

Natural Language Processing & Speech Technology

At Appen, we provide high-quality annotated training data to power the world's most innovative machine learning and business solutions. We help build intelligent systems capable of understanding and extracting meaning from human text and speech for a diverse range of use cases, such as chatbots, voice assistants, search relevance and more.

Many of our annotation tools feature Smart Labeling capabilities, which leverage machine learning models to automate labeling and enable contributors to work quickly and more accurately.



Pre-Labeling

Receive an initial 'best guess' hypothesis using our library of models before the job runs.



Speed Labeling

Leverage machine learning models for quick and accurate annotation while the job runs.



Smart Validators

Use machine learning models to verify human judgments before contributors submit the job.



The Appen platform is super neat and easy to navigate compared to most of its competitors. (...) Support has been super helpful. I get responses usually within minutes, if not, the following day. GumGum is especially happy with the Japanese annotation quality and support, which Appen has improved tremendously over the past year. What's been super helpful is to tell my customer success manager what it is I want to achieve and look to Appen to help me with the job design, creation, and coding."

Erica Nishimura, Data Curator, GumGum



- Search Classes
- Full Name
- First Name
- Last Name
- Academic Institute
- State
- Organization
- Academic Field
- Business Title
- Year

This report is based on a ^{Year} 2018 webinar given by ^{First Name} Wendy ^{Last Name} Chapman , ^{Business Title} PhD , Chair ,
Department of ^{Academic Field} Biomedical Informatics ,
University of ^{State} Utah School of ^{Academic Field} Medicine , and
^{First Name} Mike ^{Last Name} Dow , Technical Director Health Catalyst ,
entitled , " Tapping Into the Potential of Natural
Language Processing in Healthcare . "

Text Annotation

Enhance your NLP model's understanding of nuanced human speech. Speed Labeling capabilities include built-in multi-language tokenizers to assist human annotation efforts.

Target entity extraction and span labeling with options to bring your model outputs to accelerate contributor annotations. Expand on your Natural Language Processing (NLP) labeling by connecting named entities or parts of speech within relationships.

Text Utterance Collection

Filter out unusable utterances automatically before they are collected saving you time and money and reducing error rates by up to 35% using our Smart Validators.

Gather large volumes of high-quality, customized text utterances for training chatbots and other conversational AI models.

Duplicate Detection 🗑️

Block submission of the same input for this question:

In this job, across all contributors

In this job, across all contributors, across a unique prompt value

Prompt Column

Select Column ▾

Coherence Detection 🗑️

Ensures coherent input . A higher threshold results in stricter model evaluation.

Coherence Threshold

0 0.60 1

Language Detection 🗑️

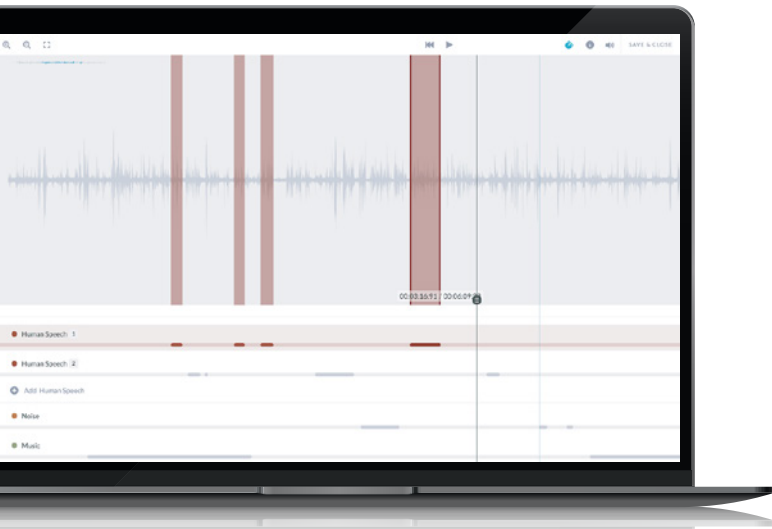
Ensures correct use of language. A higher threshold results in stricter model evaluation.

