MENULANCE

WHITE PAPER

Data: the secret sauce driving AI food translation

Dr Roberto Mariani explains why training data is the secret sauce to baking the perfect machine translation. Learn 8 best practices Menulance pioneered to curate and transform Genesis—a training dataset achieving 97% translation accuracy.



CO-AUTHORED BY DR ROBERTO MARIANI AND ANTHONY COUNDOURIS **TRANSLATED** INTO ITALIAN AND SPANISH BY DR FRANK BADRINES

WORD FROM THE CTO

We've assembled Genesis, a multilingual dataset comprising 166 million lines of dishes and ingredients. It’s part of a project that spanned a decade. Translation engines trained on Genesis achieve an impressive 83% BLEU score for English to Spanish translation. To provide context, a skilled human translator typically scores around 60% on the BLEU scale.

The success is attributed to Genesis’s narrow training dataset, short sentence training methodology, and comprehensive sentence fragmentation.

In contrast, mainstream tools like Google Translate and Bing falter in food translations due to training on general dictionaries, resulting in inaccurate translations.

We represent a game changer for cuisine enthusiasts and the AI community. The core of AI effectiveness lies not just in technological expertise, but in the meticulous curation of clean, relevant datasets.

Let’s translate the world’s food.

ABSTRACT

Artificial Intelligence (AI) is changing the way we consume food. Artificial intelligence acts as a contemporary extension of traditional media, such as restaurant menus or cookbooks, by seamlessly integrating various sources like text, images, maps, and video data. This integration enhances the accessibility of information, making the understanding of food more comprehensive.

AI-driven language translation systems assist travelers in communicating with locals, comprehending menus, and participating in food classes by eliminating language barriers. This fosters cross-cultural dialogue and enriches the overall food tourism experience.

Digital technologies and AI offer tourists personalized and immersive experiences, such as interactive food tours, virtual cooking classes, and augmented reality storytelling. These advancements boost engagement and create enduring memories for tourists.[1]

AI plays a crucial role in documenting and digitizing traditional recipes, cooking techniques, and cultural practices, thereby preserving and safeguarding culinary heritage, particularly in regions where traditional knowledge is at risk of being lost.

Furthermore, AI is revolutionizing restaurant management and the sharing of recipe suggestions. AI and digital innovations can streamline restaurant operations, including inventory control and reservation systems, leading to reduced wait times and an enhanced dining experience for both patrons and restaurateurs.

AI-powered platforms can analyze regional ingredients, culinary customs, and global food trends to offer better food and recipe suggestions. This inspires travelers to explore new flavors and culinary techniques.

However, amid the excitement surrounding AI's potential, there's a persistent challenge: the tendency to overpromise and underdeliver. This arises from a fundamental imbalance, where the focus on creating cutting-edge technology often overlooks a crucial aspect—the importance of curating clean datasets upon which AI engines are trained.

I was formerly working with a team of scientists to create a mobile, speech-conversant robot called Jijo-2 at Japan's Electrotechnical Laboratory. The Jijo-2 project started in 1997 and was a collaboration involving multiple scientists and

engineers, focusing on integrating various technologies such as speech recognition, navigation, and facial recognition to enable natural interaction in office environments.

I was tasked with solving facial recognition. We tried every known AI engine at the time to help Jijo-2 recognize and match a face with a name. With each engine we tried, the results were the same: any engine trained on our dataset achieved an average of 44% accuracy.

It wasn’t until we began focusing on data and not the engine that we saw accuracy improve. At the time, this was a breakthrough in thinking. Scientists were obsessed with AI engines. Jijo-2 proved that curated datasets were far more important.

This white paper explores the importance of high-quality data. We use Menulance's training dataset called Genesis to show how data quality can make or break AI applications.

INTRODUCTION

In the mid-1800s, during the Industrial Revolution, British farmers began using steam-powered machines like threshers and plows to grow more food for a fast-growing population. These machines boosted productivity and brought new hope to agriculture.

By the late 1800s, farm equipment had become highly advanced, but farmers hit a limit. The machines couldn't improve yields any further.

So, farmers shifted focus—from machines to the soil. In regions like Yorkshire and East Anglia, they began experimenting with crop rotation, fertilization, pest control, and selective breeding. They also improved irrigation and drainage to create better growing conditions.

Innovators like Jethro Tull helped lead this change with tools like the seed drill and new soil techniques, laying the foundation for a more scientific approach to farming. These changes helped increase crop yields not by making better machines, but by better understanding and working with the land.

By the end of the 19th century, agricultural progress was no longer just about machines—it was about soil, science, and strategy.

The same is true for machine translation today. The engines are already powerful. To improve further, we must look beyond the machine and focus on the data it learns from.

BRAISED CROCODILE IN MARMALADE SAUCE

GROWTH 🍔

The culinary world has exploded

Food is not merely a biological necessity; it is an integral aspect of life, playing a central role in numerous cultures. Food embodies the entire social environment and serves a commemorative function, connecting contemporary times to ancestral practices and preserving a society's cultural heritage.

As the old saying goes, *we are what we eat*, and activities related to food—such as cooking, eating, and discussing it—occupy a significant portion of our daily lives. Author A. Sonnenfeld, in *Food: A Culinary History from Antiquity to the Present,* suggests that food is perhaps the most distinctive expression of an ethnic group, a culture, or, in modern times, a nation.[2]

Culture not only dictates what people eat but also shapes how they consume food.[3] According to Claude Lévi-Strauss, cooking functions as a language, translating its unconscious structure. Norms surrounding meal sequences and constituents reflect social connections among diners, and culinary codes influence the significance assigned to different functions of food. In Japan, for instance, the appearance of food is considered as crucial as its taste, echoing Goethe's belief that a meal should please the eye first and then the stomach.

Food sharing is found in religious rituals

The phrase *breaking bread* holds profound biblical significance, rooted deeply in Christian tradition. Beyond its historical context, breaking bread embodies principles of fellowship and community within Christianity. The Bible depicts the early Christian church engaging in the practice of sharing meals, alongside prayer and the apostles' teaching. This communal act signified unity among believers and strengthened their bonds of faith.

