical interoperability — file or data formats and/or communication protocols. Abstract metadata, quality or service level standards are discussed only marginally, and only as long as they have bearing on real machine-to-machine interoperability in localization.

**BCP 14 (previously also known as RFC 2119, and now includes RFC 8174)**

**Short Description:** BCP stands for Best Current Practice. BCP 14 defines the standardization specific meaning of normative keywords such as MUST, MUST NOT, OPTIONAL, REQUIRED, RECOMMENDED and so on. RFC stands for request for comments, and these are numbered sequentially. RFC 2119 is the most common normative reference in other specifications throughout information technology standardization bodies. RFC 8174 was added to address the ambiguity of uppercase versus lowercase in RFC 2119 keywords. Localization-related standards such as ITS and XLIFF use BCP 14 keywords to make their normative statements that create the basis of conformance statements, testing and verification.

**Owner:** Internet Engineering Task Force (IETF), a nonmembership standardization body. Contributors are individuals who implicitly commit themselves by contributing without signing any formal contract. IETF creates internet-related technical standards, protocols, processes and nonnormative informational content. IETF is backed by the Internet Society.

**Intellectual Property Rights (IPR) Mode:** Reasonable and Non-Discriminatory (RAND), an IPR mode that allows owners charging for use of essential patents, provided that the charge is “reasonable” and “non-discriminatory.” This is a bit vague, but 2013 saw some groundbreaking legal development with regard to standards offered under RAND IPR policies. In a Washington District Court, Judge James L. Robart ruled in a RAND case between Microsoft
and Motorola that an initial licensing offer from Motorola had not been made in good faith and therefore constituted a breach of contract, in particular the obligation to license essential patents under RAND conditions to all implementers. The methodology of setting the rates in this case was so well explained and logically constructed that it is being used as precedent not only by other US courts but also outside of the United States and by policy makers.

**Current version and work in progress:** For a very long time (20 years), BCP 14 had only RFC 2119 as its content, so many standard makers treated BCP 14 and RFC 2119 as synonyms. With the addition of RFC 8174 (RFC 2119 Clarification) it became important to distinguish if a standardization work product references BCP 14 as a whole, or just RFC 2119. RFC 8174 makes it clear beyond any reasonable doubt that the IETF notion of normative keywords is typograph-driven. For example, “MUST” or any other of the defined keywords has its normative meaning only as far as it is printed in uppercase, as opposed to the ISO notion of normative language that is concept-driven and in fact forbids typographical distinction of normative keywords.

RFC 2119 was released for unlimited distribution in March 1997. RFC 8174 updates RFC 2119 and was released for unlimited distribution in May 2017.

However, it is undesirable to make changes to this BCP because so many normative texts across IETF, W3C, OASIS and so on depend on the meaning of the normative keywords as set out here. It is questionable whether adding RFC 8174 was a good thing, as it deepened the rift between the IETF and ISO notions of normative keywords. It is too early to tell what the impact of the clarification was. Before, if upperringing was omitted on a keyword in a spec, you could fix it in later publishing stages as it was considered editorial. In theory, it sounds clear to tell that only uppercased keywords are keywords, but in practice, it will make corrections difficult. Changing from plain to uppercase has been made into a material change by this supposed clarification. This clarification has some potential to cause interpretation damage. Until uppercasing was made mandatory, editors were dutifully aiming to avoid the keywords and use non-normative semantic variants. Now, many will stop doing that, citing RFC 8174, causing a lot of confusion in ordinary readers. It may well happen that this clarification will cause more changes to this BCP in the feature.

**BCP 47: Tags for Identifying Languages**

**Short Description:** BCP 47 is a normative IETF track that compiles recommendations on how to create a unique language tag from codes defined in several other normative sources, including ISO codes. It is frequently referenced by OASIS, Unicode and W3C standards.

**Owner:** IETF

**IPR Mode:** RAND

**Current version and work in progress:** RFC 5646 was released for unlimited distribution in September 2009. RFC 4647, Matching of Language Tags, was released for unlimited distribution in September 2006. BCP 47 is a persistent name that always points to the latest release, no matter what the current RFC number.

Additions to the BCP 47 component standard ISO 639-3 are under periodic review at its registration authority (RA), SIL International. The status of 2020 changes was determined and applied in December 2020. It is important to note that ISO 639-3 does not contain the actual list of the alpha-3 codes as the ISO 639-2 does. The reason is simple: ISO 639-2 only contains several hundred of language codes for languages with “large bodies of literature”, and is maintained by the Library of Congress. The actual code lists in the standard are only updated at the times of the standards revision, while ISO 639-3 covers almost 8,000 languages following the annually published Ethnologue reports by SIL. The actual alpha-3 codes for ISO 639-3 are continually published at SIL International’s publicly accessible pages at [https://iso639-3.sil.org/code_tables/639/data](https://iso639-3.sil.org/code_tables/639/data).

BCP 47 itself is stable, which is important for backward and forward compatibility. New component tags are being continuously registered via registration authorities specified in the standard. Other current developments are connected to Unicode extensions for BCP 47.

**BCP 47 Extension T: Transformed Content**

**Short Description:** Extension T is possible via the extensibility mechanism defined in BCP 47 (RFC 5646) itself. Extension T has normative status within the Unicode Consortium, as it is being maintained as part of CLDR, which is its major normative deliverable. This extension allows for additional tags specifying from which other language, locale or script the content at hand had been transformed. Extension T is not recommended for usage in structured environments such as XML, where this type of metadata can be specified using markup solutions rather than a single text field. Note that Extension T is appending the information about the originating language or locale with a leading “t,” which means that the BCP tag starts with the target locale and the source locale is appended. This makes sense given the structure of BCP 47 tags, but may be perceived as contrary to the customary listing order of source and target languages, so “EN-t-IT,” for example, actually means that the tagged content is English but was transformed from Italian, not the other way around.

**Owner/Maintainer:** IETF/Unicode Consortium.
Standards

January/February 2021

IPR Mode: RAND.

Current version and work in progress: Informational RFC 6497, published in February 2012. While RFC 6497 itself is stable, Extension T data and field definitions are regularly maintained as part of the CLDR release cycle.

BCP 47 Extension U: Unicode Locale Extension for BCP 47

Short Description: Extension U is possible via the extensibility mechanism defined in BCP 47 (RFC 5646) itself. It has normative status within the Unicode Consortium as it is being maintained as part of CLDR.

Owner/Maintainer: IETF/Unicode Consortium.

IPR Mode: RAND.

Current version and work in progress: Informational RFC 6067 was published in December 2010, and is maintained by the Unicode Consortium as part of CLDR.

Extension U is regularly maintained as part of the CLDR release cycle. The Unicode Localization Interoperability Technical Committee — known since early 2018 as the CLDR Technical Committee Subcommittee (CLDR TC/ULI SC) — delivers the input exception data for sentence-breaking mechanisms for different locales for the periodic CLDR releases. As a result, sentence-breaking behaviors driven by different exception data can be specified through assigned keys under the extension’s U mechanism.

Notably, the canonicalization algorithm included in RFC 6067 is slightly out of date. The current provision is specified in UTS #35. The canonicalization algorithm should be therefore updated in RFC 6067.

CLDR

Short Description: Unicode Common Locale Data Repository, http://cldr.unicode.org, is a standard repository

A Few Terms

bidirectional: a mixture of characters within a text where some are read from left to right and others from right to left. Bidirectional or bidi refers to an application that allows for this variance.

content management system (CMS): a system used to store and subsequently find and retrieve large amounts of data. CMSs were not originally designed to synchronize translation and localization of content, so many have been partnered with globalization management systems.

CHM: an extension for the Compiled HTML file format, most commonly used by Microsoft’s HTML-based help program.

Extensible Markup Language (XML): a programming language/specification pared down from SGML, an international standard for the publication and delivery of electronic information, designed especially for web documents.

intellectual property rights (IPR): rights relating to creations of the human intellect, primarily encompassing copyrights, patents and trademarks.

Organization for the Advancement of Structured Information Standards (OASIS): an IT standardization consortium based in the state of Massachusetts. It works on the development, convergence and adoption of open standards for a variety of areas. Its foundational sponsors include IBM and Microsoft. Localization buy-side, toolmakers and service providers are also well represented.

Simple Object Access Protocol (SOAP): a messaging protocol that allows programs that run on disparate operating systems (such as Windows and Linux) to communicate using Hypertext Transfer Protocol (HTTP) and its Extensible Markup Language (XML).

technical committee (TC): standardization bodies usually own, create, maintain and update technical standards through purpose-specific technical committees. In organizational structures such as OASIS, Unicode and ISO, they are called technical committees, while in others such as W3C they are not. They may also be referred to as an Industry Specification Group, Working Group, Special Interest Group and so on.

translation management system (TMS): sometimes also known as a globalization management system, a TMS automates localization workflow to reduce the time and money employed by manpower. It typically includes process management technology to automate the flow of work and linguistic technology to aid the translator.

Web Service Definition Language (WSDL): an XML format for describing network services as a set of endpoints operating on messages containing either document-oriented or procedure-oriented information.

World Wide Web Consortium (W3C): an international community that develops and owns many standards, including XML and HTML.

XML Schema Definition (XSD): a W3C recommendation that specifies how to formally describe the elements in an Extensible Markup Language (XML) document.
of internationalization building blocks, such as date, time and currency formats, sorting (collation) rules and so on. CLDR is not a standard in a classical sense. It is, as the name suggests, a repository that is being constantly updated and released on a rolling basis following its data release process.

**Owner:** Unicode Consortium.

**IPR Mode:** RAND.

**Current version and work in progress:** Version 38 was released on October 28, 2020.

A limited winter submission period for version 39 started in November 2020, and the full summer submission period for version 40 will start in May 2021. CLDR is being released on a regular semiannual schedule, whereas the cycle starting in the fourth quarter of each year is focused on tooling and bug fixing, and usually skips the public data submission phase.

### Common Translation Interface (COTI)

**Short Description:** The Association of German Manufacturers of Authoring and Content Management, shortened to DERCOM in German, created this standard.

DERCOM claims that it had to step up to protect and promote the interests of its membership and, at the same time, to let translation providers use the translation management systems (TMS) of their choice. DERCOM argues that no such interface exists, and it is true that a common interface is better than the present jungle of proprietary APIs on both ends that can only provide interoperability after some custom development on both ends for each new interface. Arguably this can be addressed by an Enterprise Service Bus or messaging architecture, but not every content management system (CMS) owner has the resources or technical muscle to do that.

COTI level 3 provides a sophisticated state machine that provides synchronous support for translation orders, cancellations and even updates of existing orders, which is notoriously tricky. While DERCOM is right that no translation web service interface has been standardized, it is surprising that COTI only standardized the business metadata interface and doesn’t say anything about payload standardization, standardization over a canonical data model such as XLIFF lends itself, but COTI actually does not mandate any data model restrictions with regard to the payload. So actually this specification only solves the CMS part of the equation. The bundle is thrown over the wall and it’s up to the language service provider to figure out what is localizable, what is not and so on.

While German CMS providers might be happy to implement always-synchronous Simple Object Access Protocol (SOAP) web services, the general trend seems to be rather toward RESTful interfaces and microservices, and it might hinder the adoption of COTI that level 3 strictly enforces synchronous SOAP calls. REST refers to Representational State Transfer, a widely known web services architectural best practice that has never been officially standardized. Unlike SOAP, which is a specific standardized protocol, there is no single REST standard or even a single best practice. REST is rather an architectural style for interoperability.

