Activity Report 2018

Team ALMAnaCH

Automatic Language Modelling and ANAlysis & Computational Humanities

Inria teams are typically groups of researchers working on the definition of a common project, and objectives, with the goal to arrive at the creation of a project-team. Such project-teams may include other partners (universities or research institutions).

RESEARCH CENTER
Paris

THEME
Language, Speech and Audio

Table of contents

Team ALMAnaCH

Creation of the Team: 2017 January 01

Keywords:

Computer Science and Digital Science:

A3.2.2. - Knowledge extraction, cleaning
A3.3.2. - Data mining
A3.3.3. - Big data analysis
A3.4.1. - Supervised learning
A3.4.2. - Unsupervised learning
A3.4.6. - Neural networks
A3.4.8. - Deep learning
A9.1. - Knowledge
A9.2. - Machine learning
A9.4. - Natural language processing
A9.7. - AI algorithmics

Other Research Topics and Application Domains:

B1.2.2. - Cognitive science
B1.2.3. - Computational neurosciences
B9.1.1. - E-learning, MOOC
B9.5.6. - Data science
B9.6.5. - Sociology
B9.6.6. - Archeology, History
B9.6.8. - Linguistics
B9.6.10. - Digital humanities
B9.7. - Knowledge dissemination
B9.7.1. - Open access
B9.7.2. - Open data
B9.8. - Reproducibility

1. Team, Visitors, External Collaborators

Research Scientists

Pierre Boullier [Inria, Emeritus]
Laurent Romary [Inria, Senior Researcher, HDR]
Benoît Sagot [Team leader, Inria, Researcher, HDR]
Djamé Seddah [Inria (détachement), Researcher, from Feb 2018]
Tommaso Venturini [Inria, Advanced Research Position, from Sep 2018 until Oct 2018, HDR]
Éric Villemonte de La Clergerie [Inria, Researcher]

Faculty Member

Djamé Seddah [Univ Paris-Sorbonne, Associate Professor, until Jan 2018]

Post-Doctoral Fellows

Yoann Dupont [Univ d'Orléans, from Jun 2018]

Murielle Fabre [Inria, from Nov 2018]
Ilia Markov [Inria, from Jun 2018]

PhD Students

Jack Bowers [Vienna Academy of Sciences]
Loïc Grobol [Ecole Normale Supérieure Paris]
Mohamed Khemakhem [Inria]
Louis Martin [Facebook, from Jun 2018]
Benjamin Muller [Inria, from Oct 2018]
Pedro Ortiz Suárez [Inria, from Oct 2018]
Mathilde Regnault [Ecole Normale Supérieure Paris]
Jose Rosales Nuñez [CNRS (LIMSI), from Jun 2018]

Technical staff

Wigdan Abbas Mekki Medeni [Inria, until Apr 2018]
Achraf Azhar [Inria]
Alix Chagué [CNRS (LARHRA), from Oct 2018]
Elias Benaïssa [Inria, until Apr 2018]
Farah Essaidi [Inria, Nov 2018]
Luca Foppiano [Inria]
Ganesh Jawahar [Inria, from Mar 2018]
Tanti Kristanti [Inria]
Alba Marina Malaga Sabogal [Inria]
Benjamin Muller [Inria, from Apr 2018 until Sep 2018]
Marie Puren [Inria, until Aug 2018]
Charles Riondet [Inria]
Dorian Seillier [Inria]
Lionel Tadonfouet [Inria]
Emilia Verzeni [Inria, until Apr 2018]

Interns

Rebecca Blevins [Inria, from Jun 2018 until Jul 2018]
Marie-Laurence Bonhomme [Inria, from Mar 2018 until Jul 2018]
Damien Biabiany [Inria, apprentice (“apprenti”), from Dec 2018]
Pauline Brunet [Inria, from Mar 2018 until Aug 2018]
Alix Chague [Univ Denis Diderot, from Apr 2018 until Jul 2018]
Farah Essaidi [Inria, from May 2018 until Sep 2018]
Amal Fethi [Ecole Normale Supérieure Cachan, from Apr 2018 until Aug 2018]
Pedro Ortiz Suárez [Inria, from Apr 2018 until Sep 2018]

Administrative Assistants

Christelle Guiziou [Inria, until Nov 2018]
Meriem Henni [Inria, from Nov 2018]

2. Overall Objectives

2.1. Overall Objectives

The ALMAnaCH team [1] brings together specialists of a pluri-disciplinary research domain at the interface between computer science, linguistics, philology, and statistics, namely that of **natural language processing**, **computational linguistics** and **digital and computational humanities**.

[1] ALMAnaCH was created as an Inria team (“équipe”) on 1st January, 2017.

Computational linguistics is an interdisciplinary field dealing with the computational modelling of natural language. Research in this field is driven both by the theoretical goal of understanding human language and by practical applications in **Natural Language Processing** (hereafter NLP) such as linguistic analysis (syntactic and semantic parsing, for instance), machine translation, information extraction and retrieval, human-computer dialogue. Computational linguistics and NLP, which date back at least to the early 1950s, are among the key sub-fields of **Artificial Intelligence**.

Digital Humanities (hereafter DH) is an interdisciplinary field that uses computer science as a source of techniques and technologies, in particular NLP, for exploring research questions in social sciences and humanities. **Computational Humanities** aims at improving the state of the art in both computer sciences (e.g. NLP) and social sciences and humanities, by involving computer science as a research field.

ALMAnaCH is a follow-up to the ALPAGE project-team, which came to an end in December 2016. ALPAGE was created in 2007 in collaboration with Paris-Diderot University and had the status of an UMR-I since 2009. This joint team involving computational linguists from Inria as well as computational linguists from Paris-Diderot University with a strong background in linguistics proved successful. However, the context is changing, with the recent emergence of digital humanities and, more importantly, of computational humanities. This presents both an opportunity and a challenge for Inria computational linguists. It provides them with new types of data on which their tools, resources and algorithms can be used and lead to new results in human sciences. Computational humanities also provide computational linguists with new and challenging research problems, which, if solved, provide new ways of addressing research questions in the humanities.

The scientific positioning of ALMAnaCH therefore extends that of ALPAGE. We remain committed to developing state-of-the-art NLP software and resources that can be used by academics and in the industry, including recent approaches based on deep learning. At the same time we continue our work on language modelling in order to provide a better understanding of languages, an objective that is reinforced and addressed in the broader context of computational humanities, with an emphasis on language evolution and, as a result, on ancient languages.

This new scientific orientation has motivated the creation of a new project-team at the crossroads between different scientific networks, and in particular:

- The École Pratique des Hautes Études, with which collaboration has already started on a number of topics related to Digital and Computational Humanities; [2]
- The Berlin Brandenburg Academy of Sciences in Berlin which hosts the national lexicographic project in Germany, funded by the German Ministry of Education and Research (BMBF)
- CNRS's Institut des Sciences de la Communication (Institute for Communication Sciences), on topics pertaining to Digital Social Sciences; [3]
- If confirmed, the PRAIRIE Institute (PaRis Artificial Intelligence Research Institute), whose goal

[2] When the ALMAnaCH team was created in January 2017, two EPHE permanent members were involved: Marc Bui, Directeur d'Études Cumulant, a specialist of computational humanities and of the computational modelling of the concept of proximity, and Daniel Stökl Ben Ezra, Directeur d'Études, a specialist of digital and computational humanities, Hebrew and Aramaic language, literature, palaeography and epigraphy. Since then, discussions and joint research endeavours have been initiated, showing the great potential of such a collaboration. Joint project proposals were submitted, one of which successfully, and we plan to work on future proposals in coming months and years. Yet after extensive discussions within all members involved in the team as well as with Éric Fleury, the head of Inria Paris, and François Jouen, Dean of the Natural Sciences department at EPHE, we came together to the conclusion that ALMAnaCH was not the optimal level for setting up a large-scale collaborative environment between both institutions, as the potential for collaboration between Inria Paris and EPHE goes well beyond NLP and text-based digital humanities. Discussions on a future Framework Agreement between EPHE and Inria Paris have started, in which ALMAnaCH will play a key role. In this context, several EPHE non-permanent members are still hosted at Inria Paris, within ALMAnaCH offices, in order to ease joint collaborations.

[3] ALMAnaCH hosted Tommaso Venturini, then on a fixed-term Senior Researcher Position, in September and October 2018, in the context of his involvment in one of ALMAnaCH's projects, the SoSweet project on Twitter-based sociolinguistics. He was granted a permanent position as CNRS Chargé de Recherches at the Institut des Sciences de la Communication starting in November 2018, and we intend to further collaborate in the future.

will be to act as a catalyst for research in Artificial Intelligence and for exchanges and between academia, industry and higher education in this domain, in which NLP plays a key role.

3. Research Program

3.1. Overview and research strands

One of the main challenges in computational linguistics is **to model and to cope with language variation**. Language varies with respect to domain and genre (news wires, scientific literature, poetry, oral transcripts...), sociolinguistic factors (age, background, education; variation attested for instance on social media), geographical factors (dialects) and other dimensions (disabilities, for instance). But language also constantly evolves over all possible time scales. [4] Addressing this variability is still an open issue for NLP. Commonly used approaches, which often rely on supervised and semi-supervised machine learning methods, require huge amounts of annotated data. They are still struggling with the high level of variability found for instance in **user-generated content** or in **non-contemporary texts**.