In Judaism, the act of breaking bread is central to the observance of Shabbat, the weekly day of rest and spiritual rejuvenation. During the Shabbat meal, families traditionally gather to share bread, often in the form of challah, which is blessed and then broken before being distributed among those present. This act symbolizes unity, gratitude, and the importance of communal connection within Jewish tradition.

In Hinduism, the act of sharing food holds significant importance, particularly in the context of the ritual known as *Prasad* or *Prasadam*. Prasad refers to food that has been offered to a deity during worship, after which it is distributed to devotees as a blessed gift. Sharing Prasad symbolizes spiritual grace, blessings, and the bond between the worshipper and the divine.

Observant Muslims follow dietary guidelines outlined in Islamic law, known as Halal. This includes permissible (Halal) and prohibited (Haram) foods. For example, Muslims are prohibited from consuming pork and alcohol, and they must ensure that meat is slaughtered according to Islamic principles. The act of ensuring that food is permissible according to Islamic law is itself a significant aspect of Islamic practice.

In essence, food serves as a powerful vehicle for cultural preservation and expression, embodying the traditions, history, and identity of a society.

Fusion cuisine has emerged

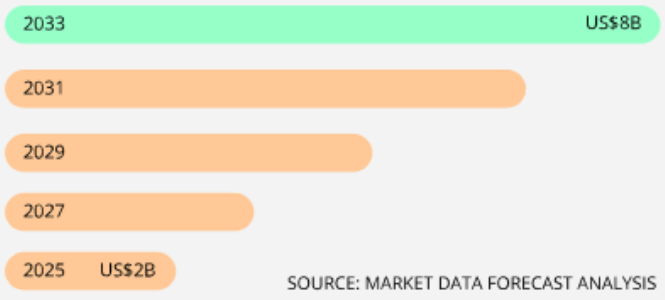
Immigrants have long contributed their food traditions from their native countries to the local cuisine of their new homes. For example, cookbook writer and anthropologist Claudia Roden notes that the iconic British dish *fish n' chips* was introduced to Britain in the 16th century by Jewish immigrants who were forced to leave Spain and Portugal.[4]

Globalization has not only facilitated the spread of global brands and fast-food chains but has also fostered an appreciation for local cuisines and ethnic ingredients. Fusion cuisine has emerged as a significant trend, blending ingredients and techniques from historically distinct culinary traditions. This global culinary exchange has increased the demand for wholesome foods and organic produce.

Mediterranean flavors are often paired with spices from the East, traditional British pies and puddings are reinvented with ingredients like basil or wasabi, and street food has been elevated to haute cuisine status.[5] Fusion cuisine can be seen as a form of translation or transcreation, blending elements from multiple cultures into unique culinary creations.

Food tourism

Culinary or food tourism is the pursuit of unique and memorable eating and drinking experiences, both near and far. Food tourism is helping popularize translation apps like Menulance. In 2024, food tourism hit $1.81B and is anticipated to have a value of $7.91B by 2033.



Through food and drink, travelers get a sense of the local culture, heritage, and people. These apps become indispensable tools, enhancing the overall food experience.[6]

- 70% of people spend more money on food abroad than they do at home, which reveals the importance of a unique food experience when traveling;
- 70% of people pick a destination based on the food and drink there;
- Food tourism statistics show that 81% of people think that food helps them understand a culture, according to *Jersey Island Holidays*.

Tourism offices from Singapore, Taiwan and South Korea are promoting the quality, taste and health benefits of their dishes to the travelers. While Europe is largely dominating the food tourism market, it is very interesting to see that Asian countries are now promoting their local dishes to foreigners.

Travelers switch to discovery mode

In a food tourism study conducted in 2020, researchers discovered that tourists alter their eating habits while on vacation. For instance, travelers visiting sun-and-sea destinations consume significantly more seafood and fish but less legumes, meat, fast food, cereals, and their derivatives. Both international and domestic tourists in sun-and-sea locales emphasized that sampling local cuisine significantly enhances their overall tourism experience.

Essentially, travelers enter a *discovery mode* mindset, where they are open to trying new dishes and are willing to deviate from their usual diet. They often embark on extensive research and exploration before making dining decisions.

The same study found:

- Before visiting their destination, 80% of travelers will research food and drink;
- In the last two years, 93% of tourists have at least one unique food or drink activity;
- 77% of millennials travel to have memorable eating and drinking experiences, according to *Jersey Island Holidays*.

The food porndemic

Many people share pictures of the food they are eating across social media platforms. A search on Instagram for #food yields at least 300 million posts, while #foodie results in at least 100 million posts, underscoring the undeniable significance of food in our society.[7]

This trend reflects a broader cultural shift where images of appetizing food are increasingly pervasive. Referred to as *food porn* or *gastro porn*, these high-resolution images populate best-selling cookbooks, Instagram, X (Twitter), commercials, and magazine ads.[8]

In Britain, food has become akin to a new form of cultural currency, surpassing the allure of fashion or music preferences. Food magazines capture food in such intimate detail that readers feel immersed in the sensory experience, whether it's the texture of a pepper or the aroma of freshly baked bread. Londoners, once defined by their choice of designers or music tastes, now identify themselves by their dining preferences and the recipes they experiment with.

Food, as a universally enjoyed element, plays a central role in cultural interactions. Influential figures like Jamie Oliver, Michael Pollan, Mark Bittman, and Rachael Ray advocate for a return to the kitchen, emphasizing the importance of sharing meals despite religious and dietary differences. The proliferation of intercultural exchanges, migration, and media exposure has heightened the demand for translated cookbooks and menus, with blogs and websites serving as platforms for sharing recipes globally.

The influence of food television channels like Food Network and UKTV Food, along with popular cooking shows such as MasterChef and The Great British Bake Off, has been instrumental in disseminating knowledge about new gastronomic traditions. These programs not only instruct viewers on the preparation of exotic dishes but also fuel interest in cookbooks and food magazines.