Level 1 is actually the specification of the payload ZIP package, its structure and manifest. Sound familiar? People just keep reinventing this very wheel, albeit each time with subtle differences that don’t allow for automation among those “standard” ZIP packages. Electronic Dossier, Linport, TIPP and now COTI level 1 — all of these just defined a folder structure that they think a translation request and response should contain without talking to each other. Alan Melby and others invested some diplomatic effort into making the Electronic Dossier people talk to the Linport people and merge the Linport project with TIPP. But even before this could bear fruit, DERCOM brought another “standard” folder structure.

COTI level 2 mandates automated exchange of the structured ZIP packages over hot (watched) folders. Whatever is placed in a hot folder is up for grabs at the receiving end. The TMS watches the CMS’s hot folder for new projects and the CMS watches the TMS’s hot folder for responses. Sound like the 1990s?

The interesting contribution to the topic of standardized translation web services is in the level 3 compliance that requires SOAP and doesn’t allow or specify any other admissible bindings, which would certainly be advisable.

**Owner:** DERCOM. DERCOM is not a standardization body but a German trade association. They nevertheless decided to provide German-producer-centric standardization, and, as they say, “non-proprietary” solution to CMS and TMS interoperability.

**IPR Mode:** Unclear. The specification, documentation (CHM) and validation artifacts (WSDL, XSD) are available for free, but no restrictions seem to have been specified with regard to licensing of IPR essential to implementing the specification. The specification says to look up “the intellectual property rights section of the technical committee web” at www.dercos.de but there seems to be no such section. This unfortunately means that not even (F) RAND restrictions apply for licensing offers made for essential IPR by their respective owners. This is at least until DERCOM actually comes out with publicly visible IPR information.

**Current version and work in progress:** COTI version 1.1.1 was put out July 07, 2017, as a minor update from version 1.1 put out April 29, 2016. The 1.0 version was first approved on May 25, 2014, and was subject to only minor editorial fixes by September 2015. Version 1.1 brought
some additional material such as the XML Schema, but also changes in the WSDL artifacts.

There doesn’t seem to be a specific plan to further develop the specification. Rather it seems that DERCOM makes changes and fixes as its membership gradually progresses in fulfilling its obligation to implement COTI level 1 through COTI level 3 compliance. The WSDL changes in the version 1.1 were most probably driven by implementation experience of the DERCOM membership.

**International Components for Unicode (ICU)**

**Short Description:** Until May 2016, ICU was an IBM-driven open source project, and in fact the most important reference implementation of both Unicode and CLDR. As ICU provides internationalization libraries, it actually consists of two subprojects, ICU4C and ICU4J, that provide libraries for C and C++ and Java respectively. Spectacularly, ICU is the common denominator of both Android and iPhone operating systems. ICU 58 is the first ICU version released under the ICU TC governance.

The ICU project implicitly defines a “message format” (See Message Format). Since the ICU message format has been implemented in number of internationalization libraries and usage of those messages has been known to cause interoperability issues on the boundary with localization, a new Message Format Working Group (MFWG) has been formed under the Unicode CLDR TC to work on the message format successor.

**Owner:** Unicode Consortium.

**IPR Mode:** RAND.

**Current version and work in progress:** ICU 67 for Unicode 13.0 and CLDR 37 is the current version. ICU4J 67.1 released on April 22 2020. The URL https://unicode-org.github.io/icu-docs/apidoc/released/icu4j/ always points to the current official version. The legacy URL now redirects to the common cover page for both ICU4J and ICU4C.

ICU (both the Java and C projects) is being developed continually to address bug fixes and dependencies. Major releases are driven by the releases of the Unicode Standard and changes in CLDR data. New major releases came have been released in late 2020 just after stabilization of CLDR 38 at the end of October. Since all ICU infra migrated from SVN to GitHub back in summer 2018, the project transitioned their spec reference mechanism to unicode-org.github.io — GitHub enabled and hosted documentation webpages.

**Internationalization Tag Set (ITS)**

**Short Description:** ITS 2.0 comprises 19 metadata categories compared to a mere seven in ITS 1.0. Additionally, ITS 1.0 metadata categories were primarily designed for internationalization of XML content. Nevertheless, as abstract data categories, ITS can be implemented in non-XML environments. Importantly, ITS 2.0 normatively specifies usage of the old and new ITS data categories for XML and HTML 5 content, and new categories have been introduced that explicitly address the localization roundtrip, such as the localization quality assurance related data categories Localization Quality Rating and Localization Quality Issue. Importantly, ITS 2.0 is listed in the JTC 1 Big Data Standards Roadmap (ISO/IEC TR 20547-5) as a key automation enabler for big data and analytics dealing with human language.

Owners of ITS decorated content want their internationalization and localization related metadata to inform the roundtrip and make it to the target content in ameaningfully processed state that allows for drilling down into the process and for reconstructing the audit trail. Localization workflow managers should pay attention to information flows directed by the ITS data categories introduced by their customers up in the tool chain. There is also potential to introduce automated or semi-automated ITS decoration steps before extraction, or to introduce relevant XLIFF mappings of ITS data categories recorded on translations during the localization roundtrip that can be imported back into ITS in XML or HTML on merging back localized content. Categories such as Translate (which had become a native HTML 5 attribute), Elements Within Text, Locale Filter, Target Pointer or External Resource should drive extraction and merging back of localizable content. Terminology and Text Analysis disambiguation markup should be passed on to human and machine translators. Proper interpretation of directionality markup is a must for sound handling of bidirectional content using Arabic or Hebrew scripts. Self-reported machine translation (MT) confidence should be passed on to the content recipients, possibly along with quality assurance related metadata (Localization Quality Issue and Localization Quality Rating).

These considerations are especially valid when hooking up existing localization workflows upward into the tool chain. Existing workflows should introduce mappings of ITS data categories used in source content, so that the metadata flow is not broken throughout the content life cycle.

The current ITS 2.0 categories are Translate (flag indicating translatability or not); Localization Note (for alerts, hints, instructions); Terminology (to identify terms and nonterms and optionally provide definitions); Directionality (manages left to right/right to left display behaviors of content portions); Language Information (BCP47 language tags on relevant content portions); Elements Within Text (shows which elements break flow
Current collaborations and liaisons were set up to promote the standardization process of the ITS 2.0. The World Wide Web Consortium (W3C) ITS Interest Group (IG) was chartered to specify a real-time interoperability mechanism to make sure that translations fit restricted database fields, forms and so on.

**Owner/Maintainer:** Multilingual Web/LT Working Group/ITS Interest Group (IG). The World Wide Web Consortium (W3C) ITS IG has been the informal main- tainer of ITS 2.0 after the MultilingualWeb-LT Working Group mandate expired; however, interest groups in W3C cannot create normative deliverables. Therefore a new Working Group will have to be formed to commence work on the successor standard, as soon as the ITS IG identifies industry need for a successor version.

However, the status of the ITS IG is unclear, as its charter expired on December 31, 2018, and it is not clear at the moment if the charter can be extended. In case ITS IG lapses, stakeholders of ITS should initiate a W3C community group to look after the adoption and maintenance of ITS 2.0. The main purpose of such a group is to see if there's a need and momentum to create a new feature or maintenance release of the Internationalization-Tag Set.

**IPR Mode:** Royalty Free (RF), an IPR mode that mandates and guarantees royalty free use of essential patents in order to implement a standard.

**Current version and work in progress:** The current version is 2.0, published as a full W3C Recommendation on October 29, 2013. Work on another major version is not imminent. That said, ITS 2.0 covered a number of new areas with nonnormative mappings. These included Resource Description Framework (RDF) and XLIFF mappings. See XLIFF for inclusion of the ITS functionality as an XLIFF 2.1 module. Current collaborations and liaisons include OASIS DITA, DockBook, XLIFF and XLIFF OMOS TCs, as well as GALA TAPICC.

ITS IG is working on further elaborating and maintaining other informative mappings, such as NIF and MQM. However, the MQM mapping aspect transitioned to the MQM W3C Community Group. ITS IG also maintains ITS extensions that support categories that could not be normatively specified for various reasons within the ITS 2.0 publishing schedule; for instance, the Readiness data category.

**JSON Localization Interchange Fragment Format (JLIFF)**

**Short Description:** JLIFF aims to be compliant with the abstract object model defined in UML diagrams and a prose specification as XLIFF OM. The TC currently works on XLIFF unit JSON representation examples, JSON schema and JSON templates to represent the XLIFF OM UML diagrams and XLIFF unit XML examples.

**Owner:** OASIS XLIFF Object Model and Other Serializations (OMOS) TC. JLIFF is currently available as a JSON Schema that is reasonably stable but not yet officially published by OASIS.

JLIFF's main intended use case is the real time interoperability of TMS and CAT tools that support the XLIFF 2 data model. Importantly and unlike XLIFF, JLIFF can support exchange of fragments at unit, group and file levels. For full XLIFF interoperability, it also supports exchange of whole XLIFF Document JLIFF equivalents, but this is not the main use case. It is important to stress that in order to preserve data integrity the lowest interchange level is that of logically separate bitext units and not arbitrary segments.

**IPR Mode:** Non-Assertion (RF).

**Current work in progress:** In summer 2018, JLIFF 2.0 and 2.1 schemas were completed by the committee on the XLIFF OMOS TC GitHub repository. The TC started working on a prose specification to be reviewed hopefully in early 2021. JLIFF schema has been attracting implementers even before becoming stable. Notably, Vistatec released an open source implementation of core JLIFF in March 2018 at https://github.com/vistatec/JliffGraphTools. Reportedly, TDC has implemented a RESTful API that is capable of three-way transposition among JLIFF, XLIFF 2, and XLIFF 1.2. It seems that the capability of JLIFF to create fragments compatible with the XLIFF 2 data model is exactly what implementers need for real time API-based interoperability.

This is also the assumption behind the GALA TAPICC Track 2, which was chartered to specify a real time translation API (JRARTEBU – JLIFF REST API for Real Time Exchange of Bitext Units).
The XLIFF OMOS TC was chartered in December 2015 and started to develop a JSON serialization in parallel to the creation of the abstract object model for the XLIFF 2 family of standards that it is developing as its first priority. The objective is to create such a JSON serialization of XLIFF and XLIFF fragment data categories that would allow lossless interchange between JLIFF and XLIFF based tool stacks. The development of the JLIFF specification and artifacts is being conducted on this GitHub repository: https://github.com/oasis-tcs/xliiff-omos-jliff. The repository is public but the contributors need to be or become OASIS XLIFF OMOS TC members. However, anyone can raise issues or comments via the associated issues or wiki at https://github.com/oasis-tcs/xliiff-omos-jliff/wiki, as well as minor bug fixes via pull requests. It’s up to the repository maintainers if those pull requests will be merged or not, which is the usual GitHub collaboration workflow.

**Message Format (MF)**

**Short Description:** ICU message format has been in use for 20 years through ICU APIs (both in ICU4C and ICU4J). It is at the core of many internationalization libraries across operating systems and development frameworks. A couple of chief issues of the current format were identified as part of the initial requirements analysis: The design is not modular, doesn’t allow for deprecation of obsolete or addition of new required features. Capability to use internal selectors within messages often leads to translatability issues, typically in cases where the target language is more morphologically complex than the source language and hence the target language requires more message variants than the source message postulates. This causes situations where the linguist is not able to produce all the variants required in the target language or structure their translation, so that it could remain grammatical in the target language in all cases after the variables and selectors had been resolved in application runtime.