ALMAnaCH tackles the challenge of language variation in two complementary directions, to which we position a specific activity related to language resources:

3.1.1. Research strand 1

We focus on linguistic representations that are less affected by language variation. It obviously requires us to **stay at a state-of-the-art level in key NLP tasks** such as part-of-speech tagging and (syntactic) parsing, which are core expertise domains of ALMAnaCH members. It also requires improving the **generation of semantic representations (semantic parsing)**. This also involves the **integration of both linguistic and non-linguistic contextual information** to improve automatic linguistic analysis. This is an emerging and promising line of research in NLP. We have to identify, model and take advantage of each available type of contextual information. Addressing these issues enables us to develop new lines of research related to conversational content. Applications include chatbot-based systems and improved information and knowledge extraction algorithms. We especially focus on challenging such specific data sets as domain-specific texts or historical documents, in the larger context of the development of digital humanities.

3.1.2. Research strand 2

Language variation must be better understood and modelled in all its possible realisations. In this regard, we put a strong emphasis on **three types** of language variation and their mutual interaction: **sociolinguistic variation** in synchrony (including non-canonical spelling and syntax in user-generated content), **complexity-based variation** in relation with language-related disabilities, and **diachronic variation** (computational exploration of language change and language history, with a focus ranging from Old to all forms of Modern French, as well as Indo-European languages in general). In addition, the noise introduced processes such as Optical Character Recognition (OCR) and Handwritten Text Recognition (HTR) systems, especially in the context of historical documents, bears similarities with that brought by non-canonical input in user-generated content. This noise constitutes a more transverse kind of variation stemming from the way language is graphically encoded, which we call **language-encoding variation**. [5]

[4] We do not view multilinguality as a case of language variation. Yet multilinguality, a consequence of language diversity, obviously underlies many aspect of ALMAnaCH's research activities.

[5] Other types of language variation could become research topics for ALMAnaCH in the future. This could include dialectal variation (e.g. work on Arabic) as well as the study and exploitation of paraphrases in a broader context than the above-mentioned complexity-based variation.

3.1.3. Research strand 3

Language resource development is not only a technical challenge and a necessary preliminary step to create evaluation data sets for NLP systems as well as training and for machine learning models. It is also a research field in itself, which concerns, among other challenges, (i) the development of semi-automatic and automatic algorithms to speed up the work (e.g. automatic extraction of lexical information, low-resource learning for developing pre-annotation algorithms, transfer methods to leverage tools and/or resources existing for other languages, etc.) and (ii) the development of formal models to represent linguistic information is the best possible way, thus requiring expertise at least both in NLP and in typological and formal linguistics. Language resource development involves the creation of **raw corpora from original sources** as well as the (manual, semi-automatic or automatic) development of **lexical resources** and **annotated corpora**. Such endeavours are domains of expertise of the ALMAnaCH team. This research strand 3 benefits to the whole team and beyond, and both feeds and benefits from the work of the other research strands.

3.2. Automatic Context-augmented Linguistic Analysis

This first research strand is centred around NLP technologies and some of their applications in Artificial Intelligence (AI). Core NLP tasks such as part-of-speech tagging, syntactic and semantic parsing is improved by integrating new approaches, such as (deep) neural networks, whenever relevant, while preserving and taking advantage of our expertise on symbolic and statistical system: hybridisation not only couples symbolic and statistical approaches, but neural approaches as well. AI applications are twofold, notwithstanding the impact of language variation (see the next strand): (i) information and knowledge extraction, whatever the type of input text (from financial documents to ancient, historical texts and from Twitter data to Wikipedia) and (ii) chatbots and natural language generation. In many cases, our work on these AI applications is carried out in collaboration with industrial partners (for which cf. Section 7.1). The specificities and issues caused by language variation (a text in Old French, a contemporary financial document and tweets with a non-canonical spelling cannot be processed in the same way) are addressed in the next research strand.

3.2.1. Context-augmented processing of natural language at all levels: morphology, syntax, semantics

Our expertise in NLP is the outcome of more than 10 years in developing new models of analysis and accurate techniques for the full processing of any kind of language input since the early days of the Atoll project-team and the rise of linguistically informed data-driven models as put forward within the Alpage project-team.

Traditionally, a full natural language process (NLP) chain is organised as a pipeline where each stage of analysis represents a traditional linguistic field (in a *structuralism* view) from morphological analysis to purely semantic representations. The problem is that this architecture is vulnerable to error propagation and very domain sensitive: each of these stage must be compatible at the lexical and structure levels they provide. We arguably built the best performing NLP chain for French [63], [97] and one of the best for robust multilingual parsing as shown by our results in various shared tasks over the years [93], [90], [96], [21]. So we pursue our efforts on each of our components we developed: tokenisers (e.g. SxPipe), part-of-speech taggers (e.g. MElt), constituency parsers and dependency parsers (e.g. FRMG, DyALog-SR) as well as our recent neural semantic graph parsers [90].

In particular, we continue to explore the hybridisation of symbolic and statistical approaches, and extend it to neural approaches, as initiated in the context of our participation to the CoNLL 2017 multilingual parsing shared task [6] and to Extrinsic Parsing Evaluation Shared Task [7].

Fundamentally, we want to build tools that are less sensitive to variation, more easily configurable, and self-adapting. Our short-term goals is to explore techniques such as multi-task learning (cf. already [95]) to propose a joint model of tokenisation, normalisation, morphological analysis and syntactic analysis. We also explore adversarial learning, considering the drastic variation we face in parsing user-generated content and processing historical texts, both seen as noisy input that needs to be handled at training and decoding time.

[6] We ranked 3 for UPOS tagging and 6 for dependency parsing out of 33 participants.
[7] Semantic graph parsing, evaluated on biomedical data, speech and opinion. We ranked 1 in a joint effort with the Stanford NLP team

While those points are fundamental, therefore necessary, if we want to build the next generation of NLP tools, we need to *push the envelop* even further by tackling the biggest current challenge in NLP: handling the context within which a speech act is taking place.

There is indeed a strong tendency in NLP to assume that each sentence is independent from its siblings sentences as well as its context of enunciation, with the obvious objective to simplify models and reduce the complexity of predictions. While this practice is already questionable when processing full-length edited documents, it becomes clearly problematic when dealing with short sentences that are noisy, full of ellipses and external references, as commonly found in User-Generated Content (UGC).

A more expressive and context-aware structural representation of a linguistic production is required to accurately model UGC. Let us consider for instance the case for Syntax-based Machine Translation of social media content, as is carried out by the ALMAnaCH-led ANR project Parsiti (PI: DS). A Facebook post may be part of a discussion thread, which may include links to external content. Such information is required for a complete representation of the post's context, and in turn its accurate machine translation. Even for the presumably simpler task of POS tagging of dialogue sequences, the addition of context-based features (namely information about the speaker and dialogue moves) was beneficial [72]. In the case of UGC, working across sentence boundaries was explored for instance, with limited success, by [62] for document-wise parsing and by [82] for POS tagging.

Taking the context into account requires new inference methods able to share information between sentences as well as new learning methods capable of finding out which information is to be made available, and where. Integrating contextual information at all steps of an NLP pipeline is among the main research questions addressed in this research strand. In the short term, we focus on morphological and syntactic disambiguation within close-world scenarios, as found in video games and domain-specific UGC. In the long term, we investigate the integration of linguistically motivated semantic information into joint learning models.

From a more general perspective, contexts may take many forms and require imagination to discern them, get useful data sets, and find ways to exploit them. A context may be a question associated with an answer, a rating associated with a comment (as provided by many web services), a thread of discussions (e-mails, social media, digital assistants, chatbots—on which see below–), but also meta data about some situation (such as discussions between gamers in relation with the state of the game) or multiple points of views (pictures and captions, movies and subtitles). Even if the relationship between a language production and its context is imprecise and indirect, it is still a valuable source of information, notwithstanding the need for less supervised machine learning techniques (cf. the use of LSTM neural networks by Google to automatically suggest replies to emails).

3.2.2. Information and knowledge extraction

The use of local contexts as discussed above is a new and promising approach. However, a more traditional notion of global context or world knowledge remains an open question and still raises difficult issues. Indeed, many aspects of language such as ambiguities and ellipsis can only be handled using world knowledge. Linked Open Data (LODs) such as DBpedia, WordNet, BabelNet, or Framebase provide such knowledge and we plan to exploit them.

However, each specialised domain (economy, law, medicine...) exhibits its own set of concepts with associated terms. This is also true of communities (e.g. on social media), and it is even possible to find communities discussing the same topics (e.g. immigration) with very distinct vocabularies. Global LODs weakly related to language may be too general and not sufficient for a specific language variant. Following and extending previous work in ALPAGE, we put an emphasis on information acquisition from corpora, including error mining techniques in parsed corpora (to detect specific usages of a word that are missing in existing resources), terminology extraction, and word clustering.

Word clustering is of specific importance. It relies on the distributional hypothesis initially formulated by Harris, which states that words occurring in similar contexts tend to be semantically close. The latest developments of these ideas (with word2vec or GloVe) have led to the embedding of words (through vectors) in low-dimensional semantic spaces. In particular, words that are typical of several communities (see above)

can be embedded in a same semantic space in order to establish mappings between them. It is also possible in such spaces to study static configurations and vector shifts with respect to variables such as time, using topological theories (such as pretopology), for instance to explore shifts in meaning over time (cf. the ANR project Profiterole concerning ancient French texts) or between communities (cf. the ANR project SoSweet). It is also worth mentioning on-going work (in computational semantics) whose goal is to combine word embeddings to embed expressions, sentences, paragraphs or even documents into semantic spaces, e.g. to explore the similarity of documents at various time periods.

