



Linguistic Tools



Diacritizer™

The AppTek Diacritizer™ (Vowelizer) is a software component of the Text-To-Speech (TTS) engine. It is the component that has the knowledge to add the appropriate diacritics (short vowels) to text. In particular, the Diacritizer offers the following services:

- It marks the sentences within running text.
- It analyzes a sentence and produces a syntax tree.
- It generates the diacritized counterpart for each word in the sentence.
- It keeps track of how its output maps onto the original input text on a word -by-word basis.

WordTag™

Word Tag™ identifies all words in a document automatically and tags them with set of linguistic information (annotations.)

During tokenization, the tokens are tested to determine whether they are potential words or merely data entities (numbers, alpha-numeric text segments, etc). Data entities are tagged as such no further linguistic identification is done on the entities according to their class. As for word tokens, they undergo the following steps:

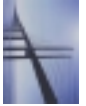
- Processing by the prefix/suffix identification and removal module.
- Consequent derivation of one or more prefix/suffix stems (PSStems).
- Generation of a part-of-speech for each stem by use of statistical information
- One or more PSStems are then passed through the linguistic tagger, which utilizes a large lexicon. The tagger will retrieve all the known linguistic information about the PSStem word: part-of-speech (Noun, Verb, Proper Noun, etc), morphology information (person, gender, etc), meanings (from multiple domains) and word use (General, Military, Computer, etc).
- All the linguistic data retrieved is then formatted in an XML document for later analysis by any additional tools. The words are tagged for POS, Meaning, Semantics, Derivations, Use, Domain, Name Entities, and Data Entities etc.

Language Recognizer

The language recognizer is a tool used to identify the language and code page of any electronic text. The system can recognize more than 54 different languages and code pages.

LexAPI™

AppTek has developed a special Applications Programming Interface for its morphology, lexicon and linguistic analyzer (LexAPI™) to facilitate the integration with third party applications. LexAPI™ adds powerful linguistic capabilities like morphology, query translation, thematic and domain search and word linguistic attributes to such applications.

















Technology to Bridge the Language Gap

Transliteration / Romanization Tool

Transliterate the phonetic representation of a given name, word or text from foreign languages into English and vice versa. The tool uses statistical information and linguistic algorithms to perform this task. Different permutations of spellings can be produced. Furthermore, different options for translation standards are available.

Running Text







AppTek has parallel corpora with millions of words of running text. This text is used for text language technology, statistical applications as well as speech research and development. They cover the following domains:

-  Aviation
-  Chemical
-  Civil Engineering
-  Construction
-  Environment
-  General
-  General/Grammar
-  Geography & Military
-  Legal
-  Medical
-  Military
-  Names
-  Physics
-  Technology
-  Telecommunication

Corpora Tagging

AppTek has worked on grammatical tagging of words and phrases of corpora, which has been diacritized. The study of the inherent grammatical features of the word combined with its behavior and function in the sentence as it relates to other words and phrases helped in tracing the changes that occur in transforming the sentence with all of its constituents into another language as target.

As a result, a correspondence was created between the tagging parameters and their source counterparts to provide the following:

-  Syntactic parsing and transfer information.
-  Grammatical tagging for multiple senses and homographs once looked up in the dictionary.
-  Source Grammatical permutations for grammatical categories when translated into the target.
-  Statistics on occurrences of grammatical phenomena in both language environments.
-  Contrastive functional structure correspondence between the grammar of words, phrases and whole sentences in both language environments.
-  Traceability of grammatical phenomena, categories and classes as they pertain to translated sentences.