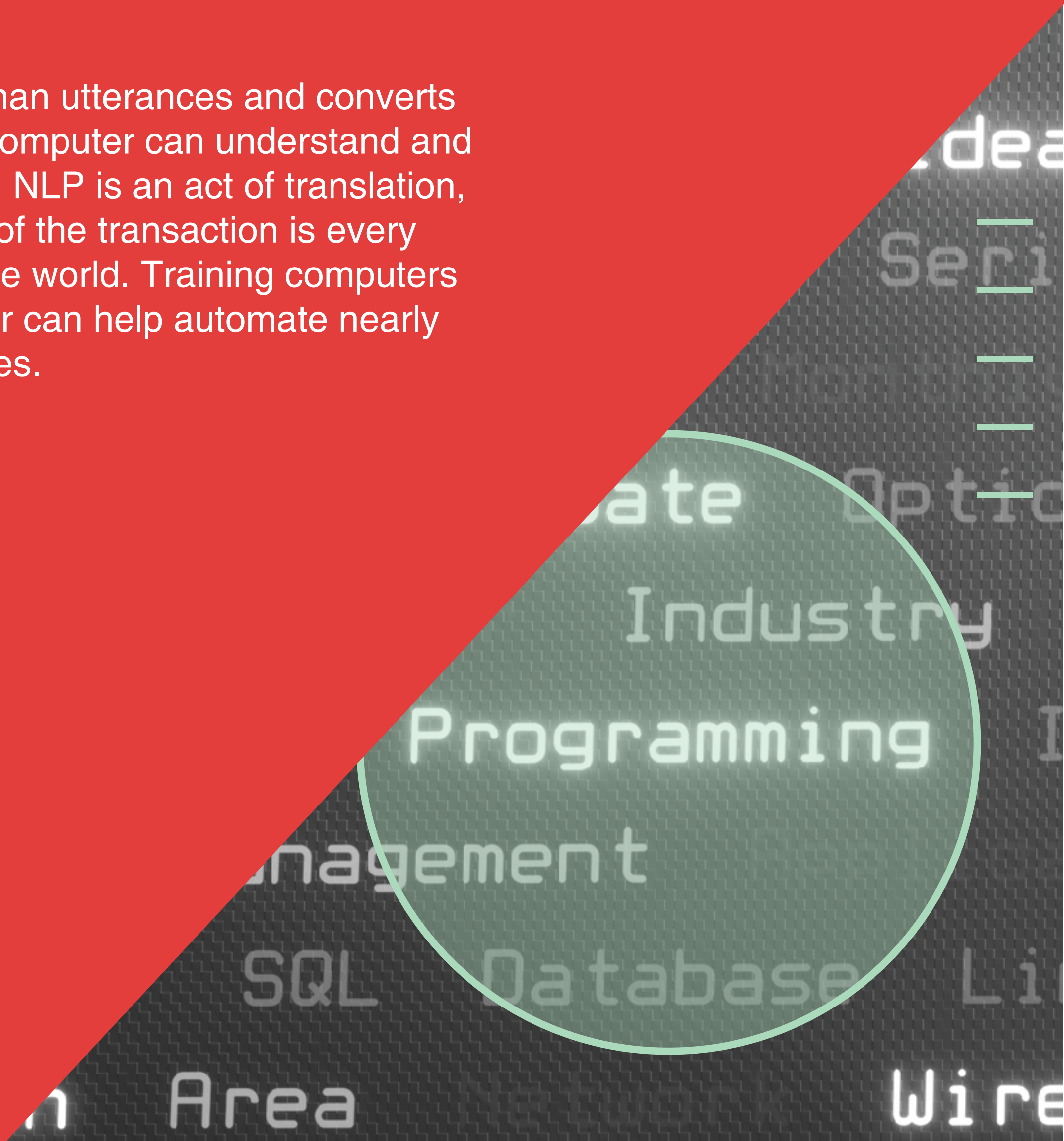
WHITE PAPER

iMerit

# ANNOTATING DATA
## FOR NATURAL LANGUAGE PROCESSING

NLP takes natural human utterances and converts them into data that a computer can understand and respond to. At its core, NLP is an act of translation, where the beneficiary of the transaction is every computing device in the world. Training computers to understand us better can help automate nearly every aspect of our lives.