**Owner:** Unicode Consortium.

**IPR Mode:** RAND.

Current version and work in progress: ICU MessageFormat, implicitly defined as part of ICU (See ICU). Reacting to issues associated with ICU MessageFormat and to demand expressed by ECMA TC39 (JavaScript), in late 2019, Unicode Consortium formed a Working Group under its CLDR TC (Unicode CLDR TC/MFWG) to work on a Message Format successor, informally known as Message Format 2.0. The group was chartered to produce a Unicode Technical Standard (See UTS) that will describe the new format. The group further refined its goals within the mandate it received from the Unicode Consortium and ECMA to produce the format, related APIs, and importantly a mapping to and from XLIFF 2 to enable an interoperable localization roundtrip of those messages.

This work is still in the inception phase, as the group was chartered and has agreed on its goals and also stated potential objectives that are considered out of its scope (no goals). The design principles discussion is currently in full swing and will probably continue for some time as first proof of concept efforts are being explored by group members to validate the slowly growing set of design principles. The new message format group set the goal to be translation roundtrip friendly and to define an XLIFF 2 mapping for localization interoperability. Importantly, this group’s effort is driven by internationalization engineers and developers across major buyer organizations such as Amazon, Apple, CaixaBank, Dropbox, Expedia, Facebook, Google, Mozilla, Oracle, and PayPal.

**Multidimensional Quality Metrics (MQM)**

**Short Description:** MQM is primarily a set of error data categories that can be used in language quality assessment, especially in bilingual translation scenarios, but also in monolingual review or authoring scenarios. The aim of MQM is to cover the whole logical space of possible errors and nonconformities related to language and locale-specific presentation of content of any kind.

It is impossible (and even if possible not likely very useful) to use all the error data categories from all levels and sub-levels. Profiling is a key method defined by MQM, so that in any particular project or job a particular profile or subset of relevant error data categories has to be defined based on stakeholder requirements and expectations. Thus MQM concentrates on quality in the sense of fitness for purpose. It specifically rejects the notion of some general or abstract quality without defining the purpose, related requirements and a resulting relevant set of possible error types. Also notable is that MQM doesn’t define any severity levels. There is an option for implementers to predefined severity levels and scoring methods based on any particular MQM subset.

Currently MQM is a proper superset of TQF. MQM and TQF had been developed separately by the QT21 EU project (coordinated by DFKI) and TAUS respectively. The EC funders effectively forced these two projects to reconcile their data models.

For effective exchange of MQM/TQF metadata within projects and jobs, inline capturing and encoding of the recorded errors is critical. This needs to be implemented as the ITS localization quality issue data category in native formats or via the same data category implemented as an XLIFF Version 2.1 module in bitext.
Looking at the above described developments it seems that MQM continues to diverge from the Localization Quality Issue Types in W3C ITS (See ITS) and hence doesn’t make progress in making the MQM data and metadata practically usable inline within source text, bitext, and target text.

### Open Lexicon Interchange Format (OLIF)

**Short Description:** OLIF is a stable and relatively widely used lexicon interchange format. It has a rich metadata structure and allows for the exchange of complex lexicon entries for various purposes, such as terminology management and MT. OLIF had been designed for use in both monolingual and multilingual context via cross-linking of “mono” elements.

**Owner:** OLIF Consortium, an ad hoc industry consortium driven by SAP and set up in 2000.

**IPR Mode:** Unclear. The specifications and schemas are available for free, but no IPR mode seems to have been specified.

**Current version and work in progress:** Version 2.1 is current. Version 3 has been in beta since 2008; no current work seems to be under way.

### Segmentation Rules eXchange (SRX)

**Short Description:** SRX is an XML vocabulary that facilitates the exchange of segmentation rules between TMX compliant systems. SRX’s relationship to Unicode is not a transparent one, and SRX can be considered incomplete from the engineering point of view. However, its proclaimed goal was not to provide a set of segmentation rules for a number of languages, but rather to provide a mechanism to exchange the rules to improve TMX interoperability. TMX often fails to guarantee its targeted lossless transfer of TM data due to segmentation differences, chief among other issues. The current SRX incarnation works on a closed world assumption, meaning it recreates (and adapts) UAX #29 rules. UAX #29 is referenced and its study encouraged, but the relationship is currently not a maintainable linkage.

According to former Unicode Localization Interoperability Technical Committee (ULI TC) chair Helena Chapman, SRX developers should not take UAX #29 (and its ICU implementation) for granted, and use SRX only for the exchange of rules that differ from the standard UAX #29 behavior. ULI TC has been collecting natural language exceptions to UAX #29 for several major languages and included these in the CLDR release cycle. CLDR, however, uses LDML as its description language, not SRX.

**Owner:** ETSI ISG LIS. ISO/TC 37.SC 4 was intended as a co-owner of SRX. However, no memorandum of understanding between the groups was signed. This work item expired and was deleted on July 22, 2013, according to the ISO standards publishing policy.

Unicode ULI TC (now CLDR TC/UL SC) agreed to host the current SRX version after ETSI ISG LIS disbanded as a publicly available specification. However, it is unclear if
ULI TC would be legally allowed to produce a new version without formally negotiating IP transfer from ETSI.

**IPR Mode:** RF for the LISA version; FRAND in ETSI (but wasn’t republished).

**Current version and work in progress:** A copublication attempt was made with ISO TC37/SC 3 as ISO CD 24621, which successfully passed the committee draft ballot, but no work progressed and the project was deleted on July 22, 2013.

There is no current work in progress due to ETSI ISG LIS dissolution. Version 2.0 was released by LISA OSCAR on April 7, 2008. Based on the published executive summary from November 2011, ETSI ISG LIS scheduled an SRX meeting for March 2012. This meeting happened behind closed doors. There are or should be dependencies with ULI TC on UAX #29 segmentation behavior modifications. SRX has potential as an XML based exchange vehicle for segmentation rules, because there is no ultimate finite solution to the segmentation issue in natural languages. In 2018, ULI TC started a note on segmentation and wordcounting that would build on the principles of UAX #29. The development of this note was abandoned in 2019 due to lack of interest.

**Simple Object Access Protocol (SOAP)**

**Short Description:** SOAP is an XML-based web services protocol. Version 1.0 was submitted to IETF in 1999 as an internet draft but never reached an RFC status, so it actually hadn’t become a standard at IETF. SOAP 1.1 had Note status at W3C and the only SOAP version that ever reached the standard status (W3C Recommendation) is SOAP 1.2.

The SOAP protocol consists of five layers: message format, transfer protocol bindings, message processing models, message exchange patterns and extensibility. Unlike REST, which is strictly HTTP-based, SOAP is protocol-neutral and can work over several lower level protocols such as HTTP (Hypertext Transfer Protocol), SMTP (Simple Mail Transfer Protocol), TCP (Transmission Control Protocol), UDP (User Datagram Protocol) or JMS (Java Message Service).

**Owner/ Maintainer:** W3C XML Protocol Working Group (closed) / no current maintainer. The SOAP 1.2 multipart recommendation was produced by the XML Protocol Working Group at W3C, and the group was closed on July 10, 2009.

**IPR Mode:** RF.

**Current version and work in progress:** 1.2 (Second Edition) was published as a W3C Recommendation on April 27, 2017. The protocol as such is not being developed or maintained at W3C or elsewhere. We list SOAP because it was resurrected in our industry by the COTI level 3 conformance requirement to implement a SOAP-based web service automation.

**TBX, also known as ISO 30042**

**Short Description:** TBX is a family of XML-based terminology markup languages that should allow for lossless exchange of terminology-related data and metadata. So far, two more lightweight versions known as TBX Basic (published in 2008 by LISA Terminology Special Interest Group) and TBX-Min (published in 2013 by LTAC Global) have been developed. TBX-Min should be targeting the use case of exchanging terminology with translators in the form of simple glossaries mappable onto UTX. However, TBX Basic is more suitable for mapping between TBX and XLIFF 2.0 with glossary modules.

TBX has been criticized for industry disconnect, for being too heavy on one hand and being too restrictive and not very suitable for MT training on the other. One of the reasons might be that TBX is supposed to be both a representation and exchange format for terminology, but it has been struggling to define a minimum set of terminology metadata suitable for practical interchange in a localisation context. Some industry implementations can hardly be considered in the spirit of the standard, as they have not enforced inclusion of even very basic metadata such as part-of-speech or they are not structurally compliant with any of the predefined data structures.

**Owner:** ISO TC 37/SC3. LTAC for all public and some private dialects; LISA (defunct as of 2011) and ETSI for legacy versions published by 2008.

**IPR Mode:** RAND within ISO. BSD 3 clause licensed an open source project at LTAC/TerminOrgs.

**Current version and work in progress:** ISO 30042:2019 (informally TBX 3) is the current version. TBX 3, ISO 30042:2019, is only the second ISO edition. Because it is the second edition at ISO, its core namespace is and must be urn:iso:std:iso:30042:ed-2. Despite that it’s the third major version of TBX considering the developments in LISA OSCAR before the ISO co-publishing agreement.

The TBX Steering Committee has agreed to focus promotion efforts on encouraging CAT tool vendors to implement a feature that allows import of files that comply with the version 3 (2019) TBX-Basic dialect, which can be found at https://ltac-global.github.io/TBX-Basic_dialect/ (always current).

Focus on just this one dialect should promote real interoperability. LTAC and the BYU TRG have been working together on a “Steamroller” application that will manipulate version 3 TBX files from dialects bigger than Basic into compliance with the Basic dialect. LTAC is also working with IATE to provide a version 3 TBX-Basic export feature. While there is a number of private dialects...
that are “bigger” than TBX-Basic, TBX-Basic should provide enough power in terms of external interoperability. On the other hand, TBX-Basic is bigger and more powerful than the other two public dialects, TBX-Core and TBX-Min. Serious L10n users (mature buyers and service providers) are well advised not to use dialects smaller than TBX-Basic.

The most important data model change that happened in ISO 30042:2019 was making the inline data model compliant with the XLIFF 2 inline data model. This was implemented by the LTAC/Terminorgs TBX Steering Committee and TC 37/SC 3 with input from OASIS XLIFF and XLIFF OMOS TCs. The modular public dialects (TBX-Core, TBX-Min, and TBX-Basic) compatible with the second edition of ISO 30042:2019 were released during 2018, as the CD and DIS version of the second ISO edition stabilized towards FDIS (no more material changes possible after FDIS, technical compatibility with the future standard was guaranteed after summer 2018), also with the new inline data model.

The second ISO edition of TBX accumulated quite a number of breaking changes. But most of those changes are dealing with modernization of the XML tooling. Apart from the inline data model change and related introduction of explicit directionality support, it is important to stress that TBX joined other modern-day standards in leaving its former monolithic design for a new modular design.

This edition of TBX specifies a nonnegotiable core and prescribes that all compliant dialects must include that core. The standard then specifies a modularity mechanism. Outside of ISO, LTAC made sure that the most common dialects are extended from the core using a telescoping principle. The simplest possible dialect is TBX-Core; the TBX-Min is composed from the Core and Min modules. TBX-Basic is TBX-Min plus the Basic module, and TBX-Linguist (considered a private dialect to allow flexibility for academic users) is TBX-Basic plus the Linguist module. The stakeholders have high hopes that this will vastly improve the so-called “blind” or plug and play interoperability. The second ISO TBX edition also defines TBX agents and provides a compliance clause targeting the document compliance, as well as specific agents’ compliance. Notably, this edition introduces namespace-based modules, albeit only in the DCT (data category as tag name) style, which is a compromise that should not disturb those who are not worried with the DCA (data category as attribute name) style. And vice versa — implementers who are afraid of implementing namespaces can stick to the old DCA style. Helpfully, both DCA and DCT are extended from the same core, and modules specified in either DCA or DCT are semantically equivalent (based on the same sets of additional data categories). DCT is easier to validate and it’s also easier to filter out unsupported modules (as those are in different namespaces) in the DCT style.