Besides general knowledge about a domain, it is important to detect and keep trace of more specific pieces of information when processing a document and maintaining a context, especially about (recurring) Named Entities (persons, organisations, locations...) —something that is the focus of future work in collaboration with Patrice Lopez on named entity detection in scientific texts. Through the co-supervision of a PhD funded by the LabEx EFL (see below), we are also involved in pronominal coreference resolution (finding the referent of pronouns). Finally, we plan to continue working on deeper syntactic representations (as initiated with the Deep Sequoia Treebank), thus paving the way towards deeper semantic representations. Such information is instrumental when looking for more precise and complete information about who does what, to whom, when and where in a document. These lines of research are motivated by the need to extract useful contextual information, but it is also worth noting their strong potential in industrial applications.

3.2.3. Chatbots and text generation

Chabots have existed for years (Eliza, Loebner prize). However, they are now becoming the focus of many concrete industrial developments, with the emergence of operational conversational agents and digital assistants (such as Siri). The current approaches mostly rely on the design of scenarios associated with very partial analysis of the requests to fill expected slots and to generate canned answers.

The next generations of such systems will rely on a deeper understanding of the requests, being able to adapt to the specificities of the users, and providing less formatted answers. We believe that chatbots are an interesting and challenging playground to deploy our expertise on knowledge acquisition (to identify concepts and formulations), information extraction based on deeper syntactic representations, context-sensitive analysis (using the thread of exchanges and profile information but also external data sources), and robustness (depending on the possible users' styles).

However, this domain of application also requires working on text generation, starting with simple canned answers and progressively moving to more sophisticated and diverse ones. This work is directly related to another line of research regarding computer-aided text simplification, for which see section 3.3.4.

3.3. Computational Modelling of Linguistic Variation

NLP and DH tools and resources are very often developed for contemporary, edited, non-specialised texts, often based on journalistic corpora. However, such corpora are not representative of the variety of existing textual data. As a result, the performance of most NLP systems decreases, sometimes dramatically, when faced with non-contemporary, non-edited or specialised texts. Despite the existence of domain-adaptation techniques and of robust tools, for instance for social media text processing, dealing with linguistic variation is still a crucial challenge for NLP and DH.

Linguistic variation is not a monolithic phenomenon. Firstly, it can result from different types of processes, such as variation over time (diachronic variation) and variation correlated with sociological variables (sociolinguistic variation, especially on social networks). Secondly, it can affect all components of language, from spelling (languages without a normative spelling, spelling errors of all kinds and origins) to morphology/syntax (especially in diachrony, in texts from specialised domains, in social media texts) and semantics/pragmatics (again in diachrony, for instance). Finally, it can constitute a property of the data to be analysed or a feature of the data to be generated (for instance when trying to simplify texts for increasing their accessibility for disabled and/or non-native readers).

Nevertheless, despite this variability in variation, the underlying mechanisms are partly comparable. This motivates our general vision that many generic techniques could be developed and adapted to handle different types of variation. In this regard, three aspects must be kept in mind: spelling variation (human errors, OCR/HTR errors, lack of spelling conventions for some languages...), lack or scarcity of parallel data aligning "variation-affected" texts and their "standard/edited" counterpart, and the sequential nature of the problem at hand. We will therefore explore, for instance, how unsupervised or weakly-supervised techniques could be developed and feed dedicated sequence-to-sequence models. Such architectures could help develop "normalisation" tools adapted, for example, to social media texts, texts written in ancient/dialectal varieties of well-resourced languages (e.g. Old French texts), and OCR/HTR system outputs.

Nevertheless, the different types of language variation will require specific models, resources and tools. All these directions of research constitute the core of our second research strand described in this section.

3.3.1. Theoretical and empirical synchronic linguistics

Permanent members involved: all

We aim to explore computational models to deal with language variation. It is important to get more insights about language in general and about the way humans apprehend it. We will do so in at least two directions, associating computational linguistics with formal and descriptive linguistics on the one hand (especially at the morphological level) and with cognitive linguistics on the other hand (especially at the syntactic level).

Recent advances in morphology rely on quantitative and computational approaches and, sometimes, on collaboration with descriptive linguists—see for instance the special issue of the *Morphology* journal on "computational methods for descriptive and theoretical morphology", edited and introduced by [60]. In this regard, ALMAnaCH members have taken part in the design of quantitative approaches to defining and measuring morphological complexity and to assess the internal structure of morphological systems (inflection classes, predictability of inflected forms...). Such studies provide valuable insights on these prominent questions in theoretical morphology. They also improve the linguistic relevance and the development speed of NLP-oriented lexicons, as also demonstrated by ALMAnaCH members. We shall therefore pursue these investigations, and orientate them towards their use in diachronic models (see section 3.3.3).

Regarding cognitive linguistics, we have the perfect opportunity with the starting ANR-NSF project "Neuro-Computational Models of Natural Language" (NCM-NL) to go in this direction, by examining potential correlations between medical imagery applied on patients listening to a reading of "Le Petit Prince" and computation models applied on the novel. A secondary prospective benefit from the project will be information about processing evolution (by the patients) along the novel, possibly due to the use of contextual information by humans.

3.3.2. Sociolinguistic variation

Because language is central in our social interactions, it is legitimate to ask how the rise of digital content and its tight integration in our daily life has become a factor acting on language. This is even more actual as the recent rise of novel digital services opens new areas of expression, which support new linguistic behaviours. In particular, social media such as Twitter provide channels of communication through which speakers/writers use their language in ways that differ from standard written and oral forms. The result is the emergence of new language varieties.

A very similar situation exists with regard to historical texts, especially documentary texts or graffiti but even literary texts, that do not follow standardised orthography, morphology or syntax.

However, NLP tools are designed for standard forms of language and exhibit a drastic loss of accuracy when applied to social media varieties or non-standardised historical sources. To define appropriate tools, descriptions of these varieties are needed. However, to validate such descriptions, tools are also needed. We address this chicken-and-egg problem in an interdisciplinary fashion, by working both on linguistic descriptions and on the development of NLP tools. Recently, socio-demographic variables have been shown to bear a strong impact on NLP processing tools (see for instance [68] and references therein). This is why, in a first step, jointly with researchers involved in the ANR project SoSweet (ENS Lyon and Inria project-team

Dante), we will study how these variables can be factored out by our models and, in a second step, how they can be accurately predicted from sources lacking these kinds of featured descriptions.

3.3.3. Diachronic variation

Language change is a type of variation pertaining to the diachronic axis. Yet any language change, whatever its nature (phonetic, syntactic...), results from a particular case of synchronic variation (competing phonetic realisations, competing syntactic constructions...). The articulation of diachronic and synchronic variation is influenced to a large extent by both language-internal factors (i.e. generalisation of context-specific facts) and/or external factors (determined by social class, register, domain, and other types of variation).

Very few computational models of language change have been developed. Simple deterministic finite-state-based phonetic evolution models have been used in different contexts. The PIElexicon project [78] uses such models to automatically generate forms attested in (classical) Indo-European languages but is based on an idiosyncratic and unacceptable reconstruction of the Proto-Indo-European language. Probabilistic finite-state models have also been used for automatic cognate detection and proto-form reconstruction, for example by [61] and [69]. Such models rely on a good understanding of the phonetic evolution of the languages at hand.

In ALMAnaCH, our goal is to work on modelling phonetic, morphological and lexical diachronic evolution, with an emphasis on computational etymological research and on the computational modelling of the evolution of morphological systems (morphological grammar and morphological lexicon). These efforts will be in direct interaction with sub-strand 3b (development of lexical resources). We want to go beyond the above-mentioned purely phonetic models of language and lexicon evolution, as they fail to take into account a number of crucial dimensions, among which: (1) spelling, spelling variation and the relationship between spelling and phonetics; (2) synchronic variation (geographical, genre-related, etc.); (3) morphology, especially through intra-paradigmatic and inter-paradigmatic analogical levelling phenomena, (4) lexical creation, including via affixal derivation, back-formation processes and borrowings.

We apply our models to two main tasks. The first task, as developed for example in the context of the ANR project Profiterole, consists in predicting non-attested or non-documented words at a certain date based on attestations of older or newer stages of the same word (e.g., predicting a non-documented Middle French word based on its Vulgar Latin and Old French predecessors and its Modern French successor). Morphological models and lexical diachronic evolution models will provide independent ways to perform the same predictions, thus reinforcing our hypotheses or pointing to new challenges.

The second application task is computational etymology and proto-language reconstruction. Our lexical diachronic evolution models will be paired with semantic resources (wordnets, word embeddings, and other corpus-based statistical information). This will allow us to formally validate or suggest etymological or cognate relations between lexical entries from different languages of a same language family, provided they are all inherited. Such an approach could also be adapted to include the automatic detection of borrowings from one language to another (e.g. for studying the non-inherited layers in the Ancient Greek lexicon). In the longer term, we will investigate the feasibility of the automatic (unsupervised) acquisition of phonetic change models, especially when provided with lexical data for numerous languages from the same language family.

These lines of research will rely on etymological data sets and standards for representing etymological information (see Section 3.4.2).

Diachronic evolution also applies to syntax, and in the context of the ANR project Profiterole, we are beginning to explore more or less automatic ways of detecting these evolutions and suggest modifications, relying on fine-grained syntactic descriptions (as provided by meta-grammars), unsupervised sentence clustering (generalising previous works on error mining, cf. [6]), and constraint relaxation (in meta-grammar classes). The underlying idea is that a new syntactic construction evolves from a more ancient one by small, iterative modifications, for instance by changing word order, adding or deleting functional words, etc.

