

---

## Preface

Research in Natural language Processing as well as practical task-oriented projects increasingly require sophisticated software architectures. Several factors contribute to that tendency: development of experimental studies on large multi-formated corpora with a need for short modeling – experimentation – evaluation cycles; combination in the same process of different models, resources and algorithms; reuse and capitalisation of these modules and resources; sophistication of the considered linguistic models... These various elements converge to promote the design and use of genuine *development environments* (or *platforms*) dedicated to Natural Language Processing. Using such environments yields important productivity gains in the design and implementation of operational systems as well as in the tuning of linguistic models themselves.

Many such platforms have been built and experimented in the last years, either by industrials, academics or inside collaborative projects. If some “majors” stand out in the landscape with respect to their audience and accumulated experience (GATE, UIMA, Open NLP, NooJ, Unitex...) one can also observe a remarkable spreading of these technologies, as echoed in the present volume. Such a “biodiversity” is neither accidental nor a product of the compelling tendency of computer scientists to always rebuild “new” systems from scratch. Instead, it results from a great diversity of constraints of both computational and linguistic nature, depending of the target tasks and use conditions, leading to complex and articulated choices:

- choice of a software architecture allowing to combine different processes: pipeline, agents, distributed Web services...
- representation of documents and annotations added by successive processes: robustness with respect to format variability, multilingualism or multimodality...
- available linguistic resources and modules (analysis models, grammars, lexicons...), facilities to integrate new or external ones;
- user interfaces proposed in order to constitute complex processing chains and visualise their effect on corpora;
- ...

These questions are discussed in the contributions gathered for the present volume. Ten different platforms are presented with the authors carefully describing their functionalities, objectives, and design principles, together with relevant applications or experiments.

A first group contains three articles each presenting a “*generalist*” platform, designed to bring together a great diversity of processes for *a priori* unspecified tasks. They are representative of what may be called the *mainstream* in the research area.

- J. Heinecke, G. Smits, C. Chardenon, E. Guimier De Neef, E. Maillebuau and M. Boualem present *TiLT : plateforme pour le Traitement Automatique des Langues Naturelles*, developed in the industrial context of Orange Labs. This platform incorporates a rich array of modules: lexical correctors, taggers, syntactic analysers (chunking and dependencies), semantic analysis (semantic nets) with special attention to multilingualism. The question of ambiguity is addressed and a multicriteria decision-aid system is described, together with some real life applications with strong constraints in terms of robustness and portability.
- In *ANTELOPE. Une plateforme industrielle de traitement linguistique*, F.-R. Chaumartin, observing that many resources and analysers are distributed inside the NLP community, presents an architecture directed towards the integration of external components inside an “homogeneous” platform, relying on general “good practice” Software Engineering principles. ANTELOPE is grounded on Melčuk’s Text-Meaning theory and provides access to different lexical semantic resources “around WordNet” in a unified way. ANTELOPE was also developed in an industrial context (the PROXEM company) and applications are evoked.
- With *Articulation des traitements en TAL. Principes méthodologiques et mise en œuvre dans la plateforme LinguaStream*, A. Widlöcher and F. Bilhaut carry out a methodological reflection on the diverse and heterogeneous treatments required by corpus-based experimental studies and the conditions for their seamless articulation inside one platform. A set of linguistic and data-processing principles are made explicit: open component-based architecture; heterogeneity of linguistic objects and structures, implying the availability of different analysis paradigms (grammars, transducers, lexicometry, lexical resources...); necessity of a unified annotation format; tools for observation of the analysed corpora... The paper shows how these principles are implemented in the LinguaStream platform, and three applications are discussed.

Three papers then concern *specific applications* or *specific stages* in a processing chain:

- H. Saggion, SUMA, *A Robust and Adaptable Summarisation Tool*, presents a generic system devoted to the development of automatic summarisation procedures within the paradigm of the GATE platform. On the one hand it relies on this platform for a number of “ordinary” tasks (documents storage, processes chaining, text annotation, visualisation...) and on the other hand it includes a set of adaptable specific components and evaluation tools, the “SUMA Toolkit”. The system also offers multilingual and multi-documents functionalities.

- The project presented by T. Hamon and A. Nazarenko in *Le développement d'une plate-forme pour l'annotation spécialisée de documents web : retour d'expérience*, concerns specifically the development of “semantic” search engines for specialised domains. The authors have designed the Ogmios platform, which incorporates classical information retrieval and extraction modules, from named-entity recognition to semantic tagging via anaphora resolution. A distributed architecture allows processing speed up. Several experiments using Ogmios are presented and evaluated, leading to methodological considerations regarding such topics as the problem of platform evaluation, the necessity of experimentation tools, the local-global mixing of treatments or the semantics of annotation.
- In *SxPipe2 : architecture pour le traitement pré-syntaxique de corpus écrits*, B. Sagot and P. Boullier raise the question of surface preprocessing steps that are needed to deal with raw corpora. They present a generic configurable platform, SxPipe2 which splits these tasks into several steps performed by specific exchangeable modules. A DAG representation allows to cope with ambiguities. A formalism of local grammars is available in order to perform pattern recognition, together with several “standard” components (spelling corrector, sentence segmenter, tokeniser, compound words and named entities recognizers...).