Home cooking back on the menu

Cookbooks sell extraordinarily well. According to Kristen Mclean of the NPD market research group, close to 20 million cookbooks fly off the shelves every year in the US alone. Not only that, but it’s one of the most stable book markets with only tiny annual fluctuations. It even survived the COVID-19 pandemic.[9]

Cookbooks are the sixth most popular book genre in the United States, making up 5% of total book sales.[10]

Nielsen BookScan data shows that cookbook sales in the US grew 8% year-on-year between 2010 and 2020, with sales numbers boosted even further by the pandemic.[11]

While there was nothing good about the pandemic itself, it boosted the cookbook market as more people tried their hands in the kitchen—remember all those cookies, banana cakes, and loaves of sourdough bread? For many, they discovered or rediscovered a passion for good home cooking that's continued even since we *got back to normal*.[12]

In the first year of the pandemic, cookbook sales spiked about 16%. Though they have dropped off a bit since then, sales remain strong. What changes more is the kinds of cookbooks people are buying at any given time.[13]

"In 2020, after the pandemic arrived, we saw a very sharp uptick in certain types of cookbooks, including cookbooks on the basics for cooking for yourself, cookbooks on bread baking, cookbooks a little bit later on cocktail making and other types of like in-home entertaining," McLean said. "So it was a real ticker tape of the psychology of the folks who were stuck at home."[14]

Today, it's a completely different list.[15]

"Now the emphasis is really on quick and easy, single-pot dishes," she said. "The type of thing you would expect if people were having to integrate cooking back into very busy lives."[16]

SIZE 🤯

610M people snack on machine translation each day

AI has given birth to machine translation, which is considered a branch or subfield of Artificial Intelligence. It falls under the broader umbrella of AI because it involves the development of systems and algorithms capable of automatically translating text or speech from one language to another.

Machine translation uses AI techniques, including natural language processing and machine learning, to enable computers to understand and generate human languages, facilitating cross-lingual communication.

Machine translation can employ different approaches such as rule-based methods, statistical models, or neural networks. Also known as automatic translation, machine translation quickly became accessible to everyone with the advent of the internet, leading numerous technology providers to offer machine translation services at no cost.

After more than 70 years of evolution, significant achievements have been made in machine translation. Machine translation applications utilize algorithms and models to process and convert text or speech in one language into equivalent text or speech in another language.

Examples of machine translation apps include Google Translate, Microsoft Translator, and Menulance. However, machine translation engines have inherent weaknesses that make traversing languages challenging. For instance, they often produce outputs consisting largely of word-for-word translations.

Additionally, machine translators do not verify their work, lacking a pause-and-repeat function to allow them to review a phrase more than once. Furthermore, they struggle with context, making it difficult to understand how a mistranslated word or phrase could alter the meaning of a passage in different contexts.

Despite their shortcomings, machine translation is here to stay.

- 610M people use Google Translate daily
- 1B people have used DeepL's services
- Bing has 100M daily active users
- Lionbridge has translated 3.5B words
- Yandex app gets 40K downloads per month
- language translation apps worth US$37B by 2033[17]

Challenges in mainstream translation

Mainstream language translation tools, such as Google Translate and Bing Translate, often struggle to accurately translate specialized terms, particularly within domains like culinary arts. The inherent flaw in their approach arises from training on vast datasets that encompass both general and specialized dictionaries, leading to inaccuracies and semantic confusion.

In the culinary dictionary, these inaccuracies can result in humorous and sometimes harmful translations.

kid on a spit

Machine translation failed to translate *kid* using the culinary dictionary, instead favoring the general dictionary.

nun fart

Machine translation failed to translate the French dish *Pet de Nonne* as a named entity and instead treated it as a grammatical dish.

In 2015, Bored Panda published a post titled “80 of the Funniest Menu Translation Fails Ever,” citing examples too rude to publish here. Read them at your own peril.

Culinary translation presents unique difficulties. Culinary terms and ingredients often have cultural nuances and regional variations. Machine translation engines may struggle to capture these subtleties accurately.

Some food items or cooking techniques may not have direct equivalents in other languages, making it challenging for the translation engine to convey the precise meaning. Culinary writing often involves wordplay, creativity, and descriptive language. Translating these elements while maintaining the intended flavor and style is a nuanced task.

PEACHES STUFFED WITH AMARETTI

CORE

Genesis: starter dough of food translation

At the core of Menulance lies Genesis, a meticulously curated training dataset featuring 166 million words and phrases related to culinary arts available in English, Spanish and Italian. Beyond its sheer volume, Genesis possesses unique mathematical properties, functioning as a computational database rigorously verified for optimal efficiency.

In a machine translation test conducted from October 2022 to January 2023, we assessed the performance of Genesis, Menulance's training dataset.

We trained three machine translation engines using a validation set of 29,600 English-to-Spanish lines randomly pulled from Genesis. The training was carried out on SMT (Statistical Machine Translation), NMT (Neural Machine Translation), and HMT (Hybrid Machine Translation) engines.

Here are the particulars of the three engines.[18]

- Error detection, Auto-correction, Detection of unknown words
- Text simplification/Dealing with accents
- Real-time search/translate for pre-existing text
- Statistical, Hierarchical, and Neural translation engines
- Maximum likelihood fusion between the engines
- Words and multi-word dictionary look-up
- Synonymic suggestion
- Bag-of-word searches

Translation engines trained on Genesis achieved an impressive 83% BLEU score for English-to-Spanish translation. BLEU (Bilingual Evaluation Understudy) scores are a standard measure used to evaluate the effectiveness of machine translation engines. BLEU is a widely used metric for evaluating the quality of machine-generated translations by comparing them to human-generated reference translations.

A higher BLEU score indicates better alignment with human references. BLEU evaluates translation quality by comparing the n-grams, or word sequences, in the machine-generated translation with those found in the reference translation.

To put Genesis's achievement into context, leading machine translation models like Google Translate or DeepL typically score between 30 to 40 percent in BLEU tests. Professional translators achieve BLEU scores of 60.

The following is an interpretation of BLEU scores.[19]

30 - 40%	Understandable to good translations
40 - 50%	High-quality translations
50 - 60%	Very high-quality, adequate, and fluent translations
> 60%	Quality is often better than human
83%	Machine translation trained on Genesis

Genesis achieves 97.6% accuracy

83% BLEU is an outstanding result. However, we wanted a second opinion.