# Introduction

*“Hey Alexa, can you set an alarm for 6:30 AM?”*
*“Alarm set for 6:30 PM”*

Alexa, Google Assistant and Siri are members of many households today. The ubiquitous voice assistants can listen to your requests and respond with suggestions or even tell jokes to keep you entertained.

These are perhaps the most recognizable applications of Natural Language Processing (NLP). NLP has existed for decades in academia, but has become more visible in recent years with the increased adoption of Machine Learning and Deep Learning techniques. NLP and its associated processes and concepts form the backbone of many applications that aim to mimic or augment human interactions.

NLP takes natural human utterances and converts them into data that a computer can understand and respond to. At its core, NLP is an act of translation, where the beneficiary of the transaction is every computing device in the world. Training computers to understand us better can help automate nearly every aspect of our lives.

NLP performs its magic during the nanoseconds between when a command is sent to Alexa and a response is created by it. The voice command is simply a sequence of acoustic information, but with natural language understanding, Alexa abstracts over this information to identify meaningful units of sound, group sounds into words, group words into grammatical and semantic units, and connect these units to a set of concepts to "understand" them. To act on this understanding, Alexa then performs the correct function, constructs an appropriate response, and synthesizes the voice response, repeating the process somewhat in reverse.

Now that the flow of natural communication has been opened up between humans and computers the possibilities are infinite. A quick scan of the technology around us reinforces the point: NLP has entered our daily lives in the smallest of ways. The past few years have seen path-breaking developments in the linguistic space. One breakthrough example is Google Duplex, a smart chatbot that can carry out routine tasks like making appointments or reservations over the phone. Another less flashy but more ubiquitous example is the auto-suggestion and auto-completion of responses while emailing and texting.

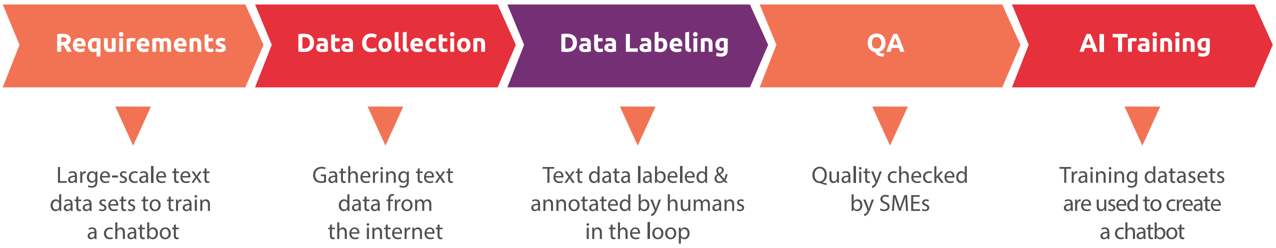
This trend shows no sign of slowing down. The NLP market is poised to grow from $4 billion in 2019 to $22.3 billion in 2025. That said, the challenges of developing successful NLP algorithms are enormous. The biggest one is that humans converse, write, and express their thoughts in an infinite number of ways, which are highly context and culture-specific.

No language is uniform across the world. Each geography has developed a dialect of its own with unique slang, cultural references and cadences of speech. "What's up?" in the US becomes "What la?" in Singapore. Humans can intuit the meanings of words through context and other indicators even in unfamiliar language situations and can recalibrate quickly. A computer cannot… unless it is explicitly trained to do so.

If a model is not taught to understand these nuances of human communication, utterances can get lost in translation resulting in ineffective algorithms. Accurate and well-structured training data, to enable supervised learning, can be the differentiator in the NLP space.