Two papers address the key issue of the *annotations* used to represent linguistic information extracted from texts, and more specifically the *fusion* of different sources of such annotations:

- In *CorpusReader : construction et interrogation de corpus multi-annotés*, S. Loiseau considers the combination of different analyses in a way that differs from mainstream platforms. Taking note of the wide-spreading and diversification of tools producing annotations at various levels (morphology, syntax, semantics...) he proposes to operate their *a posteriori* fusion rather than to produce them in one single environment. On the other hand he argues for the interest of such multi-annotated corpora in linguistics studies, inasmuch as they may reveal new relevant correlations involving different description levels. The CorpusReader platform allows to merge several annotations and to exploit the resulting documents by extracting sub-corpora, representations or quantifications.
- In *A flexible Framework for Integrating Annotations from Different Tools and Tagsets*, C. Chiarcos, S. Dipper, M. Götze, U. Leser, A. Lüdeling, J. Ritz, and M. Stede, also raise the question of merging annotations (with a particular emphasis on manually produced annotations and application to less-resourced languages). They focus on the question of interoperability of annotation formats. They introduce a pivot format, called PAULA, along with converters to and from this format, together with an ontology of linguistic annotations allowing to represent the correspondence between

alternative tagsets. The ANNIS platform performs these operations so that corpora can be visualised, queried, and evaluated across multiple layers.

The *evaluation* process can benefit from specific platforms:

- In *Sews : un Serveur d’Evaluation orienté Web pour la Syntaxe*, O. Hamon, P. Paroubek and D. Mostefa present a platform dedicated to automatic evaluation of syntactic parsers, developed and experimented in the context of the PASSAGE project. The authors describe its main functionalities and its software architecture. They discuss the problem of compatibility between different annotation models and present the retained EASY format. The recorded experiments show the considerable gain in efficiency, for participants as well for organisers, that such a platform provides.

Finally one paper reformulates in a somewhat radical way the question of implementation of NLP platforms.

- In *Cocytus : parallel NLP over disparate data*, N. Evans, M. Asahara et Y. Matsumoto observe that Unix-like operating systems already offer a wide range of tools for bringing together different processes and organise data flows; they argue that clever developments can provide efficient NLP environments as an alternative to “software” platforms. They propose the Cocytus system, based on the Inferno operating system in which they (notably) incorporated a representation of tree-like structures, adapted to flow processing, in order to encode linguistic data. Moreover, a speed-up of applications can be obtained by means of user-transparent parallelisation. A performance evaluation using the Penn Treebank has been performed and is presented.

As one can see, without being exhaustive, the whole set of these contributions bears witness of the wealth of a promising research area, source of significant productivity gains for the development of NLP applications.

Patrice ENJALBERT

Laboratoire GREYC (UMR 6072)

UFR Sciences - Université de Caen - Campus II

Bd. Maréchal Juin - B.P. 5186 14032 Caen Cedex

[Patrice.Enjalbert@info.unicaen.fr](mailto:Patrice.Enjalbert@info.unicaen.fr)

### **Coordinateurs**

Kalina BONTCHEVA, Université de Sheffield, Royaume Uni  
Patrice ENJALBERT, Université de Caen, France  
Benoît HABERT, ENS LSH et ICAR, France

### **Comité de lecture spécifique**

Jason BALDRIDGE, Université du Texas, Austin, USA  
Frédéric BILHAUT, Université de Caen, France  
Jean CARLETTA, Université d'Edimbourg, Royaume-Uni  
Farid CERBAH, Dassault Aviation, France  
Javier COUTO, INCO, Uruguay  
Robert DALE, Macquarie University, Australie  
François DAOUST, UQAM, Québec, Canada  
Thierry DECLERCK, DFKI, Allemagne  
Serge HEIDEN, ENS LSH et ICAR, France  
Nancy IDE, Vassar College, New-York, USA & LORIA/CNRS, France  
Michel JACOBSON, LACITO, France  
Diana MAYNARD, Université de Sheffield, Royaume-Uni  
Jean-Luc MINEL, MoDyCo, CNRS, France  
Sylvaine NUGIER, EDF, France  
Sébastien PAUMIER, Université de Marne-la-Vallée, France  
Étienne PETITJEAN, ATILF, France  
Thierry POIBEAU, LIPN-CNRS, France  
Laurent ROMARY, INRIA, France & MPG, Allemagne  
Vera Lucia STRUBE de LIMA, Université Catholique Pontificale du Rio Grande do Sul, Brésil  
Valentin TABLAN, Université de Sheffield, Royaume-Uni  
John TAIT, IRF, Autriche