We took the results of the BLEU test and rescored the three engines using a type of human verification called METEOR. METEOR, short for Metric for Evaluation of Translation with Explicit ORdering, evaluates translations by comparing output BLEU translations with the test dataset and calculating a score based on how closely they match in terms of meaning and structure. It is a type of human verification and can be more accurate than BLEU because it addresses some of the limitations of BLEU, such as sensitivity to synonyms and word order.

BLEU has its strengths, especially in large-scale evaluations where manual assessment might be impractical. It provides a quick and objective measure of the dissimilarity between translations. However, it may not capture certain aspects of translation quality that human evaluators can discern, such as stylistic nuances, idiomatic expressions, or cultural appropriateness.

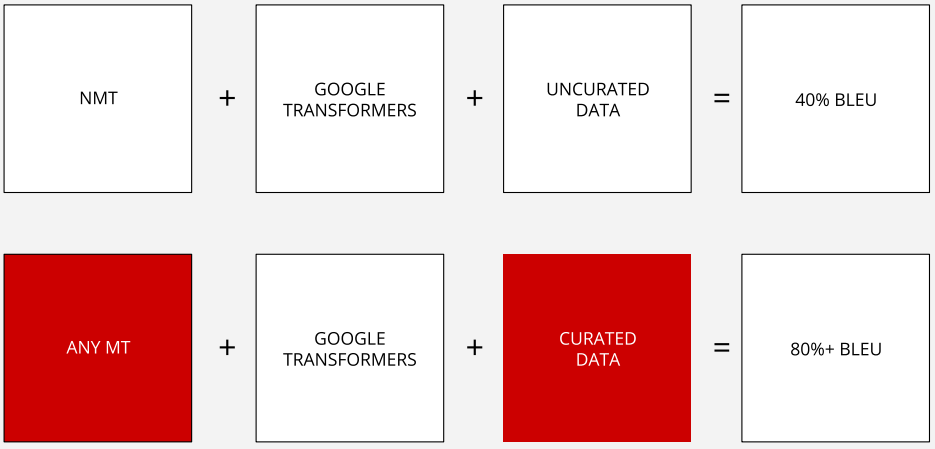
Over the course of two months, researchers manually verified the results of each engine, scoring 88,800 lines of BLEU output. The results were startling. METEOR found that 97.6% accuracy was a more realistic score for machine translation engines trained on Genesis.

The results of each manual assessment are below. Combining quantitative metrics like BLEU with human judgment ensured a more nuanced and holistic approach to evaluating translation quality.

	SMT	NMT	HMT
Accuracy of machine translation engines trained on Genesis	97.33%	96.97%	97.41%

Blame the fuel, not the engine

Both BLEU and METEOR affirm the same finding—machine translation engines themselves are not inherently flawed; rather, the accuracy of their translations is intricately tied to the quality of the training data they are exposed to. Using the same engines that run other mainstream machine translation apps, we can achieve higher accuracy by focusing on the *fuel, not the engine.*



There is a common misconception that dumping uncurated or raw data into artificial intelligence is sufficient for training and achieving unguided, robust performance. This is an oversimplification and misses the critical role of data curation and the need for a deep understanding of AI principles.

The result is engines that struggle to navigate the complexities of language, leading to suboptimal translations. To address the root cause of low accuracy in machine translation, it is imperative to emphasize the significance of high-quality, diverse, and well-curated training data.

Here is a summary table of what we discovered:

- Machine translation engines trained on poorly curated training data achieved *low* BLEU scores
- Machine translation engines trained on Genesis achieved *high* BLEU scores
- Machine translation engines trained on curated datasets scored *similarly* BLEU scores

TIRAMISU TACOS WITH MARSALA MOUSSE

HOW

8 best practices learnt curating Genesis

What follows are 8 best practices we learnt curating Genesis. If you are embarking on an MT project, these lessons are invaluable.

1 LIMIT TRAINING TO A SINGLE DOMAIN

General-purpose translation models, trained on a wide range of topics, often encounter difficulties interpreting context-specific meanings. For example, the word *kid* typically means *child*, but in culinary contexts, it usually refers to a *baby goat*. Due to statistical bias, general models are prone to choosing the most frequently encountered meaning, incorrectly translating food dishes containing specialized terms, like misinterpreting *kid in wild mushroom sauce* as involving a child.

To mitigate such ambiguity, Menulance curated Genesis, a dataset focused exclusively on culinary vocabulary, which comprises fewer than 10% of the entire English lexicon. This focused approach ensures the translation engine accurately recognizes specialized food terms and context-specific meanings, dramatically reducing errors caused by semantic ambiguity.

Research supports this domain-specific approach. Studies by Koehn and Knowles (2017)[20] highlight that NMT models trained on domain-specific data achieve higher BLEU scores and greater fluency than general models trained across diverse topics. Their findings emphasize that restricting training data to a specific domain enhances translation quality and minimizes confusion from irrelevant contexts.

Ultimately, the success of machine translation relies not only on data quantity but also on its quality, specificity, and relevance. Carefully curated datasets like Genesis allow machine translation engines to deliver precise, reliable, and contextually accurate translations.

2 TRAIN USING SHORT SENTENCES

Genesis is trained in food dishes and ingredients—a domain naturally suited to short sentences. If culinary language typically involved longer sentences or paragraphs, curating Genesis would have been significantly more complex since lengthy sentences introduce greater complexity and a higher risk of errors when training machine translation engines.

When a sentence is short, it has fewer words that make up its meaning. People naturally find short sentences easier to understand because they have fewer words to figure out.[21] This principle extends to training machine translation engines; short sentences are simpler to learn and translate than longer ones.

The average sentence length in the Genesis dataset is around 11 words. General-purpose machine translation engines often train on hundreds of words. They are designed to handle paragraphs, leading to overly complex algorithms that must cater to *long-range dependencies,* where meaning extends across multiple sentences.