Text data is available everywhere. The key lies in selecting the most relevant text types for a specific use case and carefully annotating them to build large-scale training datasets that can power a linguistic algorithm.

### How Machines Learn Language



Requirements
Large-scale text data sets to train a chatbot

Data Collection
Gathering text data from the internet

Data Labeling
Text data labeled & annotated by humans in the loop

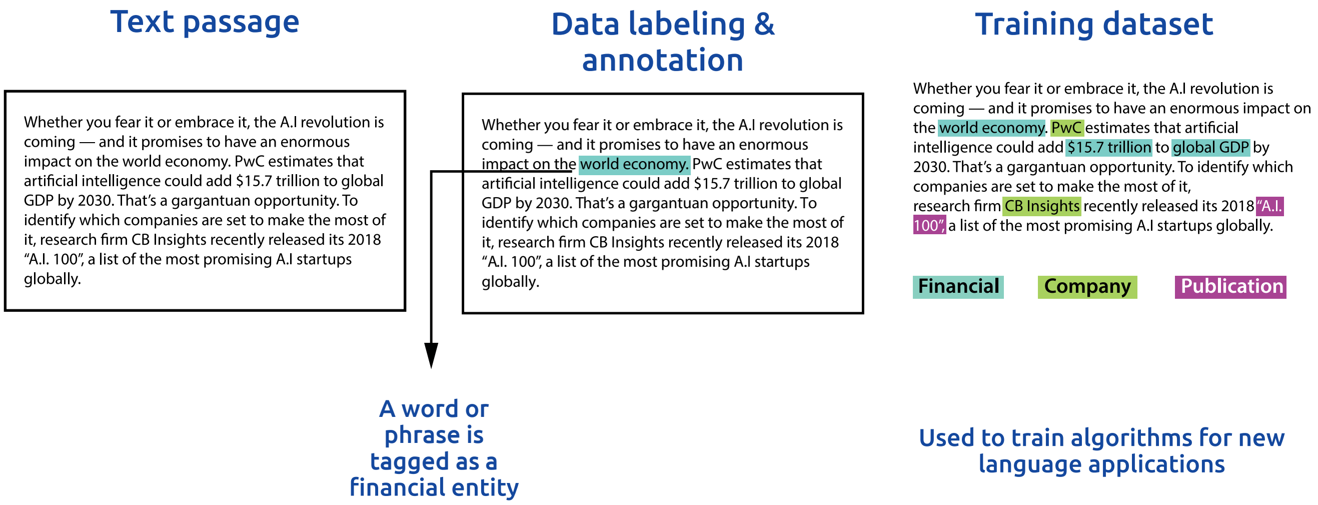
QA
Quality checked by SMEs

AI Training
Training datasets are used to create a chatbot

The building of a chatbot is a technical and complex process. The first steps involve teaching an algorithm to understand human language, specific to the requirements of the project. A chatbot can help a user order pizza or set up a demonstration of a SaaS product. Once the purpose of the chatbot is identified, huge datasets of utterances are compiled from suitable sources. These utterances have to be tagged for the what (entities), why (intent), how (actions involved), and when (time of the action) of the specific project.

This is where data labeling and annotation enter the picture. Expert labelers carefully read and comprehend vast text datasets, and annotate them as per project guidelines. Once the quality of the datasets has been measured and stamped as acceptable they are used to train the chatbot algorithm and shape how the bot will interact with users.

## Where Data Labeling Fits in



Text passage

Whether you fear it or embrace it, the A.I revolution is coming — and it promises to have an enormous impact on the world economy. PwC estimates that artificial intelligence could add $15.7 trillion to global GDP by 2030. That's a gargantuan opportunity. To identify which companies are set to make the most of it, research firm CB Insights recently released its 2018 "A.I. 100", a list of the most promising A.I startups globally.

