Segmentation of Proper Names in Scientific Papers.

The aim of this semester project is to evaluate and improve Proper Name segmentation methods in texts. It will be done in collaboration with Frontiers, which will provide real-world data.

Frontiers is a leading open-access publisher of academic journals. They process large amounts of unstructured information (in textual form) to build knowledge databases. N*ormalization* is one critical aspect in this process, in particular for proper names. For example, to properly parse names like "*Alexandro Gomez Ricardo*", "*Yvan Samuel Jacquard*", "*J-P Richard*" or "*Charles-Phillipe Edouard de la Clairefontaine*", to distinguish them from other capitalized strings in the scientific papers and to properly segment them into first name and family name.

This implies to understand the linguistic rules or key features of proper name compositions. For example, to understand that Spanish names often contain two family names (in the above example: "*Gomez Ricardo*", whereas European names often contain a second surname (in the above example: "*Samuel*").

The goal of this semester project is to develop a complete solution to handle proper names and affiliations. The project would include:

1. the review & evaluation of current state-of-the-art solutions;
2. the development of an evaluation corpus that includes difficult names and
3. the application of the solutions found in 1) to the evaluation corpus from 2).