A culinary example of a long-range dependency might appear in a recipe or food review spread across multiple sentences. Consider the following:

“Gently sauté the onions and garlic in olive oil until translucent. Set them aside. In the same pan, brown the chicken thighs. Return the onions and garlic to the pan, add tomatoes and simmer until the chicken is tender.” Here, *them* in the second sentence and *the onions and garlic* in the fourth sentence refer back to the ingredients mentioned at the very beginning.

Amazon Alexa, the voice-controlled virtual assistant, employs short sentence training and canonical decomposition-like techniques to enhance its language understanding capabilities. By breaking down user commands into smaller units and focusing on short, contextually rich sentences, Alexa can accurately interpret and respond to a wide array of user queries and commands.[22]

3 AVOID TRAINING USING PUBLIC DATA

High-resource languages like English, Spanish, and Mandarin benefit from vast amounts of linguistic data,[23] including extensive parallel corpora (collections of aligned texts in both the source and target languages) and comprehensive linguistic resources (well-annotated corpora, grammars, and dictionaries), which all enhance translation quality.

In contrast, low-resource languages like Haitian Creole, Maltese, and Kinyarwanda face significant challenges due to the scarcity of training data and linguistic resources. This lack of data often leads to lower accuracy in machine translation outputs, as these languages do not benefit from the same robust parallel corpora or comprehensive reference materials available for high-resource languages.[24]

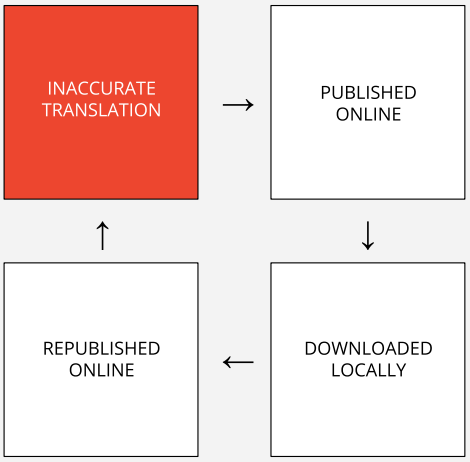
A study by UCLA Medical Center in 2021 highlighted that Google Translate's accuracy ranged from 55% to 94%, depending on whether the target language was a high- or low-resourced language.[25]

	BLEU score
Spanish	94%
Mandarin Chinese	81.7%
Farsi	67.5%
Armenian	55%

The correlation between high-resource languages and accurate training data, while true in a general sense, did not aid the curation of Genesis. Despite Genesis being trained in high-resource languages (English, Spanish, and Italian), we faced significant challenges because online culinary information was unreliable and full of inaccuracies.

In our experience, whether the language was high- or low-resourced was irrelevant—the available culinary data was flawed and unfit for Genesis.

The common practice of using these inaccurate culinary translations to train mainstream machine translation engines has resulted in a self-reinforcing feedback loop where inaccuracies are continually published, downloaded, and republished in the public domain.



Once an inaccurate translation is introduced, its repetition in multiple media leads to an illusion of correctness simply by virtue of its frequency. Over time, these repeated errors become entrenched as *truth* because they are encountered so often that users and even subsequent machine translation systems begin to accept them without question.

Today, inaccurate food translation exists in all manner of media: restaurant menus, food labels, e-commerce sites, mainstream dictionaries, niche dictionaries, blogs, and printed books.

This vicious cycle illustrates how critical data curation is, as breaking the cycle requires ensuring that each piece of training data is accurate and verified, preventing errors from propagating through Genesis. No matter which machine translation engine is used—SMT, HMT or NMT—if trained on poorly curated datasets such as these, the engines will struggle to produce accurate translations.

Recognizing that no machine translation engine can perform well with such dirty data, we embarked on the painstaking task of manually curating our own dataset from scratch. Starting with an initial collection of 140,000 sentences in 2012, we discovered that 40-50% of online culinary information was inaccurate and had to be manually fixed by professional translators.

Our focus was on precision and quality rather than merely relying on the volume of data. This approach is aligned with best practices in other industries; for example, Waymo, Alphabet's self-driving car subsidiary, has spent over a decade continuously

collecting and meticulously annotating sensor data to train its autonomous driving algorithms.

Similarly, IBM Watson for Oncology has achieved success by training on carefully curated datasets composed of medical literature, clinical trial data, and expert insights. Tesla's Autopilot system also owes its real-world effectiveness to the continuous refinement of massive amounts of high-quality, annotated data gathered from its fleet, which ensures safety and performance.[26]

4 GATHER DENSE TRAINING DATASETS

Big datasets are crucial for training machine translation engines, with many experts suggesting that 20 million examples is a good starting point. However, data density is equally critical. Simply having a large dataset isn't helpful if it repeats the same translation pair millions of times. A well-curated training dataset requires dense, balanced, and diverse data to ensure comprehensive learning.

Dense, massive datasets also address a common issue faced by machine translation engines: overfitting. Overfitting occurs when a machine translation engine memorizes its training data and fails to generalize beyond it. A machine translation engine that overfits cannot handle translation requests outside its training set, causing it to generate inaccurate translations.[27]

Consider a Tesla self-driving car trained only on a single road in Los Angeles. Would you trust it to drive on unfamiliar roads in Australia? Certainly not. The vehicle must experience thousands of different roads across numerous cities, varying traffic patterns, and diverse weather conditions before it can reliably generalize its driving capabilities. Similarly, a machine translation engine must learn from a broad, dense dataset to translate accurately and effectively handle phrases it has never previously encountered.

Dense datasets are imperative because they provide diverse and high-quality training examples for the translation model. These datasets are rich in linguistic variations, covering a wide range of food styles, contexts, and syntax. The richness of dense datasets allows machine translation models to capture the complexities of language more effectively, resulting in higher translation quality and improved fluency in the translated text.

The original Genesis dataset of 4M lines was sparse and unevenly balanced. To resolve this, we expanded the dataset using a structured approach. This allowed us to systematically build a substantially larger and denser dataset of 166M lines.

For instance, we produced variations like:

braised crocodile in marmalade sauce,
braised duck in marmalade sauce, and
braised pork in marmalade sauce.

