UMR 8094

Langues, Textes, Traitements Informatiques, Cognition

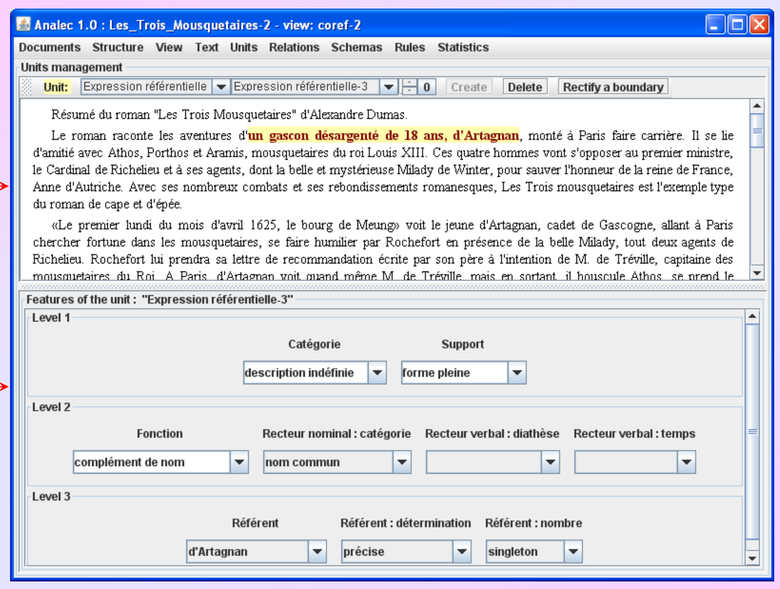
ANALEC: a New Tool for the Dynamic Annotation of Textual Data

Frédéric Landragin, Thierry Poibeau and Bernard Victorri

ANALEC is a new tool which aim is to bring together corpus annotation, visualization and query management. The main idea is to provide a unified and dynamic way of annotating textual data. ANALEC allows researchers to dynamically build their own annotation scheme and use the possibilities of scheme revision, data querying and graphical visualization during the annotation process. Each query result can be visualized using a graphical representation that puts forward a set of annotations that can be directly corrected or completed. Text annotation is then considered as a cyclic process. Statistics like frequencies and correlations make it possible to verify annotated data on the fly during the annotation.

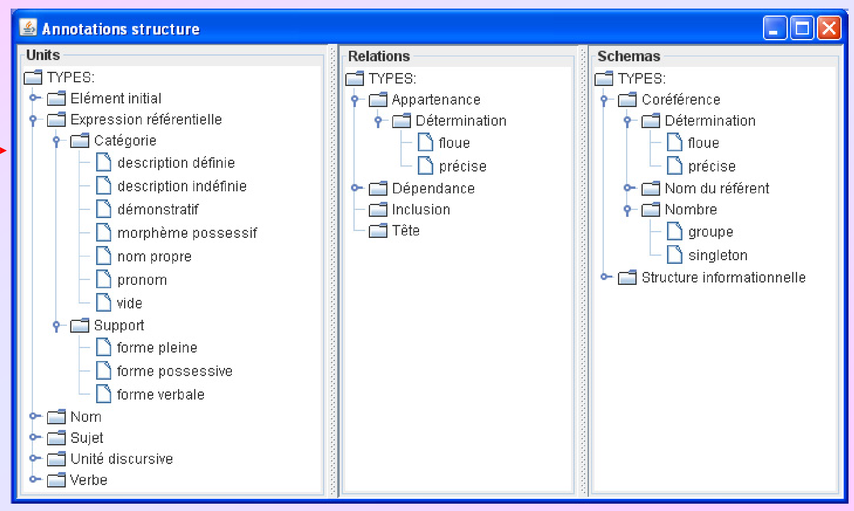
Corpus visualization zone →

Corpus annotation zone: →
In this example, the features are classified using three levels