Data labeling & annotation

Whether you fear it or embrace it, the A.I revolution is coming — and it promises to have an enormous impact on the world economy. PwC estimates that artificial intelligence could add $15.7 trillion to global GDP by 2030. That's a gargantuan opportunity. To identify which companies are set to make the most of it, research firm CB Insights recently released its 2018 "A.I. 100", a list of the most promising A.I startups globally.

A word or phrase is tagged as a financial entity

Training dataset

Whether you fear it or embrace it, the A.I revolution is coming — and it promises to have an enormous impact on the world economy. PwC estimates that artificial intelligence could add $15.7 trillion to global GDP by 2030. That's a gargantuan opportunity. To identify which companies are set to make the most of it, research firm CB Insights recently released its 2018 "A.I. 100", a list of the most promising A.I startups globally.

Financial Company Publication

Used to train algorithms for new language applications

# NLP Functions & Data Types

Named Entity Recognition

Named Entity Linking

Sentiment & Intent Analysis

Salience Analysis

## Named Entity Recognition and Linking

In all communication we constantly reference entities like proper names, entities of interest (companies, animals, diseases), and locations. We assume that the listener also knows these entities. With a computer these basic assumptions are thrown out the window. We have to provide a roadmap for these words to the algorithm. This is where Named Entity Recognition (NER) steps in. In NER we annotate the passages of text by identifying and classifying these entities of interest. These annotations then go into training an algorithm. Once trained the algorithm can recognize and classify these entities into the desired categories, adding a crucial layer of meaning to text. Entities can be domain-specific or generic.

Named Entity Linking (NEL) is a process of disambiguation by associating an ambiguous entity to its correct definition in a knowledge base. Consider the sentence “Apple stock rose by 2%”. The algorithm must learn that this refers to the company Apple, and not the fruit. The same word may be used differently even in one document. “Nestled among former apple orchards in Silicon Valley, the Apple R&D team is hard at work on the next generation of cameras.”

Named Entity Recognition, combined with auxiliary functions like NEL, arms algorithms with cognition capabilities.

# Of sections and subsections

Journalism deals with volumes of written material and has leveraged the power of NER. A software that can scan articles and extract the entities it discusses can help a publication categorize its material more accurately and thereafter help readers find content more easily.

# A well-deserved vacation

A memorable vacation is the perfect blend of stunning landscapes, exciting local activities and cuisine, and relaxation. As Airbnb moves from being an accommodation platform to an end-to-end travel solution, it is using knowledge graphs to help users plan trips. In a knowledge graph, entities, attributes, and relationships are captured from a dataset and displayed in a graphical form. This is a powerful way of extracting useful insights as this provides context, and many large companies including Google, LinkedIn, and Amazon make use of the technique. In Airbnb’s case, the technique is being harnessed as they scale the platform to include more experiences, restaurants and homes, and the relationship between them helps provide a holistic experience.

# Your call is important to us

Customer support is an exceptional key focus area to any company offering products and services. In today’s world feedback can be shared on the phone, online, and in person, and it’s important to direct users correctly to the right support platforms and service agents. Recognizing product and service names as well as locations can help streamline this process and ensure redressal is swift and satisfactory.

# Sentiment and Intent Analysis

Humans can instinctively understand if a person is happy, sad, angry, or even emotionless while communicating. In personal interactions linguistic cues like intonation, non-verbal cues and body language come to our aid, and while online, emojis, punctuation, and GIFs step in. Context also offers valuable clues.

Teaching a computer to understand this spectrum of human sentiment is a valuable NLP application. Sentiment analysis mines unstructured text datasets and classifies word and phrases as positive, negative, or neutral. This captures the voice of the customer, a valuable commodity for a business.

Datasets categorized by sentiment can be leveraged to drive marketing and public relations campaigns, improve customer service, and build on product or service features. Companies can listen to trending online chatter about their brands and pick up key signals rapidly, prioritizing actionable feedback.