This increased density meant the machine translation engine encountered numerous similar yet distinct sentence patterns, greatly improving its ability to generalize.

The expression *sponsali con prosciutto* is often mistranslated by mainstream engines as *wedding with ham* because the term *sponsali* refers predominantly to *wedding* instead of *chives* on the web. When translating the term *sponsali,* the word *wedding* will have a higher chance to be selected instead of chives, and this will lead to a contextually wrong culinary translation. One way to avoid this problem is to increase the number of italian-english sentence pairs with simultaneous appearances of the words *sponsali-chives,* increasing the chances that *sponsali* gets effectively translated into *chives*.

While large datasets are necessary, their density and diversity are what truly enhance the performance of machine translation models. Ensuring the data is well-curated and representative of various linguistic nuances is key to achieving superior translation accuracy.

5 IDENTIFY DEGREES OF FREEDOM WITHIN TRAINING DATA

Gathering dense training datasets is critical, but equally important is determining precisely which data points should be collected. This selection depends on clearly defining the dataset's degrees of freedom (DoF), which represent the specific variables and variations that affect the outcomes an AI or machine translation model must handle.

For instance, during the development of the Jijo-2 facial recognition project, we identified nine critical DoF, including the type of camera used, its position, lighting conditions, 3D poses, and subject-specific attributes such as age and facial expression. Each of these variables influenced data collection significantly, ensuring that the dataset comprehensively covered real-world variations and allowed our AI models to generalize effectively.

Conversely, in a project I completed with Unilever in 2019, recognizing shampoo and conditioner bottles required fewer DoF. Factors such as lighting conditions, bottle orientation, and product placement were essential, but facial expressions or camera types were irrelevant. Clearly defining the DoF allowed us to focus on the necessary data, increasing the accuracy and efficiency of the AI recognition model.

For a Tesla self-driving car we spoke of earlier, DoF within the training data include various environmental and situational factors that influence driving. These encompass road types (highways, residential streets, dirt roads), road conditions (wet, icy, dry), traffic density (low to congested), lighting (daytime, dusk, night, artificial lighting), weather variations (sunny, rainy, snowy, foggy), obstacle types (pedestrians, cyclists, animals, vehicles, construction zones), as well as geographic and regional differences (road signs, driving conventions).

Capturing comprehensive data across these diverse scenarios ensures that the vehicle can safely and reliably navigate situations it hasn't explicitly encountered during training.

	DoF
Tesla Self-Driving Car	Road types, road conditions, traffic density, lighting, weather variations, obstacle types, geographic and regional differences.

Jijo-2 AI Model	Camera type, camera position, lighting conditions, 3D poses, age, facial expression, other subject-specific attributes.
Menulance's Genesis	Ingredients, dish names, categories of food, adjectives, gender.

In machine translation, particularly within specialized domains such as culinary language, DoF becomes critical in capturing linguistic variations effectively. For example, culinary datasets could include variations in ingredients, cooking methods, presentation styles, dish names, regional terms, origin, and descriptive nuances. Without clearly defining these DoF, datasets may lack the variability necessary to train translation engines accurately, limiting their effectiveness when encountering new or slightly altered inputs.

To illustrate, consider translating a dish such as *braised crocodile in marmalade sauce.* DoF here could include cooking methods (braised, grilled, roasted), meat types (crocodile, alligator, chicken), and sauce variations (marmalade sauce, honey glaze, peppercorn sauce).

Koehn notes that clearly identified DoF enable the efficient expansion and management of datasets, significantly improving a model's adaptability to domain-specific nuances and reducing semantic inaccuracies (Koehn & Knowles, 2017).[28]

Moreover, clearly identifying DoF supports scalability and future-proofing. If the relevant variables are known from the start, expanding or refining the dataset to accommodate new conditions, products, or contexts becomes straightforward.

6 EXTRACT NAMED ENTITIES FROM TRAINING DATA

Machine translation models rely heavily on the frequency of words appearing in their training data, making them vulnerable to inaccuracies when encountering rare, specialized, or newly coined terms.[29] This issue is especially significant within specialized domains such as culinary translation, where unique dish names and technical terminology are prevalent.

We addressed this challenge by systematically replacing or adding named entities in Genesis. Named entities are dishes known globally, often retaining their original names across languages, such as *Spaghetti alla carbonara.* Whether the language is Spanish, Italian or English, it remains *Spaghetti alla carbonara.*

Mainstream translation engines often do not curate dishes like *Spaghetti alla carbonara* correctly as a named dish, leading to overly literal or nonsensical translations. For example, the Italian phrase *alla carbonara* literally means *in the style of the charcoal burner.* Without context, a mainstream machine translation engine might translate *Spaghetti alla carbonara* into something like *Charcoal Burner's Spaghetti* or *Spaghetti with Coal,* losing the culinary significance and appetizing nature entirely.

When adding a dish such as Spaghetti alla Carbonara, we perform two critical tasks in Genesis:

- Fragment the sentence into subsentences, such as *spaghetti* and *alla carbonara,* which aids the model in understanding individual elements.
- Preserve the complete dish Spaghetti alla Carbonara as a single atomic entity.

This approach significantly enriches the dataset, enhancing the machine translation engine's capability to translate accurately by increasing exposure to slight variations in recognizable culinary terms.

Today, Genesis contains tens of thousands of such named entities, strengthening the model’s accuracy through this structured categorization and enrichment process.

Conversely, dishes such as *Braised crocodile served with marmalade sauce* are typically localized, descriptive, and treated as a grammatical dish. Unlike named entities, these dishes rely heavily on translating each descriptive subsentence. By fragmenting these dishes, the machine translation model separately analyzes each term's meaning and grammatical relationships, ensuring the translated phrase accurately conveys the dish's intended meaning and culinary context, even if the entire phrase is unfamiliar to the target audience.

Research supports this structured approach to named entities and grammatical parsing.[30] According to Koehn and Knowles (2017), clear domain-specific categorization and controlled vocabulary expansion significantly improve NMT outcomes.