GALA and ttt.org still host the latest LISA OSCAR version including the legacy TBX-Basic. TBX-Basic 3.1 (the version maintained by LTAC/TerminOrgs, last updated Sep 12 2014) is the latest version of TBX-Basic compatible with ISO 30042:2008.

Although ISO 30042:2008 (informally TBX 2) was superseded by ISO 30042:2019 (informally TBX 3), this TBX 2 based Basic dialect version is still widely used. It is somewhat confusing that this legacy version of the Basic dialect is versioned as 3.1, although it is not forwards compatible with the current TBX 3 version, which is ISO 30042:2019.

ISO 30042 will undergo another systematic review at ISO by 2024, but no plans regarding that have been formulated yet. As ISO 30042:2019 is fairly fresh, most of the current work concentrates on evangelizing the advantages of the new modernized and modular design to the implementers. ISO TC 37/SC 3 launched a couple of companion projects to ISO 30042 to provide additional information and guidance. These projects are ISO WD TR 24633 Building an RNG schema for TBX Core and ISO WD TS 24634 Management of terminology resources — Representation of concept relations and subject fields in TBX.

Work in liaison with the OASIS XLIFF TC and later with OASIS XLIFF OMOS TC resulted in development and testing of a TBX-Basic to XLIFF 2 with glossary module mapping. The mapping upgraded to the modular second ISO edition based on TBX-Basic — semantically equivalent with the old TBX-Basic — is now being specified on the standards track within OASIS XLIFF OMOS TC. See the latest editor’s draft at https://tools.oasis-open.org/version-control/browse/wsvn/xliff-omos/trunk/XLIFF-TBX/xliff-tbx-v1.0.pdf.

Translation API Cases and Classes (TAPICC)

Short Description: The first TAPICC symposium, on October 26, 2016, brought a four-way categorization of use cases that were considered in scope of the Translation API (TAPI) standardization: 1) Exchange of a payload blackbox accompanied with rich business or project metadata. In this area, TAPICC wants to abstract common business metadata used in other initiatives past and current including Linport, TIPP, COTI, XLIFF 1.2 project group, OASIS Translation Services TC (closed) and so on. 2) Bidirectional real-time exchange of XLIFF unit data in arbitrary serializations. 3) Enriching of XLIFF units in arbitrary serializations with terminology, translation suggestions, text analysis, process, quality assurance metadata and so on. 4) Exchange of data for layout representation purposes during the bitext management process.
RESTful API that will work with data models specified by the
ies and services. T1/WG4 was chartered to design an actual
and consuming XLIFF 2 including documenting public librar
encoded metadata and the business level imposed metadata.
payload will be allowed to be exchanged via the TAPICC
surface (XLIFF 2 for T1/WG1 and JLIFF for T2/WG1), and
ics (XLIFF EMBP) are in version 1.0. https://galaglobal.github.io/TAPICC/T1/
Version 1.0 of the XLIFF EBMP was contributed by
GALA to the OASIS XLIFF TC for publication as a TC note.
Several other deliverables (including the asynchronous
API bringing together inputs from all T1 TAPICC groups)
have not yet reached stability but engaged in a number of
public consultations via GALA webinars and so on. Further
progress was stalled by GALA not being able to organize the
planned spring 2020 TAPICC face to face meeting as GALA
2020 pre-conference in San Diego, California.

TAPICC is in a staged development phase, where the
Track 1, chartered on February 8, 2017, has four active working-
groups. T1/WG1 was chartered to develop a consensual
set of business metadata that will enable efficient payload
exchange. T1/WG 2 was chartered to specify what kind of
payload will be allowed to be exchanged via the TAPICC
API, and how to resolve potential conflicts between payload
encoded metadata and the business level imposed metadata.
T1/WG3 was chartered to work on best practices of creating
and consuming XLIFF 2 including documenting public libraries
and services. T1/WG4 was chartered to design an actual
RESTful API that will work with data models specified by the
other T1 working groups, most importantly the T1/WG1.

The joint purpose of the T1 groups is to enable asyn-
chronous API exchange within the complex localization
supply chain, while relying on an XLIFF 2 based canonical
data model.

In October and November 2018, TAPICC chartered
and launched a call for volunteers for T2/WG1 — widely
publicized in tcworld 2018 in Stuttgart on November 14,
2018. T2/WG1 - JLIFF REST API for Real Time Exchange
of Bitext Units (JRARTBU) has currently been assigned
one Work Package: WP1 Unit Exchange. Batch or buffered
exchange is out of scope of WP1, although not out of scope
of Track 2 or T2/WG1. SOAP and any other protocols or
bindings are out of scope of T2/WG1, although not out of
scope for TAPICC Track 2. TAPICC Track 2 is going to
reuse the OASIS XLIFF OMOS TC specified JLIFF format,
the JSON serialization of the XLIFF 2 based object model.
This synchronicity of data models will ensure semantic
and behavioral interoperability between the asynchronous
T1 and real-time transactional T2 bitext interchange.

Translation Memory eXchange (TMX)

Short Description: TMX has been arguably the most
important and most widely implemented localization
standard format. TMX is a simple XML vocabulary that
was designed to provide lossless translation memory (TM)
exchange. However, several obstacles prevented TMX
from reaching the set goal. Level 1 implementations are
too low a common denominator to actually secure loss-
less interoperability, because of segmentation differences
(that should in theory be addressed by SRX) and because
of absence of inline markup on Level 1. Level 2 stipulates
lossless exchange of native inline codes that are however
ignored by many tools and encoded as abstract placeholders.
TMX is now far behind industry developments, but
it will continue to be important for some time as a legacy
format, mainly for collecting MT training corpora from
legacy tools and repositories.

Owner: ETSI ISG LIS.
IPR Mode: RF in the LISA published versions, and
FRAND in the ETSI ISG LIS version.

Current version and work in progress: There is no
current work in progress, as the ETSI ISG LIS group was
closed. Unfortunately, OASIS and ETSI were not able to
agree on transferring the TMX IP to the OASIS XLIFF
OMOS TC that had been chartered to take over the own-
ership of TMX.

1.4b is the current version with LISA numbering. It
is also referred to as ETSI GS LIS 002 V1.4.2 (2013-02).
Based on the published executive summary from Novem-
ber 2011, ETSI ISG LIS wished to coordinate TMX 2.0
development with XLIFF 2.0 definition of inline markup
codes. During the existence of ETSI ISG LIS there was
no technical development on the TMX front. The group only republished the latest LISA version on the ETSI Group Specification template. At FEISGILT in June 2013 in London the ETSI ISG LIS chair made a public consultation of a possible related work item that ETSI ISG LIS would like to undertake the standardization of fuzzy matching calculations for text segments. At the time of ETSI closing ISG LIS (September 2015), there were informal talks about OASIS XLIFF OMOS TC taking over the TMX ownership, maintenance and development. The XLIFF OMOS TC was chartered and scoped to be able to take over TMX ownership. However, formal IP transfer between ETSI and OASIS has not been negotiated to the present day. Because of the legal matters being stuck, delegates of the June 2016 FEISGILTT and XLIFF Symposium discussed developing an XLIFF 2 profile with mandatory usage of the Translation Candidates module as a maintainable replacement of the TMX functionality.

In 2017, Andrzej Zydroń proposed developing a note within OASIS XLIFF TC that would describe how to use an XLIFF 2 profile instead of the obsolete and unmaintained TMX format. However, this work hasn’t progressed due to lack of interest and technical consensus.

Unicode Bidirectional Algorithm (UAX #9)

**Short Description:** Default text flow of Arabic and Hebrew scripts is right to left. However, text written in these scripts often contains portions with left-to-right directionality, such as names of companies or products. That is why such text is called bidirectional (bidi). Many characters have strong directionality properties, but there are also characters with weak directionality behavior and neutral characters whose directionality depends on context. In practice, normally invisible control characters (markers) need to be used in order to encode bidi in plain text. Simply put, UAX #9 is a detailed normative account of Unicode bidi behavior (mainly) in plain text.

In theory, the characters that the Unicode Bidirectional Algorithm makes use of to explicitly set text flow direction should not be used within markup context. Instead, the bidi flow control characters should be replaced with appropriate markup controlling the text flow. In practice, many tools ignore directionality markup and apply UAX #9 in full (including the control characters) even in structured and markup environments. This may be due to the fact that UAX #9 has a long tradition (since Unicode 2 in 1996). There is also a standardization gap, as Unicode control and stateful characters are not provided with clear processing requirements to be applied on entering markup environments. UTR #20 unfortunately provides only an abstract guidance rather than clear and unambiguous processing requirements. While XLIFF 2.0 has its own directionality attributes, it does not have attributes corresponding to inline bidi overrides or embeddings, so these are allowed as UAX #9 control characters if needed.

**Owner:** Unicode Consortium.

**IPR Mode:** RAND.

**Current version and work in progress:** Revision 42 was released for Unicode 13.0 on February 2, 2020. UAX #9 is being constantly revised to be up to date with the current Unicode release, and www.unicode.org/reports/tr9 links to the current official version. Additionally, the link www.unicode.org/reports/tr9/proposed.html always points to the latest proposed version if such a proposal exists. Major versions happened between Revision 27 and Revision 29. Revision 29 sent ripples that profoundly influenced handling of bidi text in the world of markup languages. It led to introduction of new direction handling elements and attributes in HTML and XML vocabularies. Because directionality handling in HTML and several XML vocabularies was in flux in 2013, ITS 2.0 does not contain normative provisions for directionality handling. XLIFF 2 does contain up-to-date directionality markup and a valid guidance how to combine it with directionality control characters if necessary when modifying segmentation in compliance with the Revision 35 of UAX #9.

Unicode in XML and other Markup Languages (UTR #20) [Withdrawn]

**Short Description:** Unicode Technical Reports (UTR) are persistent names that always point to the actual revision. Unicode, as its main target is plain text, contains many control, formatting and other characters. This document gives a normative overview and general guidelines of which characters should and should not be used in markup context. In general, any Unicode character that is XML illegal or would require additional metadata for interpretation should come with a markup handling/replacement recommendation, or processing requirement. Authoring tools, XML editors and browsers are generally encouraged to ignore inappropriate or deprecated Unicode characters, so their preservation on crossing of plain text/markup boundary will often lead to harmful loss of data or metadata. In general, plain text is linear and requires special control characters or specific application behavior to encode metadata and/or styling information that can be handled with structured markup in XML or HTML environments.

**Owner:** W3C (Internationalization Core Working Group). UTR #20 was withdrawn at Unicode in March 2016, and maintenance of the document transferred to W3C.

**IPR Mode:** W3C (RF).

**Current version and work in progress:** Now withdrawn both at Unicode Consortium and W3C. Last released as a
W3C Working Group Note on July 13 2017. Albeit dated, this is historically an influential document.

No current work is in progress, although W3C accepts and collects Github issues that are being collected against the latest editor’s draft http://w3c.github.io/unicode-xml/, currently dated July 8, 2017, and identical with the last published version. This important note would deserve a refresher for rapid new Unicode releases.

**Unicode Locale Data Markup Language (LDML or UTS #35)**

**Short Description:** This specifies an XML vocabulary for encoding locale specific generic data categories — dates, amounts, decimals, units of measure, currency symbols and so on. Its main purpose is to enable the creation and maintenance of CLDR but is also used directly in programming frameworks such as .NET.

**Owner:** Unicode Consortium.

**IPR Mode:** RAND.

**Current version and work in progress:** Version 37, Revision 59, was released on April 15, 2020. Version 38, Revision 60 is proposed here: https://unicode-org.github.io/cldr/ldml/tr35.html.