3.3.4. Accessibility-related variation

Language variation does not always pertain to the textual input of NLP tools. It can also be characterised by their intended output. This is the perspective from which we investigate the issue of text simplification (for a

recent survey, see for instance [94]). Text simplification is an important task for improving the accessibility to information, for instance for people suffering from disabilities and for non-native speakers learning a given language [79]. To this end, guidelines have been developed to help writing documents that are easier to read and understand, such as the FALC ("Facile À Lire et à Comprendre") guidelines for French. [8]

Fully automated text simplification is not suitable for producing high-quality simplified texts. Besides, the involvement of disabled people in the production of simplified texts plays an important social role. Therefore, following previous works [67], [88], our goal will be to develop tools for the computer-aided simplification of textual documents, especially administrative documents. Many of the FALC guidelines can only be linguistically expressed using complex, syntactic constraints, and the amount of available "parallel" data (aligned raw and simplified documents) is limited. We will therefore investigate hybrid techniques involving rule-based, statistical and neural approaches based on parsing results (for an example of previous parsing-based work, see [58]). Lexical simplification, another aspect of text simplification [73], [80], will also be pursued. In this regard, we have already started a collaboration with Facebook's AI Research in Paris, the UNAPEI (the largest French federation of associations defending and supporting people with intellectual disabilities and their families), and the French Secretariat of State in charge of Disabled Persons.

Accessibility can also be related to the various presentation forms of a document. This is the context in which we have initiated the OPALINE project, funded by the *Programme d'Investissement d'Avenir - Fonds pour la Société Numérique*. The objective is for us to further develop the GROBID text-extraction suite [9] in order to be able to re-publish existing books or dictionaries, available in PDF, in a format that is accessible by visually impaired persons.

3.4. Modelling and Development of Language Resources

Language resources (raw and annotated corpora, lexical resources, etc.) are required in order to apply any machine learning technique (statistical, neural, hybrid) to an NLP problem, as well as to evaluate the output of an NLP system.

In data-driven, machine-learning-based approaches, language resources are the place where linguistic information is stored, be it implicitly (as in raw corpora) or explicitly (as in annotated corpora and in most lexical resources). Whenever linguistic information is provided explicitly, it complies to guidelines that formally define which linguistic information should be encoded, and how. Designing linguistically meaningful and computationally exploitable ways to encode linguistic information within language resources constitutes the first main scientific challenge in language resource development. It requires a strong expertise on both the linguistic issues underlying the type of resource under development (e.g. on syntax when developing a treebank) and the NLP algorithms that will make use of such information.

The other main challenge regarding language resource development is a consequence of the fact that it is a costly, often tedious task. ALMAnaCH members have a long track record of language resource development, including by hiring, training and supervising dedicated annotators. But a manual annotation can be speeded up by automatic techniques. ALMAnaCH members have also work on such techniques, and published work on approaches such as automatic lexical information extraction, annotation transfer from a language to closely related languages, and more generally on the use of pre-annotation tools for treebank development and on the impact of such tools on annotation speed and quality. These techniques are often also relevant for Research strand 1. For example, adapting parsers from one language to the other or developing parsers that work on more than one language (e.g. a non-lexicalised parser trained on the concatenation of treebanks from different languages in the same language family) can both improve parsing results on low-resource languages and speed up treebank development for such languages.

[8] Please click here for an archived version of these guidelines (at the time this footnote is begin written, the original link does not seem to work any more).

[9] https://github.com/kermitt2/grobid

3.4.1. Construction, management and automatic annotation of Text Corpora

Corpus creation and management (including automatic annotation) is often a time-consuming and technically challenging task. In many cases, it also raises scientific issues related for instance with linguistic questions (what is the elementary unit in a text?) as well as computer-science challenges (for instance when OCR or HTR are involved). It is therefore necessary to design a work-flow that makes it possible to deal with data collections, even if they are initially available as photos, scans, wikipedia dumps, etc.

These challenges are particularly relevant when dealing with ancient languages or scripts where fonts, OCR techniques, language models may be not extant or of inferior quality, as a result, among others, of the variety of writing systems and the lack of textual data. We will therefore work on improving print OCR for some of these languages, especially by moving towards joint OCR and language models. Of course, contemporary texts can be often gathered in very large volumes, as we already do within the ANR project SoSweet, resulting in different, specific issues.

ALMAnaCH pays a specific attention to the re-usability [10] of all resources produced and maintained within its various projects and research activities. To this end, we will ensure maximum compatibility with available international standards for representing textual sources and their annotations. More precisely we will take the TEI (*Text Encoding Initiative*) guidelines as well the standards produced by ISO committee TC 37/SC 4 as essential points of reference.

From our ongoing projects in the field of Digital Humanities and emerging initiatives in this field, we observe a real need for complete but easy work-flows for exploiting corpora, starting from a set of raw documents and reaching the level where one can browse the main concepts and entities, explore their relationship, extract specific pieces of information, always with the ability to return to (fragments of) the original documents. The pieces of information extracted from the corpora also need to be represented as knowledge databases (for instance as RDF "linked data"), published and linked with other existing databases (for instance for people and locations).

The process may be seen as progressively enriching the documents with new layers of annotations produced by various NLP modules and possibly validated by users, preferably in a collaborative way. It relies on the use of clearly identified representation formats for the annotations, as advocated within ISO TC 37/SC 4 standards and the TEI guidelines, but also on the existence of well-designed collaborative interfaces for browsing, querying, visualisation, and validation. ALMAnaCH has been or is working on several of the NLP bricks needed for setting such a work-flow, and has a solid expertise in the issues related to standardisation (of documents and annotations). However, putting all these elements in a unified work-flow that is simple to deploy and configure remains to be done. In particular, work-flow and interface should maybe not be dissociated, in the sense that the work-flow should be easily piloted and configured from the interface. An option will be to identify pertinent emerging platforms in DH (such as Transkribus) and to propose collaborations to ensure that NLP modules can be easily integrated.

It should be noted that such work-flows have actually a large potential besides DH, for instance for exploiting internal documentation (for a company) or exploring existing relationships between entities.

3.4.2. Development of Lexical Resources

ALPAGE, the Inriapredecessor of ALMAnaCH, has put a strong emphasis in the development of morphological, syntactic and wordnet-like semantic lexical resources for French as well as other languages (see for instance [5], [1]). Such resources play a crucial role in all NLP tools, as has been proven among other tasks for POS tagging [86], [83], [96] and parsing, and some of the lexical resource development will be targeted towards the improvement of NLP tools. They will also play a central role for studying diachrony in the lexicon, for example for Ancient to Contemporary French in the context of the Profiterole project. They will also be one of the primary sources of linguistic information for augmenting language models used in OCR systems for ancient scripts, and will allow us to develop automatic annotation tools (e.g. POS taggers) for low-resourced

[10] From a larger point of view we intend to comply with the so-called FAIR principles (http://force11.org/group/fairgroup/fairprinciples).

languages (see already [98]), especially ancient languages. Finally, semantic lexicons such as wordnets will play a crucial role in assessing lexical similarity and automating etymological research.

Therefore, an important effort towards the development of new morphological lexicons will be initiated, with a focus on ancient languages of interest. Following previous work by ALMAnaCH members, we will try and leverage all existing resources whenever possible such as electronic dictionaries, OCRised dictionaries, both modern and ancient [85], [70], [87], while using and developing (semi)automatic lexical information extraction techniques based on existing corpora [84], [89]. A new line of research will be to integrate the diachronic axis by linking lexicons that are in diachronic relation with one another thanks to phonetic and morphological change laws (e.g. XIIth century French with XVth century French and contemporary French). Another novelty will be the integration of etymological information in these lexical resources, which requires the formalisation, the standardisation, and the extraction of etymological information from OCRised dictionaries or other electronic resources, as well as the automatic generation of candidate etymologies. These directions of research are already investigated in ALMAnaCH [70], [87].

An underlying effort for this research will be to further the development of the GROBID-dictionaries software, which provides cascading CRF (Conditional Random Fields) models for the segmentation and analysis of existing print dictionaries. The first results we have obtained have allowed us to set up specific collaborations to improve our performances in the domains of a) recent general purpose dictionaries such as the Petit Larousse (Nénufar project, funded by the DGLFLF in collaboration with the University of Montpellier), b) etymological dictionaries (in collaboration with the Berlin Brandenburg Academy of sciences) and c) patrimonial dictionaries such as the Dictionnaire Universel de Basnage (an ANR project, including a PhD thesis at ALMAnaCH, has recently started on this topic in collaboration with the University of Grenoble-Alpes and the University Sorbonne Nouvelle in Paris).

In the same way as we signalled the importance of standards for the representation of interoperable corpora and their annotations, we will keep making the best use of the existing standardisation background for the representation of our various lexical resources. There again, the TEI guidelines play a central role, and we have recently participated in the "TEI Lex 0" initiative to provide a reference subset for the "Dictionary" chapter of the guidelines. We are also responsible, as project leader, of the edition of the new part 4 of the ISO standard 24613 (LMF, Lexical Markup Framework) dedicated to the definition of the TEI serialisation of the LMF model (defined in ISO 24613 part 1 'Core model', 2 'Machine Readable Dictionaries' and 3 'Etymology'). We consider that contributing to standards allows us to stabilise our knowledge and transfer our competence.

3.4.3. Development of Annotated Corpora

Along with the creation of lexical resources, ALMAnaCH is also involved in the creation of corpora either fully manually annotated (gold standard) or automatically annotated with state-of-the-art pipeline processing chains (silver standard). Annotations will either be only morphosyntactic or will cover more complex linguistic levels (constituency and/or dependency syntax, deep syntax, maybe semantics). Former members of the ALPAGE project have a renowned experience in those aspects (see for instance [92], [81], [91], [76]) and will participate to the creation of valuable resources originating from the historical domain genre.