Similarly, a recent work by Sennrich and Haddow (2016) highlights the effectiveness of subword segmentation and systematic tagging, which help models better handle rare or previously unseen terms by breaking down and categorizing complex lexical items.[31]

Furthermore, extracting named entities greatly improves translation consistency. Consistent naming of dishes and ingredients is essential in culinary branding, marketing, and user experience, especially across multilingual menus or cookbooks. Properly named entity recognition ensures that dish names remain uniform throughout translated texts, building familiarity and trust among consumers.[32]

7 IMPROVE WORD ALIGNMENT WITH VALID SUBSENTENCES

Adding verified subsentences to Genesis enhanced word alignment accuracy in instances where an SMT engine is used. Word alignment is a fundamental step in SMT that ensures words and phrases in one language correctly correspond to their most accurate equivalent in another language. It functions similarly to a bilingual dictionary, where each word or phrase in a source language is systematically mapped to the most probable translation in the target language.

For example, in a parallel corpus (a dataset of bilingual sentences), word alignment guarantees that *apple* in English correctly matches *mela* in Italian.

To achieve precise word alignment, SMT engines analyze large datasets and automatically generate phrase pair tables during training. These tables record how frequently a phrase in one language appears with a phrase in another, assigning probability scores based on their statistical co-occurrence. When an SMT engine translates a new sentence, it refers to these probability scores, selecting the most probable word or phrase match to ensure accurate translation.

A dish like *Braised crocodile in marmalade sauce* does not have an established, universally accepted translation. Since the system has never seen this phrase before, it fragments the sentence into subsentences, assigning probability scores based on how frequently each phrase pair appears in the training data:

source (English)	target (Italian)	probability score
braised crocodile	coccodrillo brasato	0.90
braised crocodile	coccodrillo stufato	0.10
in marmalade sauce	in salsa di marmellata	0.85
in marmalade sauce	in salsa alla marmellata	0.15
braised crocodile in marmalade sauce	coccodrillo brasato in salsa di marmellata	0.80

braised crocodile in marmalade sauce	coccodrillo stufato in salsa alla marmellata	0.20

However, named entities, such as S*paghetti alla carbonara,* appear in phrase tables as fixed, high-confidence translations. Because this dish is recognized in Genesis as a named entity, its phrase pair receives a probability score of 1.00:

source (English)	target (Italian)	probability score
spaghetti alla carbonara	spaghetti alla carbonara	1.00

Each phrase pair probability is determined by data frequency—a higher probability (closer to 1.0) means a more common and reliable translation. Since *Spaghetti alla Carbonara* is a named entity, it achieves 100% accuracy. However, *Braised crocodile in marmalade sauce* is less common, leading to a lower probability (0.80), reflecting its rarity in the dataset.

One major challenge in SMT is *noise* caused by invalid subsentences. SMT engines automatically extract phrase pairs, but this process is not perfect—many meaningless or incorrect phrase pairs can be introduced.

For example, if an SMT model encounters the sentence:
Wine of Australia and caviar

It may mistakenly extract the phrase pair:
Australia and caviar → Australia e caviale

While grammatically correct, this phrase has no real-world culinary meaning. However, because SMT engines lack contextual awareness, they treat it as a valid translation, which then pollutes future translations.

This noise distorts probability calculations, increasing the risk of incorrect translations. Over time, repeated exposure to invalid subsentences can bias the system toward producing nonsensical outputs, further reducing translation accuracy.

We systematically removed invalid subsentences from Genesis to improve the performance of SMT engines. This was only feasible because Genesis is trained using

short sentences—if the dataset contained long paragraphs, verifying each fragment would be nearly impossible.

Other phrase-based machine translation approaches also use phrase pair tables, but their reliance and approach vary compared to SMT. Phrase-based Machine Translation (PBMT), an evolution of SMT, heavily relies on phrase pair tables. It segments sentences into phrases rather than just individual words and aligns these across languages. Adding verified subsentences improves PBMT's ability to piece together translations correctly. Like SMT, PBMT benefits significantly from high-quality phrase tables, as cleaner data enhances translation accuracy.

HMT combines rule-based (RBMT) and statistical (SMT/PBMT) techniques. Some HMT models use phrase pair tables alongside linguistic rules to provide more structured translations. The extent to which phrase tables are used in HMT depends on the balance between statistical learning and rule-based logic within the system.

Unlike PBMT and HMT, NMT does not use phrase pair tables. Instead, it relies on deep learning to translate entire sentences holistically. However, if the training data lacks valid subsentences, NMT models may hallucinate incorrect translations due to weak contextual understanding. While NMT does not directly use phrase tables, it indirectly benefits from subsentences, as higher-quality data reduces noise during training, leading to improved translation accuracy.

8 MEASURE TRAINING DATA BY REGULARLY USING METEOR

BLEU remains the standard for measuring machine translation accuracy. BLEU is calculated as an average over a large test set, giving a reliable score because it rewards the words with the highest probability of being used.

However, BLEU is not without limitations. BLEU evaluates translations based on matching n-grams (word sequences) with those found in the reference translation. If a rare word in the machine-generated translation has a synonym in the reference, but it's not an exact match, BLEU may penalize the translation even if the meaning is preserved.

BLEU combines 1-gram, 2-gram, 3-gram and 4-gram into a statistic that scores the difference between the expected output to the predicted output. If the same words are present, then you get a higher score. A synonym is a morphologically (very) different word, so you will have 1-gram, 2-gram, 3-gram, 4-gram mismatch, so a lower BLEU because the words are morphologically different.

Addressing this limitation is challenging because BLEU lacks semantic understanding and relies on exact word matches. Researchers and practitioners are aware of this issue and often consider it when interpreting BLEU scores. It underscores the importance of using multiple evaluation metrics and human judgment to obtain a more nuanced assessment of machine translations.

There are several other means of measuring machine translation including METEOR. BLEU is a necessary standard for transparent and reproducible machine translation evaluation tasks and that’s why it is used in the international competitions: independent agencies like NIST can rerun the experiment on the trained engines and confirm the scores.