**Unicode Regular Expressions (UTS #18)**

**Short Description:** Unicode Technical Standards (UTS) are persistent names that always point to the actual revision. UTS #18 gives general guidelines for regular expression engines on how to comply with the Unicode Standard. Three levels are specified, of which two are default (one the minimum feasible for programmers, the other more end-user friendly) and the finest is language specific. Algorithms conformant to this specification will give different results for different versions of Unicode

**Owner:** Unicode Consortium.

**IPR Mode:** RAND.

**Current version and work in progress:** Version 21 was released on June 17, 2020. No new revisions have been proposed for at the time of writing.

The URL www.unicode.org/reports/tr18 always points to the current official version. Once a new draft is proposed, www.unicode.org/reports/tr18/proposed.html will link to it.

**Unicode Text Segmentation (UAX #29)**

**Short Description:** Unicode Standard Annexes (UAX) are persistent names that always point to the actual revision number. UAX #29 is the key normative source of segmentation rules. Apart from sentence boundaries, which are most relevant for computer-aided translation tools interoperability, it defines more basic grapheme cluster and word boundaries. The segmentation rules are given in more or less natural language as an inductive succession of rules. The specification states itself that the same set of rules can be given using regular expressions. Unfortunately, no finite set of regular expression-based rules can ensure 100% successful sentence segmentation of English text, the main reason being the semantic ambiguity of the full stop. Apart from closing sentences, the same character is being used for closing abbreviations, decimal points and so on. Interestingly, in Hebrew this problem virtually does not exist, as Hebrew does not overload the full stop with abbreviation function.

Although UAX #29 cannot possibly achieve completeness, it is still beneficial to implement it as the basic set of rules, and apply more fine-grained exception rules on top of it. The ULI TC does not plan to influence the

**IPR Mode:** RAND.

**Current version and work in progress:** Unicode 13.0 was published on March 10, 2020. As of Unicode 7.0, there is an aggressive yearly schedule for major releases. Unicode 8.0 appeared in June 2015, Unicode 9.0 was released June 2016, and so on. Unicode 14 is on track for the first half of 2021.

Unicode 13 added four new scripts, (154 total), 55 new emoji characters, and after adding 5,930 characters in this release, the new total of Unicode characters stands at 143,859. Character groups and whole new scripts continue to be added as per worldwide communities’ requirements. Relatively heavy editorial reshuffle in 7.0 enables the upcoming rapid cycle for major releases. The Unicode core had to adapt to changes introduced in the significant changes in Revision 29 (See UAX #9). Changes in Unicode 11 brought updates to UTS #10 Unicode Collation Algorithm; UTS #39 Unicode Security Mechanisms; UTS #46; and UTS #51 Unicode Emoji.

Each major release of the Unicode standard triggers synchronization of four standalone Unicode technical standards: UTS #10 (Unicode Collation Algorithms), UTS #39 (Unicode Security Mechanisms), and UTS #46 (Unicode IDNA Compatibility Processing, which is compatible processing of non-ASCII URLs), and UTS #51 (Unicode Emoji).

**Standards**
default UAX #29 segmentation behavior and releases the
locale specific segmentation exceptions as part of CLDR.

**Owner:** Unicode Consortium.
**IPR Mode:** RAND.

**Current version and work in progress:** Revision 33 was released for Unicode 11.0 on May 22, 2018, and www.unicode.org/reports/tr29 always links to the current official version. Additionally, www.unicode.org/reports/tr29/proposed.html always points to the latest proposed version, as long as one exists. The Revision 34 for Unicode 12.0 has not been proposed yet at the time of writing.

ULI exceptions use the UAX #29 behavior as a baseline. UAX #29 contains an informative pointer to CLDR released segmentation exceptions developed by ULI TC that can be used to modify the locale independent baseline segmentation behavior described in UAX #29.

**Universal Terminology Exchange (UTX)**

**Short Description:** UTX is a simple bilingual glossary format that was originally targeting MT training. It is a simple tab delimited format, and an XML version also exists, incited by criticism of the localization standardization community. UTX embedding has been specified for XLIFF:doc. TBX-Min mapping was proposed by LTAC Global. Although rather minimal, this tiny terminology exchange standard specifies the part-of-speech field as mandatory and also provides an optional field for term status tracking with predefined values.

**Owner:** The Asia-Pacific Association for Machine Translation (AAMT). Although the AAMT is not a traditional standardization body, their standardization working groups own and maintain UTX.

**IPR Mode:** The documentation is available under Creative Commons 4.0. No IPR mode has been specified.

**Current version and work in progress:** Revision 1.2 Minimal Specification was released February 18, 2018. There are some header changes, and the 1.2 version supports multiple sub-glossaries as part of one glossary. Editing in spreadsheet software became easier. As far as work in progress, the latest developments consisted in the development of TBX-Min mapping by LTAC Global (see TBX).

**XLIFF Object Model (XLIFF OM)**

**Short Description:** An abstract object model defined in UML diagrams and a prose specification. Definition of an XLIFF equivalent data model but also of interoperable file, subfile, and unit fragments. This requires detailed specification of data integrity dependencies that was not needed in XLIFF 2, and that is always being exchanged as a complete XML file.

The main goal is to be able to transfer the XLIFF inline data model into arbitrary XML and non-XLIFF serializations. As such this is groundwork on an interchange data model that is needed to ground any messaging, bus or webservices architecture or a standard Translation API (see TAPICC).

**Owner:** OASIS XLIFF OMOS TC.
**IPR Mode:** Non-Assertion (RF).

**Current work in progress:** XLIFF OMOS TC was chartered in December 2015 to address the industry need of an abstract object model for the XLIFF 2 family of standards. The objective is to provide a serialization independent description of XLIFF and XLIFF fragment data categories that would allow development of non-XML serializations of the same data model. The goal is to make sure that arbitrary serializations of localization interchange data will be interoperable with XLIFF 2, vice versa and among each other. The object model is being developed on a GitHub repository at https://github.com/oasis-tcs/xliff-omos-om. The repository is public but the contributors need to be or become OASIS XLIFF OMOS TC members. Anyone can raise issues or comments via the associated issues or wiki, as well as minor bug fixes via pull requests. It’s up to the repository maintainers if those pull requests will be merged or not, which is the usual GitHub collaboration workflow. UML diagram development on that repository is powered by the open source Papyrus project.

**XML-based Text Memory (xml:tm)**

**Short Description:** xml:tm is a namespace application, which means it is not designed to form an independent document that could exist on its own. Instead, it is designed to be injected as a relatively heavy explicit internationalization apparatus into any well-formed XML document containing human readable language. Unfortunately, it is hardly possible to call this specification a standard due to a very low number of implementations — two, to be exact. The standard is being pushed by only one company without wider industry consensus. It was developed by XTM’s Andrzeic Zydroń and donated to LISA, which published it as an OSCAR standard in early 2007. Its failure to become an actual standard should be a memento of the importance of broad consensus building while creating industry standards.

**Owner:** ETSI ISG LIS.
**IPR Mode:** RF in LISA published versions and FRAND.

**Current version and work in progress:** Zydroń has exposed the 2.0 version for public comment on XTM International’s web page. Its ETSI status is unclear from publicly available sources. The last version by LISA, 1.0, was released on February 26, 2007.
XML Localization Interchange File Format (XLIFF)

**Short Description:** In February 2018, XLIFF TC delivered the first “dot” release of XLIFF 2, which is backwards compatible with XLIFF 2.0. XLIFF 2.1 provides full ITS 2.0 support and advanced validation support. The 2.1 version also deprecated the Change Tracking Module and provided a number of bug fixes in both its core and modules. Interestingly and usefully, XLIFF 2.0 documents are forward-compatible with the XLIFF 2.1 advanced validation artifacts.

The XLIFF 2.0 core covers about 20% of XLIFF 1.2, because in 1.2 everything was core and hence the core was too big to be implemented across the industry.

XLIFF 2.0 and 2.1 were created and approved by a representative group of industry and academic standardizers: big enterprise translation buyers (such as IBM, Microsoft and Oracle), large language service providers (such as SDL and Lionbridge), toolmakers (such as MultiCorpora and ENLASO), industry associations (PSBT, GALA, TAUS), academics (LRC at the University of Limerick) and individuals (notably chair Bryan Schnabel). Many more took interest in the final OASIS organizational approval rounds and during the extensive public reviews.

XLIFF 2 is a lean modular standard that allows for plug and play interoperability in the areas of inline markup, segmentation, glossary exchange, translation and review, engineering quality assurance and so on, effectively catering for an end-to-end mashup of best-of-breed solutions throughout the content value chain.

The translation candidates module allows for local inclusion of translation candidates coming from various sources including translation memories, MT services and various crowdsourcing scenarios. The glossary module allows for local inclusion of relevant terminology in both source and target languages. It supports both monolingual and bilingual scenarios, including terminology life cycle management scenarios and one-to-one mapping with TBX-Basic and UTX is possible. The formatting style module provides two attributes for embedding HTML-encoded preview information that can be used by agents for on-the-fly preview generation to provide context for human translators. The metadata module is a non-namespaced private extensibility mechanism. Custom metadata can be grouped and presented as key-value pairs, which is a predictable way of facilitating interoperable display. The resource data module replaces the binary module capability from 1.2. The change tracking module allows users to log change history, including the provenance of changes. The size and length restriction module provides a generalized way for specifying text size/volume restrictions. This module can cater to simple use cases, such as fitting a database field based on Unicode code point or byte count, and it also caters to complex scenarios, such as fitting specific hardware display restrictions using a specific font. The validation module provides a mechanism for specifying simple localization rules to be checked with relation to source and target, to ensure that a brand name will be included in the target text, for example.

**Owner:** OASIS XLIFF TC.

**IPR Mode:** RF on RAND terms in OASIS (freely available). RAND in ISO (sold).

**Current version and work in progress:** Version 2.1 was published as an OASIS standard on February 13, 2018. ISO 21720:2017 - XLIFF (XML Localization interchange file format) is identical to the OASIS XLIFF Version 2.0 from August 5, 2014.

After successfully delivering the ITS Module and advanced validation capabilities as part of the 2.1 version, the TC started publicly tracking feature development for Version 2.2. Development of XLIFF Version 2.2 has been transferred to Github at https://github.com/oasis-tcs/xliff-xliff-22.

The biggest and most interesting feature proposed for Version 2.2 is Rendering Requirements. A detailed proposal of XLIFF Rendering Requirements was presented at the 40th edition of the ASLING Translating and the Computer Conference on November 16, 2018. The idea is to give possibly normative guidance to CAT tool developers on how to display the XLIFF bitext and its rich metadata, but also to lower the technology complexity of rendering bitext and make interactive rendering of XLIFF possible in web browsers. Other proposed features include specification of semantic domains for placeholders, and roundtrip of segmentation metadata.

The XLIFF TC is also wrestling with the idea of how to use XLIFF for TMX, since TMX is in legal limbo and cannot be maintained or further developed. Three technical options are under consideration: 1) how to informatively describe an XLIFF 2.0 profile that only uses XLIFF 2 Core and the Translation Candidates Module; 2) describe how to use XLIFF Core storage as a translation memory (TM); 3) develop a new grammar that would reuse the XLIFF 2.0 vocabularies but define a different structure that would be more suitable for bulk exchange of TMs, including in multilingual scenarios.

The work on TBX Basic mapping both to and from XLIFF 2.0 (and higher) was transferred to the XLIFF OMOS TC.