Under the auspices of the ANR Parsiti project, led by ALMAnaCH (PI: DS), we aim to explore the interaction of extra-linguistic context and speech acts. Exploiting extra-linguistics context highlights the benefits of expanding the scope of current NLP tools beyond unit boundaries. Such information can be of spatial and temporal nature, for instance. They have been shown to improve Entity Linking over social media streams [65]. In our case, we decided to focus on a closed world scenario in order to study context and speech acts interaction. To do so, we are developing a multimodal data set made of live sessions of a first person shooter video game (Alien vs. Predator) where we transcribed all human players interactions and face expressions streamlined with a log of all in-game events linked to the video recording of the game session, as well as the recording of the human players themselves. The in-games events are ontologically organised and enable the modelling of the extra-linguistics context with different levels of granularity. Recorded over many games sessions, we already transcribed over 2 hours of speech that will serve as a basis for exploratory work, needed

for the prototyping of our context-enhanced NLP tools. In the next step of this line of work, we will focus on enriching this data set with linguistic annotations, with an emphasis on co-references resolutions and predicate argument structures. The midterm goal is to use that data set to validate a various range of approaches when facing multimodal data in a close-world environment.

4. Application Domains

4.1. Application domains for ALMAnaCH

ALMAnaCH's research areas cover Natural Language Processing (nowadays identified as a sub-domain of Artificial Intelligence) and Digital Humanities. Application domains are therefore numerous, as witnessed by ALMAnaCH's multiple academic and industrial collaborations, for which see the relevant sections. Examples of application domains for NLP include:

- Information extraction, information retrieval, text mining (ex.: opinion surveys)
- Text generation, text simplification, automatic summarisation
- Spelling correction (writing aid, post-OCR, normalisation of noisy/non-canonical texts)
- Machine translation, computer-aided translation
- Chatbots, conversational agents, question answering systems
- Medical applications (early diagnosis, language-based medical monitoring...)
- Applications in linguistics (modelling languages and their evolution, sociolinguistic studies...)
- Digital humanities (exploitation of text documents, for instance in historical research)

5. New Software and Platforms

5.1. Enqi

- Author: Benoît Sagot
- Contact: Benoît Sagot

5.2. SYNTAX

KEYWORD: Parsing
FUNCTIONAL DESCRIPTION: Syntax system includes various deterministic and non-deterministic CFG parser generators. It includes in particular an efficient implementation of the Earley algorithm, with many original optimizations, that is used in several of Alpage's NLP tools, including the pre-processing chain Sx Pipe and the LFG deep parser SxLfg . This implementation of the Earley algorithm has been recently extended to handle probabilistic CFG (PCFG), by taking into account probabilities both during parsing (beam) and after parsing (n-best computation).

- Participants: Benoît Sagot and Pierre Boullier
- Contact: Pierre Boullier
- URL: http://syntax.gforge.inria.fr/

5.3. FRMG

KEYWORDS: Parsing - French

FUNCTIONAL DESCRIPTION: FRMG is a large-coverage linguistic meta-grammar of French. It can be compiled (using MGCOMP) into a Tree Adjoining Grammar, which, in turn, can be compiled (using DyALog) into a parser for French.

- Participant: Éric Villemonte De La Clergerie
- Contact: Éric De La Clergerie
- URL: http://mgkit.gforge.inria.fr/

5.4. MElt

Maximum-Entropy lexicon-aware tagger
KEYWORD: Part-of-speech tagger
FUNCTIONAL DESCRIPTION: MElt is a freely available (LGPL) state-of-the-art sequence labeller that is meant to be trained on both an annotated corpus and an external lexicon. It was developed by Pascal Denis and Benoît Sagot within the Alpage team, a joint Inria and Université Paris-Diderot team in Paris, France. MElt allows for using multiclass Maximum-Entropy Markov models (MEMMs) or multiclass perceptrons (multitrons) as underlying statistical devices. Its output is in the Brown format (one sentence per line, each sentence being a space-separated sequence of annotated words in the word/tag format).

MElt has been trained on various annotated corpora, using Alexina lexicons as source of lexical information. As a result, models for French, English, Spanish and Italian are included in the MElt package.

MElt also includes a normalization wrapper aimed at helping processing noisy text, such as user-generated data retrieved on the web. This wrapper is only available for French and English. It was used for parsing web data for both English and French, respectively during the SANCL shared task (Google Web Bank) and for developing the French Social Media Bank (Facebook, twitter and blog data).

- Contact: Benoît Sagot
- URL: https://team.inria.fr/almanach/melt/

5.5. dyalog-sr

KEYWORDS: Parsing - Deep learning - Natural language processing
FUNCTIONAL DESCRIPTION: DyALog-SR is a transition-based dependency parser, built on top of DyALog system. Parsing relies on dynamic programming techniques to handle beams. Supervised learning exploit a perceptron and aggressive early updates. DyALog-SR can handle word lattice and produce dependency graphs (instead of basic trees). It was tested during several shared tasks (SPMRL'2013 and SEMEVAL'2014). It achieves very good accuracy on French TreeBank, alone or by coupling with FRMG parser. In 2017, DyALog-SR has been extended into DyALog-SRNN by adding deep neuronal layers implemented with the Dynet library. The new version has participated to the evaluation campaigns CONLL UD 2017 (on more than 50 languages) and EPE 2017.

- Contact: Éric De La Clergerie

5.6. Crapbank

French Social Media Bank
KEYWORDS: Treebank - User-generated content
FUNCTIONAL DESCRIPTION: The French Social Media Bank is a treebank of French sentences coming from various social media sources (Twitter(c), Facebook(c)) and web forums (JeuxVidéos.com(c), Doctissimo.fr(c)). It contains different kind of linguistic annotations: - part-of-speech tags - surface syntactic representations (phrase-based representations) as well as normalized form whenever necessary.

- Contact: Djamé Seddah

5.7. DyALog

KEYWORD: Logic programming
FUNCTIONAL DESCRIPTION: DyALog provides an environment to compile and execute grammars and logic programs. It is essentially based on the notion of tabulation, i.e. of sharing computations by tabulating traces of them. DyALog is mainly used to build parsers for Natural Language Processing (NLP). It may nevertheless be used as a replacement for traditional PROLOG systems in the context of highly ambiguous applications where sub-computations can be shared.

- Participant: Éric Villemonte De La Clergerie
- Contact: Éric Villemonte De La Clergerie
- URL: http://dyalog.gforge.inria.fr/

5.8. SxPipe

KEYWORD: Surface text processing
SCIENTIFIC DESCRIPTION: Developed for French and for other languages, Sx Pipe includes, among others, various named entities recognition modules in raw text, a sentence segmenter and tokenizer, a spelling corrector and compound words recognizer, and an original context-free patterns recognizer, used by several specialized grammars (numbers, impersonal constructions, quotations...). It can now be augmented with modules developed during the former ANR EDyLex project for analysing unknown words, this involves in particular (i) new tools for the automatic pre-classification of unknown words (acronyms, loan words...) (ii) new morphological analysis tools, most notably automatic tools for constructional morphology (both derivational and compositional), following the results of dedicated corpus-based studies. New local grammars for detecting new types of entities and improvement of existing ones, developed in the context of the PACTE project, will soon be integrated within the standard configuration.
FUNCTIONAL DESCRIPTION: SxPipe is a modular and customizable processing chain dedicated to applying to raw corpora a cascade of surface processing steps (tokenisation, wordform detection, non-deterministic spelling correction. . .). It is used as a preliminary step before ALMAnaCH's parsers (e.g., FRMG) and for surface processing (named entities recognition, text normalization, unknown word extraction and processing...).

- Participants: Benoît Sagot, Djamé Seddah and Éric Villemonte De La Clergerie
- Contact: Benoît Sagot
- URL: http://lingwb.gforge.inria.fr/

5.9. Mgwiki

KEYWORDS: Parsing - French
FUNCTIONAL DESCRIPTION: Mgwiki is a linguistic wiki that may used to discuss linguistic phenomena with the possibility to add annotated illustrative sentences. The work is essentially devoted to the construction of an instance for documenting and discussing FRMG, with the annotations of the sentences automatically provided by parsing them with FRMG. This instance also offers the possibility to parse small corpora with FRMG and an interface of visualization of the results. Large parsed corpora (like French Wikipedia or Wikisource) are also available. The parsed corpora can also be queried through the use of the DPath language.

- Participant: Éric Villemonte De La Clergerie
- Contact: Éric Villemonte De La Clergerie
- URL: http://alpage.inria.fr/frmgwiki/

5.10. WOLF

WOrdnet Libre du Français (Free French Wordnet)
KEYWORDS: WordNet - French - Semantic network - Lexical resource

FUNCTIONAL DESCRIPTION: The WOLF (Wordnet Libre du Français, Free French Wordnet) is a free semantic lexical resource (wordnet) for French.

The WOLF has been built from the Princeton WordNet (PWN) and various multilingual resources.

- Contact: Benoît Sagot
- URL: http://alpage.inria.fr/~sagot/wolf-en.html

5.11. vera

KEYWORD: Text mining
FUNCTIONAL DESCRIPTION: Automatic analysis of answers to open-ended questions based on NLP and statistical analysis and visualisation techniques (vera is currently restricted to employee surveys).

- Participants: Benoît Sagot and Dimitri Tcherniak
- Partner: Verbatim Analysis
- Contact: Benoît Sagot

5.12. Alexina

Atelier pour les LEXiques INformatiques et leur Acquisition
KEYWORD: Lexical resource
FUNCTIONAL DESCRIPTION: Alexina is ALMAnaCH's framework for the acquisition and modeling of morphological and syntactic lexical information. The first and most advanced lexical resource developed in this framework is the Lefff, a morphological and syntactic lexicon for French.