As an example, the restaurant NOLA can take pride in its ambience and customer service, but might need to work on the quality of its entrees.

Positive Neutral Negative

The best things about NOLA are the waitstaff and the atmosphere. It's a lovely space , and whoever trains the staff is a pro: they are friendly and very efficient . The food is just average . I ordered a crab cake, it was virtually flavorless and the cornbread was dry as dust . A standout for us was desert: delicious bread pudding and turtle pie. We will be coming back for the ambiance and deserts.

Human sentiment can also vary at the sentence and paragraph levels, resulting in nuanced and mixed sentiments. Correctly interpreting this is another layer of training.

Take, for instance, this review of a hotel room.

> ***"We had an okay experience at the hotel. The location was really good and the room was nice. The beds were clean and soft, but the bathroom was a little smaller than expected. The towels were not too clean."***

While labeling at the sentence level, "smaller than expected", and "not clean" are tagged as negative sentiments, but even as casual readers, we can immediately understand that the guest had an overall satisfactory experience, even if it was not excellent. We can also relate to the relative importance of a decent room in a good location and the availability of clean towels. A computer algorithm which has never had the experience of staying in a hotel room cannot instinctively see these indicators.

Careful and informed sentiment analysis labeling can guide a computer to gradually recognize these.

Intent analysis takes the capability a step further and tries to identify what the communicator intended to convey: a complaint, a query, a compliment or a suggestion.

# Salience scores

Not all parts of a communication are equally relevant. Assigning a graded salience score helps prioritize the utterances that are most relevant to a subject and context. This can provide for more efficient information retrieval and summarizing by highlighting the salient entities.

The salience litmus test can be applied to articles, social media posts, online forum messages, and any other forms of written communication For example, in an article about China and trade, entities like “China”, “Beijing”, and “Yuan” are very relevant, while “Antarctica”, which is included in the article, is not at all germane to the main subject matter.

# Labeling over 10 million NLP data points - Lessons Learned

## Case study 1

Our customer provides financial reports and insights to investors and decision-makers in private equity. They create content based on large volumes of complex, unstructured data from a variety of text and graphical sources. To streamline their processes they are working on building a contextually-aware AI engine that automates the collection and standardization of these sources, with near-human accuracy. The challenge lies in building a training and validation dataset that filters key information from noise, with a high level of precision.

As fuel for this engine, our dedicated team has performed Named Entity Recognition, Named Entity Linking, salience analysis, relation extraction, and sentiment analysis on over one million news articles since 2018.

# Takeaway: We can handle labeling at specialized domain levels

With projects of this scale, complexity exists at the domain expertise and the annotation levels. Our work stands at the intersection of these two pieces. Through our continuing engagement with this customer, the financial services team has developed significant expertise in the field, and can even identify clues and subtle indicators buried in the textual material to enable pattern recognition and a deeper subject understanding.

At the same time, our teams have also tackled complex annotations at the dialog level for other customers. They are able to capture the relationships between the parts of an event, and handle difficult semantic correlations between discourse.

## Case study 2

Our customer develops speech-recognition APIs to power consumer requests. They required robust and accurate datasets to train their Natural Language Processing engine to extract insights from text data. In order to train the algorithms used for this service the company required our teams to annotate and identify entities, sentiments, intents, and relationships in vast amounts of unstructured textual data like emails, reviews, customer interactions, and social media posts.

The Solutions team at iMerit built an NER/NEL/sentiment tool which allows annotators to securely and rapidly annotate and link entities interactively. The user-friendly tool was customized for the requirements of the project. The tool presents a two-step workflow. The first step includes the uploading and tagging of the required entities.The second step covers quality assurance, where the submitted entries are checked by experts, and rectified if required. iMerit's internal service delivery platform iMPP, was used to serve several million sentences to an in-house team of over 80 contributors.