The underlying research for this Reader was supported by the Science Foundation Ireland as part of the ADAPT Centre at Trinity College Dublin.
Translation Fights Back

The TICO-19 initiative focuses on translating COVID-19 information, often for under-resourced languages

A year after its emergence, the world is still trying to find ways to quell the COVID-19 pandemic. Different industries look at the problem from a unique perspective — something that is also true for the translation industry. As the pandemic spread across the globe, the translation industry looked at how language and language technology could address a global crisis of information. Thus, the Translation Initiative for COVID-19 (TICO-19) was born. TICO-19 is an initiative that launched at the onset of the crisis and saw translators, technologists, and researchers from Translators without Borders (TWB), Amazon, Appen, Carnegie Mellon University (CMU), Facebook, George Mason University, Google, John Hopkins University, Microsoft, and Translated joining forces, using language technology to make COVID-19 information available in as many languages as possible.

Fighting back with language technology

Wired magazine has declared that COVID-19 is history’s biggest translation challenge.

The TICO-19 effort marks a unique collaboration between public and private entities that came together shortly after the pandemic was declared. The focus of TICO-19 is to enable the translation of COVID-19-related content into a wide range of languages, many of which are underserved by commercial language technology.

If you are lucky enough to have access to information in your own language, it still might be a challenge to find reliable information or understand new terminology that has been emerging across nations. Who had even heard of social distancing this time last year? As the term gained

Eric Paquin is the chief technology officer at Translators without Borders (TWB), overseeing the organization’s language technology development and innovation. He joined TWB in February 2020, bringing over 20 years of experience from the localization and tech industry. Paquin is a French Canadian who has set up roots in Ireland.
popularity, the question became: how do you convey the concept of physical distancing across not only languages, but cultures?

For example, TWB research found that using phrases like “keep away from people” were far better understood in African languages like Swahili than the direct translation for “social distancing.” (see https://www.devex.com/news/how-do-you-say-social-distancing-in-swahili-96856)

Some would be tempted to turn to machine translation (MT) in order to get some of the meaning behind those new terms. But while machines might get a translation correct in some contexts, the meaning in the context of the COVID-19 crisis may be literally lost in translation.

Human translators, on the other hand, do realize that words may be new or may be translated differently in the context of the crisis. But as a translator, how do you access the right terminology? In commercially viable languages, you may have access to certain resources like translation memories (TMs) that have been developed early on. Eventually, machine translations also catch up.

However, if you are looking for information in less commercially viable languages, you might struggle to find the right resources. This is why TICO-19 members are working together to develop efficient and scalable language technology for 37 languages and counting, including languages that are currently under-resourced by technology, like Dari, Dinka, Hausa, Luganda, Pashto, and Zulu.

“Language technology is a powerful tool that can help people communicate more consistently, quickly, and confidently about global issues like COVID-19. Yet many languages don’t have the necessary data needed to build this innovative technology,” explained Grace Tang, TWB’s head of special projects. “We’re excited that industry leaders recognize this gap, and are working with us to develop technology that can help everyone communicate about COVID-19, no matter what language they speak.”

How do you say quarantine in Hausa, for example? Kebancewa sabida magani.

Ensuring COVID-19 terminology is globally equitable and accessible

The translated content focuses on key COVID-19 terminology, ensuring COVID-19 information is more globally accessible and equitable.

TWB brings language technology expertise to TICO-19, particularly for marginalized languages. Its language equality initiative, Gamayun, uses advanced language technology to improve two-way communication in marginalized languages. The ultimate goal is to enable everyone to give and receive information in the language and format they understand.

TWB’s TICO-19 involvement builds on previous Gamayun experience. Through this initiative, TWB has built language datasets and machine translation engines for Rohingya, Tigrinya, Kanuri, Congolese Swahili, and other low-resource languages.

Sourcing the right content and the right translations

Generally, translators working in marginalized languages have far fewer resources with which to
translate, making their jobs more challenging, which further hinders the amount of resources that can be created in those languages. In order to produce helpful tools, we had to ensure the relevance of the content and the quality of the translations.

We created the TICO-19 benchmark, a dataset specialized in the medical domain, which is intended to track the quality of current machine translation systems. The set includes data for very-low-resource languages and was produced with three criteria in mind: diversity, relevance and quality.

First, we sampled from a variety of public sources containing COVID-19-related content, representing different domains. We took special care to diversify the domains and sources of the data, selecting PubMed articles on COVID-19, an English-Haitian Creole dataset from CMU built during the humanitarian response to the 2010 earthquake, and some COVID-19 articles from Wikipedia, Wikinews, Wikivoyage, and Wikisource.

Second, to make our content relevant for relief organizations, we chose the languages to translate into based on the requests from relief organizations on the ground, and the humanitarian priorities of TWB.

Third, we established a stringent quality assurance process, to ensure that the content was translated according to the highest industry standards. The quality of the translations was checked, and revisions were made to ensure quality assurance rated above 95% across all languages, before any subsequent edits. Some low-resource languages like Somali, Dari, Khmer, Amharic, Tamil, and Marathi required several rounds of translation to reach acceptable performance.

The collective effort resulted in a collection of translation memories and technical glossaries so that language service providers (LSPs), translators, and volunteers can make use of them to expedite their work — and ensure consistency and accuracy at the same time.

To help MT practitioners advance medical and humanitarian machine translation, as well as other natural language processing (NLP) applications, we also provided mono-lingual and bi-lingual datasets and an open-source, multi-lingual benchmark set (which includes data for very-low-resource languages) specialized in the medical domain, which is intended to track the quality of current machine translation systems, thus enabling future research in the area.

The collaborators forming TICO-19 have made test and development data available to AI and MT researchers in 35 different languages: nine high-resourced, pivot languages, and 26 lesser-resourced languages, in particular languages of Africa, South Asia, and Southeast Asia, whose populations may be the most vulnerable to the spread of the virus.

The same data is translated into all of the languages represented, meaning that testing or development can be done for any pairing of languages in the set. Furthermore, the team has converted the test and development data into translation memories that can be used by localizers from and to any of the languages.

Readiness for future crises

Enabling efficient and accurate communication through translations for the majority of the world’s languages and particularly the most vulnerable ones is still a long road ahead. With this effort, we only addressed a fraction of the needs for a fraction of the world’s languages. Nevertheless, we hope to pave the way with the MT resources that we have released, and that they will have an immediate impact for the languages we covered and the communities that speak those languages. Our effort has opened up possibilities and communication channels that will allow the MT research community, both academic and commercial, to be more prepared for the next crisis.
But it does not stop here. The initiative also created opportunities for further collaborations. As more organizations and people offer their help, we are hoping to see the data expanded to even more languages with the same level of quality, producing more TMs and creating machine translation language models for low resource languages.

The research aspect is also being expanded to include voice technologies. Many of the communities we help have a low level of literacy and the only way to effectively communicate information is through voice. We are hoping that building voice datasets for these languages will help in the creation of language models and get closer to TWB’s long-term goal of achieving real-time, two-way communication through language technologies.

Be part of the effort
The initiative developed translated datasets for approximately 3,100 key COVID-19 terms and phrases. The resulting datasets, machine translation engines, and translation memories are publicly accessible through TICO-19’s GitHub and TWB’s online language data portal to make sure this specialized content can inform future machine translation initiatives.

If you are a professional translator and have already produced COVID-19 related content, you can share your translation memory and we will combine and release it with ours. Similarly, if you have compiled terminologies with COVID-19 terms, or if you find errors in our published terminologies, reach out and we will update them accordingly.

If you’re interested, you can also volunteer with Translators without Borders (TWB) as long as you are fluent in at least one language other than your native language. Whether you are interested in translating medical texts or translating for crisis response, there are engaging projects available to suit all preferences. Professional translators are especially encouraged to apply. Find out more at www.translatorswithoutborders.org.
**ASSOCIATIONS**

**European Language Industry Association (Elia)**
Elia is the European not-for-profit association of language service companies with a mission to accelerate our members’ business success. We do this by creating events and initiatives that anticipate and serve our members’ needs in building strong, sustainable companies, thereby strengthening the wider industry. Elia was founded in 2005 and has since established itself as the leading trade association for the language services industry in Europe.  

Elia Brussels, Belgium  
info@elia-association.org  
http://elia-association.org

**Globalization and Localization Association**
The Globalization and Localization Association (GALA) is a global, nonprofit trade association for the language industry. As a membership organization, we support our member companies and the language sector by creating communities, championing standards, sharing knowledge and advancing technology.  

Globalization and Localization Association  
Seattle, WA USA  
+1-206-494-4686  
info@gala-global.org  
www.gala-global.org

**CONFERENCES**

**LocWorld**
LocWorld conferences are dedicated to the language and localization industries. Our constituents are the people responsible for communicating across the boundaries of language and culture in the global marketplace. International product and marketing managers participate in LocWorld from all sectors and all geographies to meet language service and technology providers and to network with their peers. Hands-on practitioners come to share their knowledge and experience and to learn from others. See our website for details on upcoming and past conferences.  

Localization World, Ltd. Sandpoint, ID USA  
208-263-8178  
info@locworld.com | https://locworld.com

**TRANSLATION SERVICES**

**STAR Group**
Multiple Platforms  
STAR is a leader in information management, localization, internationalization and globalization services and solutions such as GRIPS (Global Real Time Information Processing Solution), PRISMA (Smart Content Services), STAR CLM (Corporate Language Management) including Transit (Translation & Localization), TermStar/WebTerm (Terminology Management), STAR MT (Machine Translation), CLM WebEdit (Web-based Translation & Review) and MindReader (Authoring Assistance). With more than 50 offices in 30 countries and a global network of prequalified freelance translators, STAR provides a unique combination of information management tools and services required to manage all phases of the product information life cycle.  

Languages: All  
STAR AG (STAR Group headquarters)  
Ramsen, Switzerland  
+41-52-742-9200  
info@star-group.net  
www.star-group.net  
STAR Group America, LLC Lyndhurst, OH USA  
216-691-7827  
lyndhurst@star-group.net

**LOCALIZATION SERVICES**

**Crestec**
Number 1 LSP that you should know  
Crestec is a world-class language service provider. CSA Research ranks Crestec as a global top 5 LSP in the technology sector and global top 6 in the manufacturing sector. Headquartered in Japan with a network of over 18 sites in the US, Europe and Asia, we offer a one-stop comprehensive solution for global communications from marketing content creation and technical writing to localization, printing and studio/shooting production.  

Languages: Japanese, Simplified Chinese, Traditional Chinese, Korean, French, German, Italian, Spanish, Dutch, Swedish  
Crestec USA Long Beach, CA, USA  
612-986-3108  
aki.ito@crestecusa.com  
https://crestecusa.com  
Crestec Europe Amsterdam, Netherlands  
+31 205854640  
sales@crestec.nl  
https://crestec.eu
**Total Solutions for Your Business**

E4NET is a total localization solutions provider including translation, DTP, recording, and specialized in Asian localization covering all major Asian and regional tier 3 languages. We have 20+ years of successful localization production experience with major projects for IBM, Microsoft, SAP, Oracle, HP, LG Electronics, Panasonic and more. E4NET is now providing patent translation services to the Korea Institute of Patent Information and translating life science projects including clinical protocols and reports. We are continuously developing and applying innovative technologies such as machine translation and associated customer services throughout our production process to maximize production/service efficiency. ISO 9001: 2015, ISO 27001 certified.

Languages: 60+

E4NET Co., Ltd. Seoul, South Korea
82-2-3465-8532
110n@e4net.net | www.e4net.net

---

**Beyond the Best**

Since 1992, the main supplier for IBM. Since 2001, the main supplier for SAP. ISIS Korea, Inc. is the one and only localization partner of IBM and SAP in Korea. If your company requires world-class localization services, ISIS Korea is your best choice. We have built an outstanding reputation in localization services for global companies, such as IBM, SAP, SAMSUNG, Apple, Dell Technologies, Amazon, Starbucks, Salesforce, Olympus and more, as a trusted partner based on our customer-oriented services and strict quality control. Customer satisfaction is our core value. We have every solution you need ready for you to request. Just visit us to find out what’s possible at www.isiskorea.com.