- Participant: Benoît Sagot
- Contact: Benoît Sagot
- URL: http://gforge.inria.fr/projects/alexina/

5.13. FQB

French QuestionBank
KEYWORD: Treebank
FUNCTIONAL DESCRIPTION: The French QuestionBanks is a corpus of around 2000 questions coming from various domains (TREC data set, French governmental organisation, NGOs, etc..) it contains different kind of annotations - morpho-syntactic ones (POS, lemmas) - surface syntaxe (phrase based and dependency structures) with long-distance dependency annotations.

The TREC part is aligned with the English QuestionBank (Judge et al, 2006).

- Contact: Djamé Seddah

5.14. Sequoia corpus

KEYWORD: Treebank
FUNCTIONAL DESCRIPTION: The Sequoia corpus contains French sentences, annotated with various linguistic information: - parts-of-speech - surface syntactic representations (both constituency trees and dependency trees) - deep syntactic representations (which are deep syntactic dependency graphs)

- Contact: Djamé Seddah

6. New Results

6.1. Syntax modelling and treebank development

Participants: Djamé Seddah, Benoît Sagot, Éric Villemonte de La Clergerie, Emilia Verzeni, Wigdan Abbas Mekki Medeni, Elias Benaïssa, Farah Essaidi, Amal Fethi.

- In 2018, members of ALMAnaCH have finalised a conversion of the biggest annotated data set for French, the French Treebank, to Universal Dependencies 2.3, the now *de facto* standard for syntactic annotations [27]. The same group was also deeply involved in a proposal co-written with others leaders of the field [25], aiming at representing morpho-syntactic ambiguities from user-generated content and morphologically-rich languages. This proposal was implemented via the development of language specific analysers and data-driven normalised lexica [26].
- As part of the ANR Parsiti project, the development of gold standards for North-African dialectal Arabic has seen great progresses and is coming to a pre-release date in the first semester of 2019. This work involved more than 24 man.months over the last 12 months and will culminate with a multi-layered corpus of about 2000 sentences that is made of user-generated content with a highly variable dialect that contains up to 36% of French words and mixed syntax with Arabic. In order to assess the quality of the translation produced by the Parsiti project, we also included a translation layer (North-African Arabic-French) as well as all expected morpho-syntactic and syntactic annotations, following the state-of-the-art in terms of annotations. Papers are currently being written and will target the main NLP conferences of early 2019.
- In parallel to the last item, we also translated to English half of the French Social Media Bank which was developed in our previous project [92]. A morpho-syntactic annotation layer was added. The crucial difficulty was to maintain a symmetry in term of style and level of languages between French user-generated content and its English counterpart. This data set is currently being used in the Parsiti project in order to evaluate the MT models currently being developed by the LIMSI partner.

6.2. Modeling of language variability via diachronic embeddings and extra-linguistic contextual features

Participants: Djamé Seddah, Benjamin Muller, Ganesh Jawahar, Benoît Sagot, Éric Villemonte de La Clergerie.

Following ALMAnaCH's participation in the 2017 CoNLL shared task on heavily multilingual dependency parsing in the *Universal Dependency* (hereafter UD) framework (we ranked 3rd/33 on part-of-speech tagging and 6th/33 on parsing), the team has taken part in the 2018 edition of the shared task. This year, most of the work was carried out by junior members of the team, for whom it was an interesting opportunity to gain experience on the development of NLP architectures and their deployment in the context of a shared task. It was also the opportunities to test new ideas.

We developed a neural dependency parser and a neural part-of-speech tagger, which we called 'ELMoLex' [21]. We augmented the deep Biaffine (BiAF) parser [64] with novel features to perform competitively: we utilize an in-domain version of ELMo features [77], which provide context-dependent word representations; we utilised disambiguated, embedded, morphosyntactic features extracted from our UD-compatible lexicons [26], which complements the existing feature set. In addition to incorporating character embeddings, ELMoLex leverages pre-trained word vectors, ELMo and morphosyntactic features (whenever available) to correctly handle rare or unknown words which are prevalent in languages with complex morphology. ELMoLex ranked 11th in terms of the Labeled Attachment Score metrics (70.64%) and the Morphology-aware LAS metrics (55.74%), and ranked 9th in terms of Bilexical dependency metric (60.70%). In an extrinsic evaluation setup, ELMoLex ranked 7th for Event Extraction, Negation Resolution tasks and 11th for Opinion Analysis task in terms of F1 score.

6.3. Modelling of language variability via diachronic embeddings and extra-linguistic contextual features

Participants: Djamé Seddah, Ganesh Jawahar, Éric Villemonte de La Clergerie, Benoît Sagot.

As part of the ANR SoSweet and the PHC Maimonide projects (in collaboration with Bar Ilan University for the latter), ALMAnaCH has invested a lot of efforts in 2018 into studying language variability (i.e. how the language evolve over time and how this evolution is tied to socio-demographic and dynamic network variables). Taking advantages of the SoSweet corpus (220 millions tweet) and of the Bar Ilan Hebrew Tweets (180M tweets) both collected over the last 5 years, we have been addressing the problem of studying semantic changes. We devised a novel attentional model, based on Bernouilli word embeddings, that are conditioned on contextual extra-linguistic (social) features such as network, spatial and socio-economic variables, which are associated with Twitter users, as well as topic-based features. We posit that these social features provide an inductive bias that is susceptible to helping our model to overcome the narrow time-span regime problem. Our extensive experiments reveal that, as a result of being less biased towards frequency cues, our proposed model was able to capture subtle semantic shifts and therefore benefits from the inclusion of a reduced set of contextual features. Our model thus fit the data better than current state-of-the-art dynamic word embedding models and therefore is a promising tool to study diachronic semantic changes over small time periods. A paper on this work is currently under review.

6.4. Standardisation of Natural Language data

Participants: Laurent Romary, Jack Bowers, Charles Riondet, Mohamed Khemakhem, Benoît Sagot, Loïc Grobol.

One essential aspect of working with human traces as they occur in digital humanities at large and in natural language processing in particular, is to be able to re-use any kind of primary content and further enrichments thereof. The central aspect of re-using such content is the development and applications of reference standards that reflect the best state of the art in the corresponding domains. In this respect, our team is particularly attentive to the existing standardisation background when both producing language resources or devloping NLP components. Furthermore, our specific leading roles in the domain of standardisation in both the Parthenos and EHRI EU projects as well as in related initiatives (TEI consortium, ISO committee TC 37, DARIAH lexical working group) has allowed to make progress along the following lines:

- Contributing to the revision of the ISO 24613 standard (Lexical Markup Framework) in the form of a multipart standard covering, for the time being, the core model (ISO 24613-1), machine readable dictionaries (ISO 24613-2), etymology (ISO 24613-3) and a TEI based serialisation (ISO 24613-4). Several members of the team have been particularly active as experts in the definition of the first two parts, which are now at publication and DIS stage respectively [11] and are co-editors of parts 3 and 4;
- Proposal for a reference TEI subset for integrating dictionary content: in the context of the DARIAH working group on lexical resources, a first release of the *TEI Lex 0* [12] was issued in September 2018 integrating the continuous work of the group over the the 2016-2018 period and already taken up by the infrastructure project ELEXIS [13] as its reference back-office format. This work is also the basis for the output format of Grobid-Dictionaries [71];
- Finalisation of the ISO proposal on reference annotation (ISO 24617-9): the team has been leading the work on the definition of the Reference Annotation Framework (RAF) [14] which is now at DIS ballot stage and already implemented in several concrete annotation projects[19], [43]. The standard is feature complete from a linguistic point of view (from simple co-reference to complex bridging anaphora phenomena) and compliant with the TEI stand-off annotation module [59] from the point of view of its implementation [66];

[11] See the ISO/TC 37/SC 4 work current work program under https://www.iso.org/committee/297592/x/catalogue/p/0/u/1/w/0/d/0
[12] https://github.com/DARIAH-ERIC/lexicalresources
[13] https://elex.is
[14] https://www.iso.org/standard/69658.html

- Large-scale implementation of international standard for the documentation of the Mixtepec-Mixtec language (see section 6.11);
- Proposing a customisation architecture for the EAD international standard: EAD (Encoding Archival Description [15]) is used worldwide in cultural heritage institution to describe and exchange collection level information. In the context of the EHRI project, where we had to design a mechanism for integrating heterogeneous implementations of EAD-based data, we used the TEI ODD specification language to re-design and subset the international EAD specification to precisely provide interoperability conditions within the project[14];
- Release of the SSK (Standardisation Survival Kit), a generic environment for describing standards-based digital humanities research scenarios: the SSK is an online platform for describing research scenarios developed within the Parthenos project[40] and now deployed as a service hosted by the French national Huma-Num infrastructure [16]. The SSK has been developed as a completely open project [17], where the scenarios are themselves described as TEI-based representations[51], [35], [50].

6.5. Entity-fishing: a generic named entity recognition and disambiguation for digital humanities projects

Participants: Marie Puren, Charles Riondet, Laurent Romary, Luca Foppiano, Tanti Kristanti.

Since several years (starting at the beginning of the EU Cendari project in 2012 [75]) we have been working on the provision of a generic named-entity recognition and disambiguation module (NERD) called *entity-fishing*[18] as a stable on-line service. The work we have achieved demonstrates the possible delivery of sustainable technical services as part of the development of research infrastructures for the humanities in Europe. In particular, our results contribute not only to DARIAH, the European digital research infrastructure for the arts and humanities, but also to OPERAS, the European research infrastructure for the development of open scholarly communication in the social sciences and humanities. Deployed as part of the French national infrastructure Huma-Num, the service provides an efficient state-of-the-art implementation coupled with standardised interfaces allowing easy deployment in a variety of potential digital humanities contexts. In 2018, we have specifically integrated *entity-fishing* within the H2020 HIRMEOS project where several open access publishers have used the service in their collections of published monographs as a means to enhance retrieval and access.

