Case Study on Word Analyser (Morphological Analyser)

The digital content on the internet in regional languages are growing exponentially. These contents are in all possible genres and in different writing styles. These pose a challenge to the dictionary industries, where people search for the meanings and the equivalent words in other languages for the new words. For addressing this need, publishing houses started to develop growing dictionaries in various languages, where they gather the new words using various web-crawling techniques, identify the root words and include them in the mono-lingual and bi-lingual dictionaries. For automatic identification of root words in morphologically rich languages, morphological analysers are required. The challenge for the publishing houses was in handling the agglutinative characteristics of compound word formation and the morphological inflection in Dravidian languages such as Malayalam and Tamil.

Computational Linguistics Research Lab (CLRG) at AU-KBC Research Centre is focused on scientific study of languages from a computational perspective, has developed computational models of various linguistic phenomena and built practical NLP systems. CLRG with two decades of experience in developing NLP tools and applications for Dravidian languages has robust shallow processing modules namely Morphological analyser, Part-of-Speech tagger, Chunker, Clause Boundary Identifier and semantic processing modules namely Named Entity Recognizers (NER), Anaphora resolution and co-reference resolution engines.

The morphological analyser in the workbench is

- Developed using a hybrid architecture, using machine learning techniques and deterministic finite state automata
- Provides libraries for customization
- Easily customisable to any language by a language expert
- Ability to analyse the unseen words.

Oxford University Press (OUP), leading publishing house across the globe and prominent dictionary developer for the languages across the globe, has used our morphological analyser for growing dictionary development in Malayalam and Tamil. This tool has increased the efficiency and speed of their work. OUP used it for their startups in dictionary building as well.

After multiple rounds of rigorous testing and third party evaluation on linguistics and software engineering aspects, OUP was satisfied with the performance. They procured the Malayalam and Tamil morphological analyser.