

# SYSTRAN

## Linguistic Development Kit (LDK)

### Create the solutions your customers are dreaming of

Because Machine Translation (MT) development requires high-end technologies, SYSTRAN has been at the forefront of linguistics and natural-language processing for decades.

Our supported languages offer key linguistic functionalities, including:

- Document filtering
- Language identification
- Segmentation and tokenization
- Language normalization
- Document classification (domain detection)
- Named-entity recognition
- Dictionaries
- Morphological analysis
- Syntactic analysis
- Transliteration
- Word-sense disambiguation (WSD)



Taken separately, each module is an efficient tool for processing language, documents or names. Combine them and you benefit from powerful multilingual capabilities for data mining or semantic search solutions.

Whatever your end-customers need (e-discovery and digital forensics, OSINT or COMINT analysis, competitive and marketing intelligence, e-reputation monitoring, sentiment analysis for more customer insight), SYSTRAN's LDK enables you to utilize and analyze both structured and unstructured multilingual content, such as user-generated content, social media, Web content and more.

Now you're fully equipped to create powerful data-mining or intelligence solutions for your innovation-loving customers!

These linguistic features are distributed as a **Software Development Kit** with a low memory footprint, making it easy for external developers and tech partners to embed them in their own apps and services.

REQUIREMENTS	Low memory footprint native C++ libraries, with bindings for most common development languages (Java, .NET, Javascript, etc.)
PLATFORM SUPPORT	Windows, Linux, Mac OS, iOS, Android



### Document Filtering

Imports various document formats for processing by other modules, and rebuilds of the document in the original format with modified or annotated content and preserved layout



### Language Identification

Automatically identifies which language documents are written in, through specific word- or sentence-sample detection



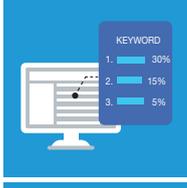
### Segmentation and Tokenization

Segments text into sentences and sentences into “tokens” (minimal processing units)



### Language Normalization

Normalizes text from blogs, emails, chat forums and user-generated content, while correcting common errors and language deviations



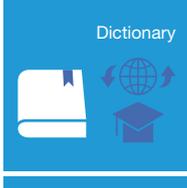
### Document Classification

Determines the document “domain” based on predefined models, and displays key “hot topic” words



### Named Entity Recognition

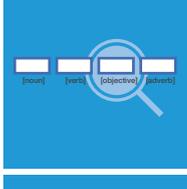
Based on the analysis of the document contents, automatically recognizes and displays person names, locations, numbers, dates and organization names



Dictionary

### Dictionary

Offers monolingual and bilingual dictionary lookups, with additional contextual information such as frequency of meanings, domains and contexts, expressions and examples



### Morphological Analysis

Provides morphological analysis for individual words, returning the list of possible lemmas and parts of speech for an inflected form



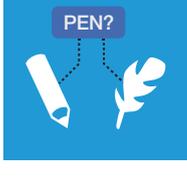
### Syntactic Analysis

Provides syntactic analysis at the sentence level, with the layers of linguistic analysis: word identification, part-of-speech tagging, and constituent and dependency analysis



### Transliteration

Makes transcription and transliteration of words or entities between languages with different scripts, and detects proper noun origins



### Word-Sense Disambiguation (WSD)

Selects best meaning of a word depending on the context