Confusing taxonomies were merged and rearranged, to ensure maximum clarity of classification. This highly curated pipeline ensured over 95% quality and a successful product launch by the customer in 2017.

## Takeaway: An expert-led training framework has a huge impact on quality.

Robust training is a critical success factor for labeling teams dealing with production-grade data annotation. Repeated exposure to the types of text and jargon they will encounter leads to superior outcomes. A strong training module at iMerit presented to its labeling and annotation team results in consistently high accuracy rates.

We have developed a training cycle that includes rigorous text analysis and comprehension modules. The team is presented with product descriptions, journalistic pieces, and an assortment of other written material. They are then assessed on their ability to break down the textual material at the entity, sentiment, and intent levels. Contributors are also equipped with search and social media scans to inform their annotation. An iterative process with feedback from our linguistic experts and the customers helps refine our labeling techniques, with a close eye on quality control. In the initial stages of a project, submissions are supervised by an expert, before heading into work on production-grade datasets. As teams gain more experience with a project and its use cases, random sampling percentages gradually reduce.

Over the course of a project an expert team specializing in linguistic data labeling is created and is defined by narrow scope, specific context, and deep subject understanding. High retention levels at iMerit ensure cumulative learnings and shared know-how of best practices.

## Case study 3

The customer provides a service that helps subscribers improve the quality of their writing by correcting mistakes and providing wording and phrase suggestions.
The company needed multiple wording suggestions to validate and refine a sentence correction algorithm. The training dataset required four correct versions for each

input sentence to check similarity and differences between human and machine-generated results. The challenge was in balancing the subjective and nuanced nature of copy-editing with the requirement for a rule-based style standardization.

Our teams were trained rigorously in the customer's style guide. Continuous assessment and peer learning, along with close collaboration with the customer in areas of ambiguity enables the team to deliver high-quality stylistic, structural and mechanical corrections on a variety of document sources.

## Takeaway: Linguistic and dialectic differences are tricky but create valuable data once cracked

While working with this customer, the text input spanned a wide range of registers to styles, from message boards written casually, and formal essays written by ESL learners. Maintaining the tonality of the different pieces while extracting the meaning and correcting errors in word choice and style and syntax and spelling was a crucial challenge that was overcome.

Verbal communication also presents a similar problem, as we experienced while tagging conversations that took place at fast food drive-through windows. The phrase "I'm all good", while spoken in the context of placing a food order means that the order is complete. But when this utterance is tagged at the word level, it sounds like positive feedback. In such cases, deciphering the intent of the sentence and understanding local turns of phrase enabled the team to correctly annotate the material. Creating a list of regional linguistic differences in the source material helped remove ambiguity and trained the experts to get into the head of the speaker.

# What's next in NLP?

The language and NLP space is expanding rapidly across a variety of use cases. Some remain in the R&D stage of testing, but many are well on their way to being moved into production.

# In conversation with Alexa

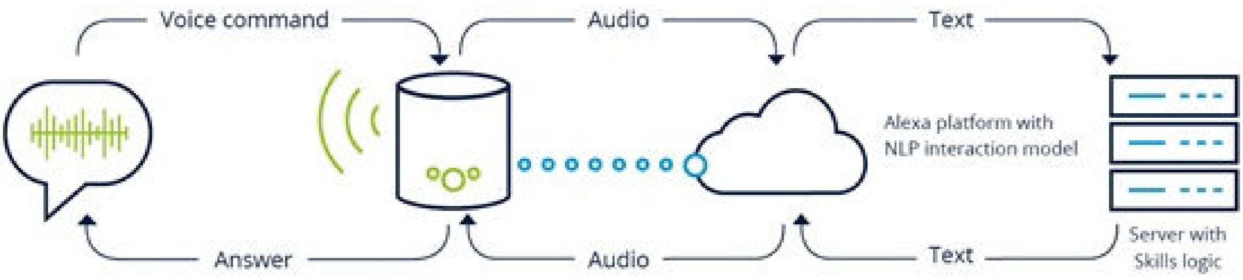
Multi-turn dialog, where you can have a longer conversation with a tool like Alexa, rather than breaking off after one exchange, is a cutting-edge area that is seeing a lot of interest and research.