ISISKOREA Inc. Seoul, South Korea
82-2-3787-0111
lion@isiskorea.com
www.isiskorea.com

---

**ORCO S.A. Localization Services**

Founded in 1983, ORCO celebrates this year its 35th anniversary. Over the years, ORCO has built a reputation for excellence and gained the trust of leading companies, such as Oracle, IBM and Carrier for the localization of their products. Our core business activities include technical, medical, legal, financial, marketing and other translations, software and multimedia localization, as well as localization consulting. We cover most European languages and our client list includes long-term collaborations with international corporations, government institutions, banks, private enterprises, NGOs and the European Union. ORCO is certified according to ISO 17100 and ISO 9001 quality standards.

Languages: Greek and European languages

ORCO S.A. Athens, Greece
+30-210-723-6001
info@orco.gr
www.orco.gr

---

**E4.**

---

**Mobico — by Saltlux Inc.**

Mobico is the new brand name of Saltlux’s technical communication services, and is also the name of the predecessor company to Saltlux, established in 1979 as Korea’s first TC business. What started as a small enterprise concentrating on creating Korean manuals and East Asian language translations evolved into a one-stop service provider for all your needs in the world of business today, including multilingual translation, localization, DTP, TW and MTPE. The relentless pursuit of progress and perfection results in the use of state-of-the-art technology and processes, which in turn lead to superior translation quality with shorter turn-around times and therefore to greater customer satisfaction.

Languages: More than 70 languages.

Saltlux, Inc. Seoul, South Korea
+82-2-2193-1725
sales@mobico.com
www.mobico.com/en

---

**RWS Moravia**

RWS Moravia is a leading globalization solutions provider, enabling companies in the IT, consumer electronics, retail, media and entertainment, and travel and hospitality industries to enter global markets with high-quality multilingual products and services. RWS Moravia’s solutions include localization, testing, content creation, machine translation implementations, technology consulting and global digital marketing services. Our customers include eight of Fortune’s Top 20 Most Admired Companies, and all of the “Fab 5 Tech Stock” companies from 2017. Our global headquarters is in Brno, Czech Republic, and we have local offices in Europe, the United States, Japan, China and Latin America.

Languages: More than 250 languages.

RWS Moravia USA: Thousand Oaks, CA USA
+1-805-262-0055
RWS Moravia Europe: Brno, Czech Republic
+420-545-552-222
moravia@rws.com
www.rws.com/moravia
A creative approach to localization

At Spark, we fuse creativity with technology to provide a localization service. Helping brands like Microsoft, Disney and Netflix to develop and deliver consistent global branding and hyperlocal content creation from marketing activation to packaging. Providing creative expertise across the entire consumer journey, continuous workflow solutions and seamless integration into business systems: making a real difference to your localization needs.

Languages: 120+

Spark - Brighter Thinking
Europe: London, UK
+44 (0)207 602 9119
hello@sparkbrighterthinking.com
www.sparkbrighterthinking.com

Joint National Committee for Languages

The Joint National Committee for Languages and the National Council for Languages and International Studies (JNCL-NCLIS) represent the interests of over 140 member organizations, associations and companies in virtually all aspects of the language enterprise — education PreK-20, research, training, assessment, translation, interpreting and localization — to the US government. The mission of JNCL-NCLIS is to ensure that all Americans have the opportunity to learn English and at least one other language.

Joint National Committee for Languages – National Council for Languages and International Studies
Garrett Park, MD USA, 202-580-8684
info@languagepolicy.org
http://languagepolicy.org

Translation Commons

Translation Commons is a nonprofit US public charity powered by translators. We are a volunteer-based online community aiming to help our language community thrive and bridge all the sectors within our industry. We facilitate cross-functional collaboration among the diverse sectors and stakeholders within the language industry and instigate transparency, trust and free knowledge. Our mission is to offer free access to tools and all other available resources, to facilitate community-driven projects, to empower linguists and to share educational and language assets.

Translation Commons Las Vegas, NV USA
(310) 405-4991
krista@translationcommons.org
www.translationcommons.org

Translators without Borders

Originally founded in 1993 in France as Traducteurs sans Frontières by Lori Thicke and Ros Smith-Thomas to link the world’s translators to vetted NGOs that focus on health and education, Translators without Borders (TWB) is a US nonprofit organization that aims to close the language gaps that hinder critical humanitarian efforts worldwide. TWB recognizes that the effectiveness of any aid program depends on delivering information in the language of the affected population.

Languages: 190 language pairs

Translators without Borders CT USA
info@translatorswithoutborders.org
www.translatorswithoutborders.org

Consoltec

Multiple Platforms

Consoltec offers FlowFit-TMS, a web-based translation management system that helps you simplify and optimize your projects, while reducing your administrative costs. FlowFit can also be used for many other project types. FlowFit provides fully customizable web portals for clients, providers and project management. Get an accurate overview of your teams’ workload in real time and select the best available providers. Manage your clients, contacts and internal/external providers effectively with the new CRM features. Use Timesheet to track the time spent on projects and tasks. Connect seamlessly to your favorite CAT tools (memoQ, SDL Studio, LogiTerm) and get comprehensive reports that provide enhanced insight on production, productivity, costs and translation memory efficiency.

Consoltec Montreal, Québec, Canada
(+1) 514 312-2485
info@consoltec.ca
www.consoltec.ca

NONPROFIT ORGANIZATIONS
Localize

Localize offers a full-featured, cloud-based content and translation management system that features advanced translation workflows, allowing content managers and translators to propose, review, and publish translations with ease. For companies without in-house translators, we provide access to high-quality, on-demand translations through our network of professional translators. Our easy to install plugin fits neatly into your existing technology stack. The technology powering the Localize Platform was built from the ground up to minimize the need for engineers in the localization process. This reduces costs by enabling nontechnical personnel to manage the localization workflow. Getting started is easy. Start your free trial today!

Languages: All
Localize Kingston, NY, USA
(415) 651-7030
sales@localizejs.com
https://localizejs.com

Plunet BusinessManager
Multiple Platforms

Plunet develops and markets the business and workflow management software Plunet BusinessManager — one of the world’s leading management solutions for the translation and localization industry. Plunet BusinessManager provides a high degree of automation and flexibility for professional language service providers and translation departments. Using a web-based platform, Plunet integrates translation software, financial accounting and quality management systems. Various functions and extensions of Plunet BusinessManager can be adapted to individual needs within a configurable system. Basic functions include quote, order and invoice management, comprehensive financial reports, flexible job and workflow management as well as deadline, document and customer relationship management.

Plunet GmbH Berlin, Germany
+49 (0)30-322-971-340
info@plunet.com
www.plunet.com

birotranslations

birotranslations specializes in life science, legal, technical, IT and automotive translations into all East European languages (Albanian, Bosnian, Bulgarian, Croatian, Czech, Estonian, Hungarian, Latvian, Lithuanian, Macedonian, Polish, Romanian, Russian, Serbian, Slovak, Slovenian, Ukrainian). We have a long-term partnership with the world’s top 100 MLVs and many end-clients all around the globe. With our experienced project managers, extensive network of expert linguists and usage of the latest CAT tool technology, your projects will be delivered on time, within budget and with the highest standards of quality. For more information, please contact Mr. Matic Berginc (details below).

Languages: Eastern European languages
birotranslations Ljubljana, Slovenia
+386 590 43 557
projects@birotranslations.com
www.birotranslations.com

MEMSOURCE

MEMSOURCE is a leading cloud-based translation management system that enables global companies, translation agencies and translators to collaborate in one secure, online location. Internationally recognized for providing an easy-to-use, yet powerful CAT tool combined with a TMS, MEMSOURCE processes two billion words per month from over 200,000 users around the world. Manage your translation projects in real-time in an intelligent platform that accepts over 50 file types and offers REST API, out-of-the-box CMS connectors and powerful workflow automation to save time and money. Join localization professionals from around the world who rely on MEMSOURCE to streamline their translation process. To start your free 30-day trial, visit www.memsourcing.com.

Languages: All
MEMSOURCE Prague, Czech Republic
+420 221 490 441
info@memsource.com
www.memsourcing.com

SMARTLING

SMARTLING

SMARTLING Translation Cloud is the leading translation management platform and language services provider to localize content across devices and platforms. SMARTLING’s data-driven approach and visual context capabilities uniquely position brands for efficiency. Seamlessly connect your CMS, code repository, and marketing automation tools to SMARTLING’s TMS via prebuilt integrations, web proxy, or REST APIs. No matter the content type, SMARTLING automation tools help you do more with less. SMARTLING is the platform of choice for B2B and B2C brands, including InterContinental Hotels Group, GoPro, Shopify, Slack, and SurveyMonkey. The company is headquartered in New York, with offices in Dublin and London. For more information, please visit SMARTLING.com.

SMARTLING New York, NY, USA
1-866-707-6278
hi@smartling.com
www.smartling.com

GLOBALWAY

GLOBALWAY

GlobalWay Co., Ltd.

As an industry-leading localization company in Korea, GlobalWay has been providing incomparable professional localization services with exceptional quality to partners all around the globe since 2003. We are here to offer language solutions including translation, voiceover, testing, DTP and engineering services. Our highly qualified in-house linguists in each field of expertise, experienced engineers and project managers will add value to your growing business. GlobalWay and its long-term global partners are ready to support you on the road to success. Are you looking for a reliable partner? Our doors are wide open for you. Should you need more information, please feel free to contact us.

Languages: 50+ more languages including Korean, English, Chinese, Japanese, German, Russian, Vietnamese, Thai, Indonesian.
GlobalWay Co., Ltd. Seoul, South Korea
+82 2 3453 4924
sales@globalway.co.kr
www.globalway.co.kr

January/February 2021
Your Partner in Asia and Beyond!
With our headquarters in Korea, our production offices in Vietnam and China, and our sales office in the US, we are in an excellent position to be your Asian language localization partner. For localizing projects from English or German into Asian languages, such as Korean, Japanese, Chinese, Vietnamese, Thai, Indonesian and Burmese, you can trust our professional translation services for IT, software, marketing/transcreation and technical projects. Since our establishment in 1990, we have been at the forefront of the localization industry as one of the Asia Top Ten and the No. 1 LSP in Korea (by CSA Research). ISO17100 certified since 2014.
Languages: More than 54 languages including Korean, Chinese, Japanese, Vietnamese, Thai, Indonesian.
Hansem Gyeonggi-do, South Korea
+82-31-226-5042
info@hansem.com
http://hansem.com/en

Medical Translations Only
MediLingua is one of the few medical translation specialists in Europe. We only do medical. We provide all European languages and the major languages of Asia and Africa, as well as translation-related services to manufacturers of devices, instruments, in vitro diagnostics and software; pharmaceutical and biotechnology companies; medical publishers; national and international medical organizations; and other customers in the medical sector. Projects include the translation of documentation for medical devices, surgical instruments, hospital equipment and medical software; medical information for patients, medical students and physicians; scientific articles; press releases; product launches; clinical trial documentation; medical news; and articles from medical journals.
Languages: 45, including all EU languages
MediLingua Medical Translations BV
Leiden, Netherlands
+31-71-5680862
info@medilingua.com
www.medilingua.com