To this end, we have set up a common layer of services on top of several existing e-publishing platforms for Open Access monographs. The *entity extraction* task was deployed over a corpus of monographs provided by the HIRMEOS partners, with the following coverage:

- 4000 books in English and French from Open Edition Books
- 2000 titles in English and German from OAPEN
- 162 books in English from Ubiquity Press
- 765 books (606 in German, 159 in English) from the University of Göttingen

The introduction of *entity-fishing* has undergone different levels of integration. The majority of the participating publishers provided additional features in their user interface, using the data generated by *entity-fishing*, for example, as search facets for persons and locations to help users narrow down their searches and obtain more precise results.

entity-fishing has been developed in Java and it has been designed for fast processing on text and PDF, with relatively limited memory and to offer relatively close to state-of-the-art accuracy (as compared with other NERD systems). The accuracy f-score for disambiguation is currently between 76.5 and 89.1 on standard datasets (ACE2004, AIDA-CONLL-testb, AQUAINT, MSNBC) (Table 1) [74].

[15] https://en.wikipedia.org/wiki/Encoded_Archival_Description
[16] http://ssk.huma-num.fr
[17] https://github.com/ParthenosWP4/SSK

Table 1. Accuracy measures

	ACE 2004	AIDA CONLL-testb	AQUAINT	MSNBC
Priors	83.1	66.1	80.3	71.1
entity-fishing	83.5	76.5	**89.1**	86.7
Wikifier	83.4	77.7	86.2	85.1
DoSeR	**90.7**	78.4	84.2	91.1
AIDA	81.5	77.4	53.2	78.2
Spotlight	71.3	59.3	71.3	51.1
Babelfy	56.1	59.2	65.2	60.7
WAT	80.0	84.3	76.8	77.7
(Ganea & Hofmann, 2017)	88.5	**92.2**	88.5	**93.7**

The objective, however, is to provide a generic service that has a steady throughput of 500-1000 words per second or one PDF page of a scientific article in 1-2 seconds on a medium range (4CPU, 3Gb Ram) Linux server.

From the point of view of the technical deployment itself, we have provided all the necessary components of a sustainable service:

- release and publish *entity-fishing* as open source software [18];
- deploy the service in the DARIAH infrastructure through HUMA-NUM [19];
- produce evaluation data and metrics for content validation.

6.6. From GROBID to GROBID-Dictionaries

Participants: Luca Foppiano, Mohamed Khemakhem, Laurent Romary, Pedro Ortiz Suárez, Alba Marina Malaga Sabogal.

GROBID is an open source software suite initiated in 2007 by Patrice Lopez with the purpose of extracting metadata automatically from scholarly papers available in PDF. Over the years, it has developed into a rich information extraction environment, and deployed in many Inria projects, but also national and international services, such as HAL (front-end meta-data extraction from uploaded scholarly publications). It is a central piece for our information extraction activities and we have been particularly active in 2018 in the following domains:

- General contributions to GROBID [20]:
  - Major refactoring and design improvements
  - fixes, tests, documentation and update of the pdf2xml fork for Windows
  - added and improved several models in collaboration with CERN (e.g. for the recognition of arXiv identifier)
  - Further tests on the specific case of bibliographic documents[32]
- Contribution to GROBID-Dictionaries [21]: the lexical GROBID extension has been implemented and tested on modern and multilingual dictionaries[23]. In the context of several collaborative activities, GROBID-Dictionaries has been applied on several documentary sources:
  - Early editions of the The Petit Larrousse Illustré in the context of the Nénufar project[45], [29]

[18] http://github.com/kermitt2/nerd
[19] http://nerd.huma-num.fr/nerd/
[20] https://github.com/kermitt2/grobid
[21] https://github.com/MedKhem/grobid-dictionaries

- Further experiments on etymological dictionaries from the Berlin Brandenburg Academy of Sciences
- Experiments on entry-based documents such as manuscript catalogues (with University of Neuchâtel)[16] and the French address Directory Bottin from the end of the XIXth Century[22]

These various experiments have been accompanied by an intense training and hand-on activity in the context in particular of the French research network CAHIERS (Huma-Num consortium), the Lexical Data Master Class and a series of workshop organised in South Africa under the auspices of a national linguistic documentation program. Finally, further alignments with the ongoing standardisation activities around TEI Lex0 and ISO 24613 (LMF) has been carried out to ensure a proper standards compliance of the generated output

The experience gained in the develoment and application of GROBID-Dictionaries has been the basis for the recently accepted ANR BASNUM project which aims at automatically structuring and enriching of the Dictionnaire universel (DU) by Antoine Furetière, in its 1701 edition rewritten by Basnage de Beauval and the doctoral work of Pedro Ortiz.

6.7. Resources, models and tools for coreference resolution

Participants: Loïc Grobol, Éric Villemonte de La Clergerie.

This year we performed many experiments, some of them detailed in [28], targeting end-to-end coreference systems for spontaneous oral French. More precisely, for several mention-pair coreference detection models, we tried to assess their sensibility to various features of coreference chains and their viability for end-to-end systems, compared to the more recent antecedent scoring models.

Also, one of our objective being to assess the usefulness of syntactic features for coreference detection, we enriched the coreference annotations of the ANCOR corpus with both automatically produced dependency syntax annotations and improved speech transcription. All these annotation were wrapped in a TEI-compliant XML format as described in [20] (see also 6.4).

Finally, we have been working on neural architectures for coreference detection, building upon some recent state of the art techniques. They are based on embeddings for general text span and we try to make them more scalable through efficient uses of the local context but also more tunable to different document types and language variation. The base idea is to complete pre-training by training on related lower-level tasks such as entity-mention detection.

6.8. Computational history through information extraction from archive texts

Participants: Éric Villemonte de La Clergerie, Marie Puren, Charles Riondet, Alix Chagué, Marie-Laurence Bonhomme.

From two different DH projects emerged some interesting research questions related to the extraction of information from archival documents, in particular the management of the diversity of document types and structures and on the contrary the acquisition of detailed information from a regular visual structure.

In the context of the ANR TIME-US, whose goal is to reconstruct the "time-budgets" of textile workers in France (18th - early 20th centuries), we worked on the creation of a digitization workflow to acquire structured textual data from a wide range of printed and handwritten materials: professional court records (like *Prud'Hommes*), Police reports on strikes or early sociological studies such as the *Monographies de Le Play* This workflow has been presented at the ADHO DH conference in Mexico (see the presentation here: [34].The set up of this workflow is a prerequisite for further experiments and processing to extract information that can be exploited by historians, such as the relation between working tasks, the time spent by workers to perform them and the price they are paid for this time.

Another project was intiated in collaboration with the EPHE and the French National Archives, in the framework of the convention signed between Inria and the Ministry of Culture. This project is called LECTAUREP (for *LECTure AUtomatique de REPertoires*, and is aimed at extracting the information recorded in the registries of Parisian notaries, held by the National Archives. This project it at the intersection of NLP and Computer Vision because one of the main objectives is to extract information from the physical layout of the documents, presented as tables. Another issue is to be able to recognize with accuracy an important diversity of handwritten scripts. The final goal of LECTAUREP is to give access to researchers the information contained in these records, in particular the name of the persons involved in cases recorded by notaries, their addresses and the nature of the case (wills, powers of attorney, wedding contracts, etc.). An initial report has been produced (see [39]), and the project will continue in 2019 with the release of the extracted information (named entities, geolocation, typology, etc) into a structured database.

6.9. Discovering correlations between parser features and neurological observations

Participants: Éric Villemonte de La Clergerie, Murielle Fabre, Pauline Brunet.

In the context of the CRCNS international network, the ANR-NSF NCM-ML project (dubbed "*Petit Prince* project") aims to discover and explore correlations between features (or predictors) provided by NLP tools such as parsers, and fMRI data resulting from listening of the novel Le Petit Prince.

In 2018, Pauline Brunet, during her Master thesis, has worked on developing the infrastructure (scripts and formats) for the integration of the features, and the use of these features for computing correlations with fMRI data. A first set of features has been identified and collected from the novel and from its processing by ALMAnaCH tools (namely FRMG as an instance of a symbolic TAG-based parser and Dyalog-SR, as an instance of an hybrid feature-based neural-based dependency parser). A first dataset of fMRI scan was received to assess the infrastructure and get some preliminary results.

The work is now being continued with the arrival as a post-doc of Murielle Fabre (November 2018). With the expected arrival of the second half of the scans, she will explore more features, use her expertise to interpret the correlations, and guide the choice of new features to be tested. Since her arrival, she has in particular focused on Multi-Word Expressions (MWEs), in particular to be comparable with results published on the English side of the project. We have also identified several kinds of parsing architectures to test, in relation with various complexity parameters: (1) LSTM (two layers), (2) RNNG (with a partile filter), (3) Dyalog-SR et (4) FRMG (TAG).

In order to be in phase (and comparable) with our US partners, we have started to assemble two French corpora: - a small corpus for domain adaptation to children's books: it will permit the fine tuning of the different parsers to a great amount of dialogues an Q&A present in *Le Petit Prince*. - a large corpus of Contemporary French oral transcriptions and texts to calculate lexical association measures (AM) like PMI (Point-wise Mutual information) or Dice scores on the MWEs found in *Le Petit Prince*. This corpus of approx. 600 millions words represents a balanced counterpart to the American COCA corpus. [22]

Both Éric de La Clergerie and Murielle Fabre attended the annual meeting of the CRCNS network (Berkeley, June 2018).

6.10. Evaluating the quality of text simplification

Participants: Louis Martin, Benoît Sagot, Éric Villemonte de La Clergerie.