Many dialog models have been trained with domain-specific datasets, where the interaction is geared to complete particular tasks. But more recently, open domain datasets with more variety and use cases are being used for conversational AI, and they need to be interpreted before being fed into the algorithm.

**User: “Alexa, where is the movie ‘Her’ playing near me?”**
**Alexa: “It’s playing at the AMC Century 50, at 6:30 PM and 10 PM.”**
**User: “Are people enjoying the movie?”**
**Alexa: “Yes, it has a review of 8.5 on IMDb”**

For this dialogue to be possible, Alexa’s algorithm has to keep track of multiple entities (the different movie, locations and reviews), intents (what the user wants to know), and actions (what the conversational agent needs to do).
In the data processing pipeline, this is a complex and multilayered operation. A major challenge here is a tool or platform to annotate data at that level, and when this can be successfully demonstrated, it opens up an array of possibilities.

## How Alexa Works



Source: ICMI, *First Things First - How Alexa Works*

# It’s all about you

Teams of developers globally are working on improving the levels of personalization that Artificial Intelligence can offer. A system where a chatbot or other AI models can learn and customize its interactions for each specific user is valuable in a number of business areas, particularly sales and marketing.

Any good sales or marketing person knows the value of tweaking their pitch for different audiences. Currently, browsing history-driven targeted ads are the most common form of online personalization, and a more advanced framework could help understand the best ways of communicating with a user, be it through variations in form, content streams, or tone. Users are more engaged when they are addressed in their own style of communication, and this is a powerful tool for businesses to leverage.

Detailed sentiment analysis of large amounts of online text can help train engines to recognize customer intents and desires to purchase, and any other feedback about products or services, and feed this directly into platforms in the marketing or sales funnel.

# Babel Fish comes to life

In the iconic book *The Hitchhiker’s Guide to the Galaxy*, the Babel fish is a small leech-like creature stuck in your ear, that can help you “instantly understand anything said to you in any form of language”. In the decades since the book was published, this has moved from the sci-fi realm to reality. Global innovators like Google and Baidu are working on perfecting headphones that offer translations in real-time. These, among several other linguistic assistance projects in the offing, require volumes of data in each and every language offered by the service.

English has been the most prevalent medium of research in NLP so far, but we will see a rising demand for more linguistic data as companies explore global markets for their products. To go back to our favorite example of Alexa, this service is currently offered in English, French, German, Italian, Japanese, Portuguese (Brazilian), Hindi, and Spanish. Within some of these languages like English and Spanish, some regional dialects are supported. Work is underway to expand upon this list.

For the Amazon voice assistant to be introduced in other languages, models which have been trained in English will need to be tweaked to work effectively and for this, training data will require robust cross-linguistic validation.

# What can iMerit do for your NLP project?

iMerit can handle your entire text-labeling pipeline. It can take unprocessed and unstructured textual data, and send it back in an annotated and meaningful form to power impactful algorithms.

Within iMerit's managed services model, the company has developed a dynamic team of linguistic labeling specialists who have been rigorously trained by subject matter experts and project managers. They work from secure facilities across Asia and the US, to provide an enterprise-grade data annotation service so that your algorithms can succeed.

iMerit's teams have gained expertise in a wide range of human communication types and can easily adapt to any datasets in the language labeling workflow. iMerit's work spans the entire range from sentiment analysis to NER, NEL and salience, in domains such as finance and e-commerce. With iterative quality control, accuracy and precision iMerit's quality is consistently high. Its feedback cycle and consultative relationship with customers brings the benefit of real human insight to millions of data points the world over.

Get in Touch

**iMerit.net** **info@imerit.net** **@iMeritDigital**