However, a better measure is METEOR—a type of human verification. Data scientists all over the world perform human verification tasks across the AI fields, from machine translation to Computer Vision. We conducted a manual verification on 29,600 test sentences and their expected output. While BLEU gave us 83%, the real accuracy was 97.6%, and it was due to the presence of synonyms.

CONCLUSION

It is important to understand that machine translation is a tool, not a substitute for human translation. The relationship between human translators and machines is far more symbiotic than competitive. 610M people use Google Translate each day—professional translators included.

Machine translation is valuable, providing a quick initial translation that human translators can then refine and enhance. It can help speed up the translation process, handle large volumes of content, and serve as a resource for translators to reference. However, the final output often benefits from the careful review and adaptation that human translators provide.

Achieving high-quality machine translation, particularly in a specialized domain like culinary language, is not about using the most advanced engine—it's about using the right data.

Genesis proves that carefully curated, domain-specific, and intelligently structured training data can produce machine translation systems that outperform uncurated data by a wide margin. In the culinary domain—where meaning, culture, and trust are everything—this isn't just a best practice. It's a requirement.

VIENNA SAUSAGE SHORTCAKE

APP

The world’s first AI that translates what’s on your plate

ENDNOTES

1. *JETIR*, Volume 10, Issue 11, "Impact of Digital Innovations and AI on Gastronomy, Tourism, and Local Food," Khilesh Patel and Dr Sunil Kumar, Nov 2023.

2. *Food: A Culinary History from Antiquity to the Present,* A. Sonnenfeld, 1999, xvi.

3. *The Translator*, "What's cooking in English culinary texts? Insights from genre corpora for cookbook and menu writers and translators," Michał B. Paradowski, Jan 2018.

4. *The ATA Chronicle*, "What's Cooking: An Introduction to Culinary Translation," July 2021.

5. *The Translator*, Vol 21, issue 3, "Food and translation, translation and food," Delia Chiaro and Linda Rossato, Dec 2015.

6. Market Data Forecast, *Culinary Tourism Market Research Report,* Jun 2024.

7. Computer Vision Foundation, "Inverse Cooking: Recipe Generation from Food Images," Amaia Salvador, Michal Drozdzal, Xavier Giro-i-Nieto, Adriana Romero, Dec 2018.

8. SBS News, "How food porn hijacks your brain," Christian Jarrett, Oct 2015.

9. Marketplace, "What is it about cookbooks?" Samantha Fields, Jan 2023.

10. GITNUX, "Must-Know Book Sales By Genre Statistics," Dec 2023.

11. *The Guardian*, "What's the point of cookbooks? Hope, love and beauty (but not cooking)," Kate Gibbs, Nov 2022.

12. Marketplace, "What is it about cookbooks?" Samantha Fields, Jan 2023.

13. Ibidem.

14. Ibidem.

15. Ibidem.

16. Ibidem.

17. FactMR, *Translation software outlook 2023-2033,* Jan 2023.

18. Menulance, Executive Summary, May 2023.

19. Google Cloud, "Evaluation models."

20. Department of Computational Linguistics, University of Zurich, Switzerland, "Domain Robustness in Neural Machine Translation," Mathias Muller, Annette Rios, Rico Sennrich, Oct 2020.

21. *Computational Linguistics*, Volume 48, Number 2, "Challenges of Neural Machine Translation for Short Texts," Yu Wan, Dec 2021.

22. Alexa Developer Documentation, "Best Practices for Sample Utterances and Custom Slot Type Values," Mar 2024.

23. Portland Community College, "The 4 Most Common Languages Spoken Around the World," Rebecca Raymond, June 2019.

24. Naarg Data Media Services, "Top 7 challenges in machine translation."

25. Phrase.com, "Research vs Practice: How Accurate Is Google Translate?" Nov 2023.

26. Bhumi Publishing, *Cutting-Edge Technology From Tesla,* Zaiba Khan, Oct 2024.

27. Wikipedia, "Hallucination (artificial intelligence)."

28. Association for Computational Linguistics, *Proceedings of the First Workshop on Neural Machine Translation,* Thang Luong, Alexandra Birch, Graham Neubig, Andrew Finch, Aug 2017.

29. LinkedIn.com, "Training AI to Translate: The Promise and Challenges of Machine Translation," Jesse Anglen, Sept 2023.

30. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, "Using Confidential Data for Domain Adaptation of Neural Machine Translation," Sohyung Kim, Arianna Bisazza, Fatih Turkmen, 2021.

31. *Artificial Intelligence Review*, "Revisiting named entity recognition in food computing: enhancing performance and robustness," Uchenna Akujuobi, Shuhong Liu, Tarek R. Besold, Aug 2024.

32. *Language Resources and Evaluation*, "ATR4S: Toolkit with state-of-the-art automatic terms recognition methods in Scala," Nikita Astrakhantsev, Dec 2017.

ABOUT THE AUTHORS

Dr Roberto Mariani, a former scientist turned tech entrepreneur, holds a PhD in Physics from the Institut Géographique National de Paris and the University of Rouen. With 30 years of experience in AI, he founded and led numerous ventures in video analytics, facial recognition, and machine translation. Roberto holds nine patents, symbolizing his pursuit of innovation.

Anthony Coundouris brings over two decades of experience in consulting for startups and holds a degree in business from the University of Technology, Sydney. As a two-time entrepreneur himself, he specializes in designing automated sales and marketing systems. Anthony is the author of the book *run_frictionless: How to Free a Founder from a Sales Role* and the creator of a decision-making framework called the 4Qs.

ABOUT THE TRANSLATOR INTO ITALIAN AND SPANISH

Dr. Frank Badrines, born in Barcelona in 1975, received his PhD in Philosophy from the University of Barcelona with a thesis on the influence of Plato and Aristotle on C. S. Lewis. He is currently a lecturer in Spanish language at the National University of Singapore and also works as a translator of Modern Greek fiction in Spain. He joined Menulance in 2013.

ABOUT US

Menulance is the first AI-powered menu translation app, trained on millions of verified terms. Discover and translate food dishes and ingredients in Spanish, Italian and English.

media@menulance.com
www.menulance.com

