



## Language Model Tool

---

### Overview

The Language Model Tool enables you to customize the vocabulary and corresponding pronunciations in the language model of the Media Mining Indexer for better speech recognition results. It helps you to fine tune the Media Mining Indexer for a more focused coverage of domain specific information.

This is a much needed tool in the context of the ever-changing nature of broadcast news content. So if you are simultaneously analyzing audio information from different domains, you can tune indexer instances for optimal indexing and analysis.

You also have the flexibility to switch between domain specific language models on the same channel depending upon the nature of content that is broadcast within particular time frames.

The toolkit can be configured to automatically update the indexing process with data from a specified source. This is also offered as an offline service by Sail Labs in case users want to avoid the process of local installation and maintenance.

---

### Features and Benefits

- Open Interface facilitates addition of data from multiple sources through an API
- Standard Interface is provided for text and html inputs
- Easy to use GUI makes updating the Indexer convenient for a non technical user
- Automatic generation of pronunciations for newly added vocabulary minimizes the need for human intervention
- Toolkit is provided with a base model trained on a corpus of several hundred million words, ensuring robust recognition even with minimal addition
- Users have the flexibility to set the importance of the new information as compared to the information provided in the base models for optimal recognition results

---

## Technical Specifications

---

### Prerequisites

The process of building the Language Models does not require the Sail Labs Indexer. In other words, it is not necessary to have the Sail Labs Indexer installed on the same machine.

---

### Hardware prerequisites

- Memory: 2GB recommended (1 GB minimum; depending on usage)
- CPU: Pentium 4 2.4 GHz or better recommended (Pentium-III/733 minimum)
- Disk space: 40GB on a fast disk recommended (minimum 7200 rpm; 10.000 rpm recommended for professional use)

Guidelines for disk space calculation:

- Base: 15MB for the *Language Model Tool* , 500MB for each language
- Run-time: 500MB for each Language Model built, 10 GB free disk space for temporary data

---

### Software prerequisites

Sun Java 1.4.0 Runtime

## How it works

The Language Model Tool generates a new version of a Language Model. The original Language Model is part of the Media Mining Indexer Language Feature. A Language Model contains:

- Spellings of words
- Pronunciations of words
  - Words can have multiple pronunciations: USA = "y-u-e-s-ay" or "y-u-e-s-o-v-ay"
  - Different words can have the same pronunciations: too, to, two
- Context of words
  - The probability of every word triple, double and single is calculated: *This stock market is strictly for professional traders with itchy fingers and confident five-minute options*
  - Context is used to resolve ambiguities: *Good morning.* - *Mourning and weeping.* *The right answer.* - *I often write letters to David Wright.*

To create a new Language Model lists of words and documents need to be prepared. The LM Tool generates pronunciations for the words. The Language Model is updated with the new words and sentences. An installation package is created.

---

## Language Options

Languages supported by the Language Model Tool are:

- Arabic
- French
- German
- Spanish
- English (US telephony)
- English (US/UK)