Language

English ▾

Leniency Threshold

0 0.60 1



Audio Annotation

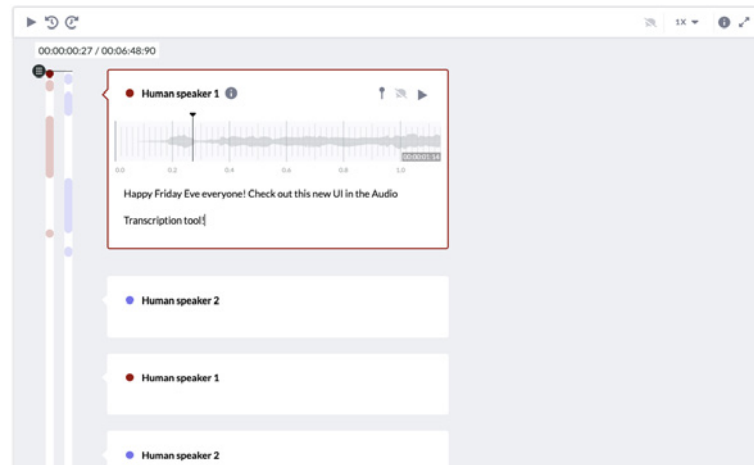
Acquire audio data ready for models or transcription with clever audio timestamping. 2x faster- improved Audio Annotation tool now twice as fast as traditional annotation tools.

Segment audio into layers, speakers and timestamps for your Audio Speech Recognition and other audio models.

Audio Transcription

Generate high-quality audio transcripts rapidly with acoustic tags in a variety of languages.

Transcribe spoken audio into text or validate machine-generated transcriptions. Leverage NLP to improve transcription quality and efficiency.





Text Evaluation and Post-Editing

Bring a level of nuance with a human touch and ensure machine generated text outputs meet your level of quality and coherence.

Evaluate the naturalness and relevance of the text generated by NLP models, such as machine translation models and other sequence models with the help of our multi-lingual specialists. We can also help post-edit machine generated text in order to make them more suitable for your use case.

Here are just a few ways our customers use our high-quality training data to solve real world business problems:



Microsoft partnered with Appen to improve the user experience and quality of BING's search results for the U.S. and international markets. They are now able to achieve high-quality search results that are accurate, timely, comprehensive, free of spam and relevant to the search query intent. **Microsoft** is able to process millions of pieces of search data every month in more than a dozen markets worldwide with ever-improving quality. To ensure efficient and consistent reporting in all markets, we also developed a proprietary data analysis and reporting tool.

With the help of our expert Linguistic team, and our recommendations for improving the evaluation process,

Microsoft is able to grow rapidly in new markets.



CallMiner, a pioneer of the artificial intelligence (AI)-powered speech analytics space partnered with Appen to train AI models to understand customer service conversations, including sentiment and emotions, and other relevant insights between organizations and their customers. The team needed a large dataset across an array of organizations to parse out the truly negative moments for Sentiment analysis.

CallMiner has been using our platform to annotate sentiment and emotion of call center data. Our platform lets them process more calls faster and with more accuracy, enabling them to expand their customer base and explore new types of conversation insights with the extra time saved.



Speech and Audio Collection

Improve machine learning at scale and fast-track your project with our data collection services and off-the-shelf data sets.

Gather large volumes of high-quality, customized speech and audio data for training voice-prompted virtual assistants, voice activated search functions, transcription services, voice-to-text capabilities and more.



London School of Economics and Political Science wanted to capture the content of the messages that political actors send to others and further, using those discoveries to calculate political party positions. They also wanted to identify indicators that would measure the sophistication, or readability, of political texts. They required a large and varied sample size of texts, and numerous human labelers to compare texts to one another, across several languages, so they chose us.

They were able to use our technology platform and our global Crowd to accomplish data labeling in a way that's fast, inexpensive, and scalable—without sacrificing data quality.



Dialpad leverages Appen for audio transcription and categorization to build their transcription models as well as verifying internal transcriptions and outputs of their models. They use our geolocation tools to make sure British contributors label idiomatic speech from the U.K.

Within a few weeks, **Dialpad** saw accuracy go up to 88% and it has stayed in the high 80s and 90s ever since, even across a large diversity of models.