[22] https://corpus.byu.edu/coca/

In 2018, our collaboration on text simplification with the Facebook Artificial Intelligence Research lab in Paris (in particular with Antoine Bordes) has started in practice. It has taken the form of a CIFRE PhD. In this context, in 2018, we dedicated important efforts to the problem of the evaluation of text simplification (TS) systems, which remains an open challenge. As the task has common points with machine translation (MT), TS is often evaluated using MT metrics such as BLEU. However, such metrics require high quality reference data, which is rarely available for TS. TS has the advantage over MT of being a monolingual task, which allows for direct comparisons to be made between the simplified text and its original version.

We compared multiple approaches to reference-less quality estimation of sentence-level TS systems, based on the dataset used for the QATS 2016 shared task. We distinguished three different dimensions: grammaticality, meaning preservation and simplicity. We have shown that n-gram-based MT metrics such as BLEU and METEOR correlate the most with human judgment of grammaticality and meaning preservation, whereas simplicity is best evaluated by basic length-based metrics [24].

6.11. Advances in descriptive, computational and historical linguistics

Participants: Benoît Sagot, Laurent Romary, Jack Bowers, Rebecca Blevins.

ALMAnaCH members have resumed their work in descriptive, computational and historical linguistics, an important way to ensure that NLP models and tools are robust to the diversity of world languages, as well as a way to apply NLP models and tools for contributing to research in linguistics. Three of 2018 advances in this regard are the following:

- In the context of the doctoral work of Jack Bowers, a first release of a global documentation of the Mixtepec-Mixtec language has been released which covers, multilayered annotated spoken and written resources as well as a reference lexical resource covering both basic word descriptions and elaborate semantic and etymological (word formation) content [13];
- Work on language description and computational morphology for Romansh Tuatschin in collaboration with Géraldine Walther (Universität Zürich) was pursued, following the work published in 2017 [99]. A new interest in the quantitative, corpus-based study of code switching in this language has emerged in collaboration with Claudia Cathomas (Universität Zürich), leading to preliminary results to be published in 2019;
- We resumed our work in (classical) etymology in collaboration with Romain Garnier (Université de Limoges, Institut Universitaire de France), with a focus not only on (Ancient) Greek and its substrates, but also, more specifically, on Anatolian languages that could be amongst said substrates. In particular, we proposed that Lydian could be the source language for a number of Greek words lacking a good etymology in the literature [31], which motivated Rebecca Blevins's internship on the development of a lexicon of the Lydian language. We also published new etymological results at the (Proto-)Indo-European level [37].

6.12. Language resources and NLP tools for Medieval French

Participants: Éric Villemonte de La Clergerie, Mathilde Regnault, Benoît Sagot.

The main objectives of the ANR project "Profiterole" are to automatically annotate a large corpus of medieval French (9th-15th centuries) in dependency syntax and to provide a methodology for dealing with heterogeneous data like such a corpus (because of diachronic, dialectal, geographic, stylistic and genre-based variation, among other types of linguistic variation). To this end, we have continued previous experiments in morpho-syntactic tagging by trying to determine which parameters and which training sets are the best ones to use when annotating a new text. We explored two approaches for syntactic annotation (i.e. parsing). On the one hand, an ongoing thesis aims at adapting the FRMG metagrammar to medieval French, notably by changing the constraints on certain syntactic phenomena and relaxing the order of words. The development of the OFrLex lexicon has started within the Alexina framework, following the Lefff lexicon for contemporary French [5]. It already allowed for preliminary experiments. On the other hand, we conducted parsing experiments with neural models (DyALog's SRNN models). Note that members of the ALMAnaCH team participated in the CoNLL dependency parsing Shared Task 2018, which included an Old French dataset (see section 6.2).

7. Bilateral Contracts and Grants with Industry

7.1. Industrial Collaborations

- **Verbatim Analysis**: this Inria start-up was co-created in 2009 by BS. It uses some of ALMAnaCH's free NLP software (SxPipe) as well as a data mining solution co-developed by BS, VERA, for processing employee surveys with a focus on answers to open-ended questions.
- **opensquare** A new Inria startup, opensquare, was co-created in December 2016 by BS with 2 senior specialists in HR consulting. opensquare designs, carries out and analyses employee surveys and offers HR consulting based on these results. It uses a new employee survey analysis tool, *enqi*, which is still under development.
- **Facebook**: A collaboration on text simplification ("français Facile À Lire et à Comprendre", FALC) is ongoing with Facebook's Parisian FAIR laboratory. In particular, a co-supervised (CIFRE) PhD thesis started in 2018 in collaboration with UNAPEI, the largest French federation of associations defending and supporting people with special needs and their families. This collaboration is expected to pave the way for a larger initiative involving (at least) these three partners as well as the relevant ministries.
- **Bluenove**: A contract with this company has been signed, which defines the framework of our collaboration on the integration of NLP tools (e.g. chatbot-related modules) within Bluenove's plateform Assembl, dedicated to online citizen debating forums. It involves a total of 24 months of fixed-term contracts (12 months for a post-doc, currently working on the project, and 12 months for a research engineer, which is still to be recruited).
- **Science Miner**: ALMAnaCH (previously ALPAGE) has been collaborating since 2014 years with this company founded by Patrice Lopez, a specialist in machine learning techniques and initiator of the GROBID and NERD (now entity-fishing) suites. Patrice Lopez provides scientific support on the corresponding software components in the context of the Parthenos, EHRI and Iperion projects, as well as in the context of the Inria anHALytics initiative, aiming at providing a scholarly dashboard on the scientific papers available from the HAL national publication repository.
- **Hyperlex** A collaboration was initiated in 2018 on NLP for legal documents, involving especially EVdLC.
- ALMAnaCH members led a proposal for the creation of an ANR LabCom with Fortia Financial Solutions, a company specialized in *Financial Technology*, namely the analysis of financial documents from investment funds. The proposal has been rejected. Meanwhile, this project is currently being extended toward a FUI with Systran, the market leader in specialized machine translation systems, and the BNP as industrial partner. The funding requested will cross the 3 millions euros bar.
- ALMAnaCH members have recently initiated discussions with other companies (Louis Vuitton, Suez...), so that additional collaborations might start in the near future. They have also presented their work to companies interested in knowing more about the activities of Inria Paris in AI and NLP.

8. Partnerships and Cooperations

8.1. National Initiatives

8.1.1. ANR

- **ANR SoSweet** (2015-2019, PI J.-P. Magué, resp. ALMAnaCH: DS; Other partners: ICAR [ENS Lyon, CRNS], Dante [Inria]). Topic: studying sociolinguistic variability on Twitter, comparing linguistic and graph-based views on tweets

- **ANR ParSiTi** (2016-2021, PI Djamé Seddah, Other partners: LIMSI, LIPN). Topic: context-aware parsing and machine translation of user-generated content
- **ANR PARSE-ME** (2015-2020, PI. Matthieu Constant, resp. Marie Candito [ALPAGE, then LLF], ALMAnaCH members are associated with Paris-Diderot's LLF for this project). Topic: multi-word expressions in parsing
- **ANR Profiterole** (2016-2020, PI Sophie Prévost [LATTICE], resp. Benoit Crabbé [ALPAGE, then LLF], ALMAnaCH members are associated with Paris-Diderot's LLF for this project). Topic: modelling and analysis of Medieval French
- **ANR TIME-US** (2016-2019, PI Manuela Martini [LARHRA], ALMAnaCH members are associated with Paris-Diderot's CEDREF for this project). Topic: Digital study of remuneration and time budget textile trades in XVIIIth and XIXth century France
- **ANR BASNUM** (2018-2021, PI Geoffrey Williams [Université Grenoble Alpes], resp. ALMAnaCH: LR). Topic: Digitalisation and computational linguistic study of Basnage de Beauval's *Dictionnaire universel* published in 1701.

8.1.2. Competitivity Clusters

- **LabEx EFL** (2010-2019, PI Christian Puech [HTL, Paris 3], Sorbonne Paris Cité). Topic: empirical foundations of linguistics, including computational linguistics and natural language processing. ALPAGE was one of the partner teams of this LabEx, which gathers a dozen of teams within and around Paris whose research interests include one aspects of linguistics or more. BS serves as deputy head (and former head) of one of the scientific strands of the LabEx, namely strand 6 dedicated to language resources. BS and DS are in charge of a number of scientific "operations" within strands 6, 5 ("computational semantic analysis") and 2 ("experimental grammar"). BS, EVdLC and DS are now individual members of the LabEx EFL since 1st January 2017, and BS still serves as the deputy head of strand 6. Main collaborations are on language resource development (strands 5 and 6), syntactic and semantic parsing (strand 5, especially with LIPN [CNRS and U.Paris 13]) and computational morphology (strands 2 and 6, especially with CRLAO [CNRS and Inalco]).

8.1.3. Other National Initiatives

- **LECTAUREP project** (2017-2018): An explorative study has been launched in collaboration with the National Archives in France, in the context of the framework agreement between Inria and the Ministry of Culture, to explore the possibility of extracting various components from digitized 19th Century notary registers.
- **Nénufar (DGLFLF - Délégation générale à la langue française et aux langues de France)**: The projects is intended to digitize and exploit the early editions (beginning of the 20th Century) of the Petit Larousse dictionary. ALMAnaCH is involve to contribute to the automatic extraction of the dictionary content by means of GROBID-Dictionaries and define a TEI compliant interchange format for all results.
- **PIA Opaline**: The objective of the project is to provide a better access to published French literature and reference material for visually impaired persons. Financed by the Programme d'Investissement d'Avenir, it will integrate technologies related to document analysis and re-publishing, textual content enrichment and dedicated presentational interfaces. