IQVIA

IQVIA NLP Multilingual Capabilities

IQVIA NLP provides varying levels of support for text mining non-English language documents. This document summarizes the most commonly available natural language processing (NLP) capabilities we have applied to non-English language documents. Our capability in this area is constantly evolving based on our customers' requirements; below is a summary of the most commonly supported capabilities:

- comprehensive, high quality NLP-based capabilities including quantitative expressions and linguistic units for English, Dutch, French, German, Italian, Mandarin Chinese, Portuguese and Spanish

- linguistic support for Nordic languages (Danish, Norwegian and Swedish).

If your preferred language is not yet included above, please speak to one of our consultants who will be able to provide further information about the current level of support available. A number of key capabilities enable our technology to analyze text in different languages; these are described below.

Pattern classes

Pattern classes enable the technology to match and normalize expressions such as dates, numbers, measurements and genetic mutations. Typically there is commonality across languages, though adaptation is required for items such as numbers, where words may be used instead of digits.

ANSI terminologies

The application of terminologies enables the technology to recognize concepts within the documents. Use of multilingual terminologies such as MedDRA (available in 13 languages) can improve recall while preserving precision by:

- searching for a concept and matching synonyms from multiple languages to improve recall

- matching English synonyms only in English text, German synonyms only in German text, etc., to improve precision.

Linguistics

A range of linguistic extraction techniques are used to transform unstructured text into structured, including:

- **tokenization**, used to break down text words

- **stemming** to identify the root of a word and recognize different variants with the same root

- **Part of Speech (PoS) tagging**, which labels words according to their grammatical role (noun, verb, proposition)

- **chunking**, which groups words together based on their PoS, for example grouping a verb and an adverb to form a verb group.

The table on the next page summarizes IQVIA NLP's multilingual support capability.

ANSI TERMINOLOGIES

	MedDRA	MeSH*	RELATIONS	SNOMED	CONTEXT
ENGLISH	●	●	●	●	●
DUTCH	●	●			●
FRENCH	●	●	●		●
GERMAN	●	●	●		●
ITALIAN	●	●			●
MANDARIN CHINESE	●				
PORTUGUESE	●	●			●
SPANISH	●	●	●	●	●

* MeSH coverage is more varied between languages than MedDRA.

	PATTERN CLASSES					LINGUISTICS
	NUMERICS	MEASURES	DATES*	AGE	UNITS	
ENGLISH	●	●	●	●	●	●
DUTCH	●	●	●	●	●	●
FRENCH	●	●	●	●	●	●
GERMAN	●	●	●	●	●	●
ITALIAN	●	●	●	●	●	●
MANDARIN CHINESE	●	●	○	●	●	●
PORTUGUESE	●	●	●	●	●	●
SPANISH	●	●	●	●	●	●

* All dates are using the English language dates pattern (default ISO patterns) with native vocabulary.

Key: ● = Available now   ○ = In development

CONTACT US

+44 (0)1223 651 910 (U.K.) | +1 617 674 3256 (U.S.)

nlp@iqvia.com

linguamatics.com

IQVIA