Translated.
Professional translation services made easy. Crafted by expert humans, powered by technology, efficiently delivered. We have delivered 1.2 million translations in 150 languages to 134,091 clients in 40 macro-domains since 1999, powering the globalization strategy of the most demanding clients. We work hard to make translation services more effective, by enhancing our production processes with great technologies and talented people. A perfect example is T-Rank™, the system that instantly matches your content with the most qualified translator for the job. We offer a wide range of linguistic services that cover all your future needs: Google Ads translation, software localization, subtitling, and APIs to integrate human translation. We open up language to everyone.
Languages: 150 languages and 40 areas of expertise.
Translated Rome, Italy
+390690254001
info@translated.com
www.translated.com

iDISC Information Technologies
iDISC, established in 1987, is an ISO 9001 and ISO 17100 certified language and software company based in Barcelona with branches and teams in Mexico, Brazil, USA, Argentina, Bolivia and Guatemala. We have dedicated teams for web content, software localization and translation of technical, business, automotive, biomedical and marketing documents. Our software development engineers and translation teams provide high-quality and on-time production solutions that are cost-efficient, flexible and scalable.
Languages: Spanish (all variants), Portuguese (all variants), Catalan, Basque, Galician, Valencian, Kiche, Quechua, Aymara, Guarani
iDISC Information Technologies, S.L. Barcelona, Spain
34-93-778-73-00
info@idisc.com
www.idisc.com

Rheinschrift Language Services
Outstanding localization requires world-class experience. Rheinschrift gives your business a native voice in the German-speaking world. We offer more than 20 years’ experience providing translations and localizations for software and hardware manufacturers as well as for the sectors of business, technology, legal matters and medicine/medical applications. Our services also range from glossaries, post-editing, project management and desktop publishing services to many other related services. Rely on Rheinschrift to deliver the most competent translations and meet your deadline, whatever it takes.
Languages: German to/from major European languages
Rheinschrift Language Services Cologne, Germany
+49 (0)221-80-19-28-0
contact@rheinschrift.de
www.rheinschrift.de

memoQ
memoQ is a technology provider that has been delivering premium solutions to the translation industry since 2004. For almost 15 years, memoQ has been dedicated to delivering innovation through diverse developments that today help hundreds of thousands of freelance translators, translation companies and enterprises worldwide. Having simplicity and more effective translation processes in mind, memoQ combines ease of use, collaboration, interoperability and leveraging in one single tool. Discover a new world with memoQ, and let our team help optimize your translation processes and make your business more successful.
Languages: All
memoQ Budapest, Hungary
+361888313
sales@memoq.com
www.memoq.com

memoQ
Hanssem
ISO 9001 & 17100 CERTIFIED

TRANSLATION TOOLS

iDISC

Medical Translations Only
MediLingua

Translated

Rheinschrift

memoQ
SDL plc

SDL is the global innovator in language translation technology, translation services and content management. Over the past 25 years we’ve helped companies deliver transformative business results by enabling powerful, nuanced digital experiences with customers around the world. SDL is the leading provider of translation software to the translation industry and SDL Trados Studio is recognized globally as the preferred computer-assisted translation tool of government, commercial enterprises, language service providers and freelance translators.

Languages: All
SDL plc Maidenhead, United Kingdom
+44 1628 417227
info@sdltrados.com
www.sdl.com
www.sdltrados.com
In 2010, Translators without Borders (TWB) was formed in response to the Haiti earthquake. We translated aid information into Haitian Creole, established a translation platform, built a community of skilled linguists, and established a non-profit organization to help with this crisis and respond to other emerging crises around the world.

In 2020, to meet the unprecedented multilingual communications challenge of the COVID-19 global pandemic, TWB rapidly mobilized. We translated millions of words of COVID-19 information, partnered with a myriad of NGOs from around the world, and developed innovative language technology.

It’s been ten years. Throughout it all, TWB has helped people overcome crisis, poverty, and many other challenges by providing the means for them to access important information and communicate in their own language.

This year, celebrate ten years with us.
Today, ten years from our founding, TWB is the globally recognized humanitarian non-profit that believes that overcoming language barriers is key to safety, security, and to building our shared humanity. We work with a network of over 30,000 volunteer translators and a wide range of partners to deliver information, power and agency to people in need.

In humanitarian crises, we create tools like glossaries and multilingual chatbots. We ensure health information reaches everyone, especially speakers of lesser-served languages. More broadly, we develop scalable technology for these languages. Without such resources, speakers of marginalized languages face an ever-widening global knowledge gap.

TWB’s ultimate goal is to shift control of communication, helping people to access information in their language and, just as importantly, to share their own ideas and raise their voices. We must build technology that facilitates listening to what people have to say, no matter what language they speak.

TWB seeks a world where knowledge knows no language barriers. Without your help, we cannot continue this work. And, if we don’t continue, millions of people will be left behind. It is time to act.

On TWB’s ten year anniversary, we want to thank you for all of the support you have given us over the years. The work is far from over. We need your help to use the power of language and effective communications to help solve the world’s most challenging problems.

Are you in?
Takeaway

The TWB Journey
Ten years and counting

Donna Parrish

Donna Parrish is principal and co-founder of LocWorld. She is also secretary of Translators without Borders. In a past life, she was publisher of MultiLingual.

It was January of 2010. A 7.0 magnitude earthquake had devastated Haiti. Since only 5% of the Haitian population was fluent in French, aid communication needed to be in Haitian Creole. This was a huge task. After this experience, Lori Thicke, who had founded the French organization Traducteurs sans Frontières, realized the world needed a larger, US-based organization to head up linguistic crisis response efforts. An organizational meeting of Translators without Borders (TWB) was held at a LocWorld in Seattle the following autumn.

We hit the ground running. Within months, Henry Dotterer and ProZ.com had set up a platform for our translators and aid organizations to use. As with most new nonprofits, the board was the organization — doing much of the operational work. Unfortunately, the world has never had a shortage of emergencies, be it typhoons, earthquakes, epidemics, refugee crises — the list is, unfortunately, very long. And the work was too much for a volunteer board. So Rebecca Petras, our first hire, moved from board member to program manager.

Growth

It’s interesting: you think you can fulfill the translation requests of aid organizations and your job is done. But the more work we did, the more work we saw that should be done. Often the language needs are hard to meet because the languages are lesser-served, and there are few available translators. We set up a translation training center in Kenya to address that need locally, and eventually trained almost 250 people.

In crisis response, critical information needs to be translated immediately. So, we began the “Words of Relief” project, creating an international network of translators that can provide immediate assistance. In the 2015 Nepal earthquake, we had translators online within hours, translating the Twitter pleas for help along with critical geographic information, no doubt saving lives.

We realized that known language data was often lacking or inaccurate for critical regions. TWB’s Language Data Initiative maps languages in various countries such as Nigeria, along with communication preferences of the populace (Figure 1).

Health language data is obviously an area of our concern. There is no word for “rape” in Swahili. You can imagine how challenging it is to help victims around the world with no access to accurate, culturally sensitive translations of health and medical terminology. Enter our Simple Words for Health program, a database of 12,000 essential medical terms that have been simplified and translated into more than 40 world languages by qualified doctors and trained medical translators. And then there’s the Project Wiki, where TWB partners to ensure that critical healthcare information reaches an estimated to 3.3 million people every month. Since the partnership began in 2012, 1,900 medical articles have been made accessible on local-language Wikipedias, with TWB providing the translations into 83 languages.

As TWB matured as an organization, we realized to truly enable communication on the world stage, we needed an executive team with experience and
long-time connections in the international aid community. At the end of 2015, we hired Aimee Ansari as our executive director, and, thanks to her guidance, we have never looked back.

In 2017, The Rosetta Foundation merged with TWB, bringing in many more translators. Our translation platform evolved into Kató, which now includes machine translation, translation memory, and enhanced quality assurance tools.

TWB had set up an impressive set of glossaries prior to 2020 for regions such as northeast Nigeria, Myanmar, Mozambique, Democratic Republic of Congo (DRC), and Bangladesh.

When the pandemic hit, we also developed a COVID-19 glossary in over 50 languages and scripts. We developed a chatbot for people in the DRC, and, as of this writing, we have translated almost six million words in 105 languages.

As our work and message spread, it was gratifying to see aid organizations coming to us first to determine the language needs for new projects. This interest and demand is envied in the commercial world!

We have been honored to be featured in multiple aid-related publications as well as such well-known outlets as BBC, CNN, Forbes, The Economist, The Lancet, The Telegraph, Wired and, of course, MultiLingual.

**Now and beyond**

What’s new with TWB? I don’t have enough room here to list everything! Gamayun, the language equality initiative, comes to mind immediately. Supported by Cisco, Microsoft, Twilio, ImpactCity, and Kinsasha Digital, Gamayun uses advanced language technology to increase language equality and improve two-way communication in marginalized languages. The initiative focuses on gathering text and speech data that can make it easier to automate marginalized languages. Using this data, we can build advanced technology-driven solutions for both text- and voice-based communication.

The ultimate goal is to shift control of communication, allowing people to access information in the language and format they understand and to share their voices. Do you see that change? We are moving from only translating information for people to also listening to what they have to say. This is critical in addressing their needs and concerns.

A new product I am excited about is our chatbot technology, which is not menu-based, but rather uses natural language processing. This is already deployed for people in the DRC. With conversations in French, Congolese Swahili, and Lingala, the chatbot helps with COVID-19 and Ebola questions, and in so doing makes us aware of educational needs, prevalent concerns, and false rumors. This chatbot lives on WhatsApp, but future ones will be hosted on whatever platforms are most commonly used in the target area. Plans are already in the works for other locales and applications.

It has been an intense and exciting ride, this TWB adventure. I am honored to be part of it. As an unprecedented and groundbreaking organization, TWB faces unique challenges fitting into established categories to get funding for our work. So, I invite you to join in the solution. When we look at the “take-away” for our lives, TWB makes a difference. 😊
Sarah Weise is the CEO of award-winning marketing research agency Bixa and the bestselling author of *InstaBrain: The New Rules for Marketing to Generation Z*. For 15 years, she has been a guide to hundreds of leading brands including Google, IBM, Capital One, Mikimoto, PBS and Real Warriors, to name a few. Sarah helps brands achieve a laser-focus on their customers and build experiences that are downright addictive. She lectures at Georgetown University’s McDonough School of Business on marketing strategy for the next generation and speaks at conferences and corporate events worldwide.

Multiple specialized tracks of highly valuable sessions, delivered by knowledgeable and experienced speakers from these leading brands and others:

- Avigilon
- Calm
- HubSpot
- KnowBe4
- Mozilla
- NetApp
- Netflix
- Nike
- Red Hat
- SAP
- Sony
- SurveyMonkey
- Tinder
- Workday
- VMware
- WIX

View the full conference program at locworld.com/lww43program

... and introducing our “Global Toolbox” days. Virtual workshops and round tables a week prior to LocWorldWide43 to further build your skill and knowledge libraries. Discover all of our Global Toolbox sessions at:

locworld.com/lww43toolbox

REGISTRATION OPEN NOW locworld.com
Meaningful conversations with thought leaders

Stream with us and join the conversation:

YouTube  Facebook  LinkedIn  Twitter  Vimeo

Watch the previous seasons on YouTube:
youtube.com/c/MultiLingualTV

previous recordings
COVID-19 has shifted people to online shopping, e-learning and digital healthcare quickly. In this new normal, it’s essential that these applications be as interesting and effective as possible. Gamification, a technique to add game-like elements to web and mobile applications, can help impact and inspire your users.

My mission is to provide gamification strategies and deployments to ensure that global enterprises have engaging, motivating and innovative e-solutions for their employees and customers.

www.rws.com/moravia