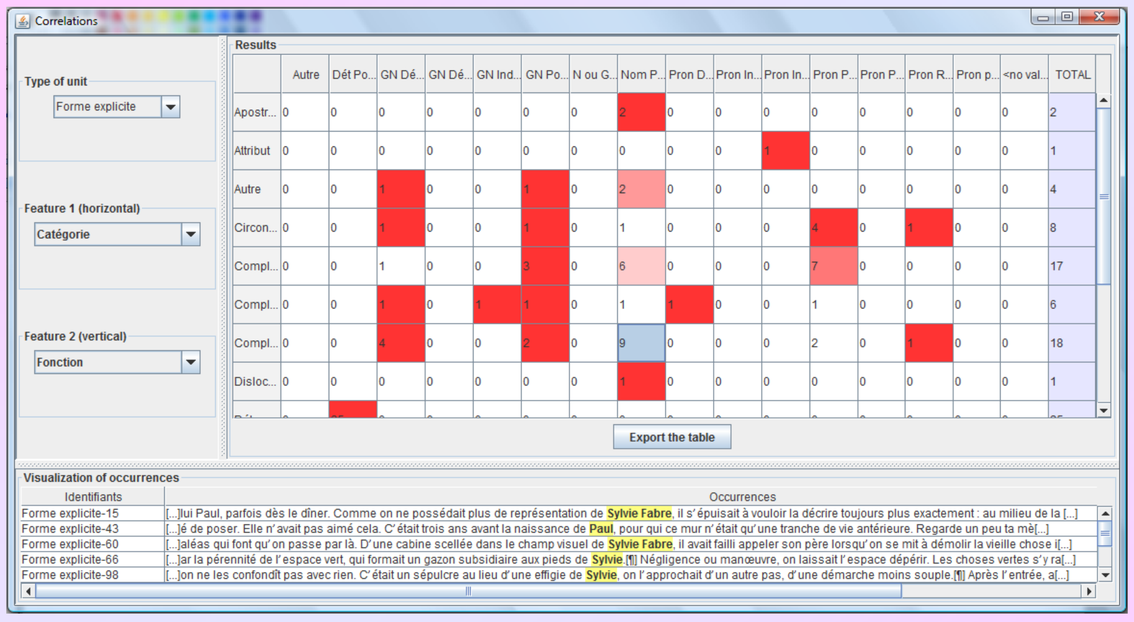
Some features can be edited, others not depending on the **'view'**

This window allows the user to navigate in **the annotation scheme** →



Each type, each element and each value of an element can be renamed and deleted

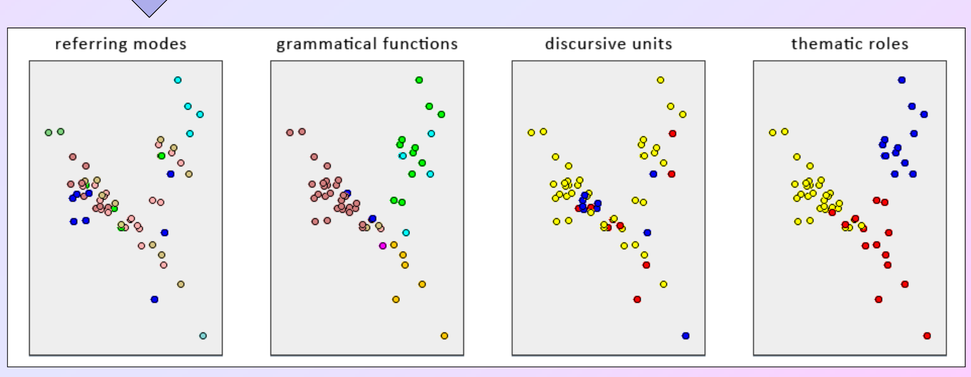
Correlation window



The user can choose a unit and two different features, and ANALEC automatically computes a table displaying their correlations. The cells in the table are coloured following the result of a classical chi-squared test

Geometrical representation window

Each point in the graph is linked to a linguistic unit, and interesting groupings may appear depending on the features taken into account



This module is especially useful to check that all the elements of a certain type are grouped together or not

Designing an annotation scheme is a major issue that requires a 'trial and error' strategy. Moreover, data annotation is not a static process but a dynamic one, which depends on corpus visualization and data queries. The goal of ANALEC is to allow researchers to dynamically build their own annotation scheme, using the possibilities of partial annotation, dynamic scheme revision, data querying and visualization during the annotation procedure itself. ANALEC includes three kinds of computation and visualization modules to analyse annotated data: frequencies, correlations, and geometrical representations. With these modules, it is possible to observe and prove linguistic hypotheses in vivo, directly from the observation of the data.

This research is partially sponsored by the contract PEPS MC4 (Modélisation Contrastive et Computationnelle des Chaînes de Coréférence) from CNRS INSHS and INS2I.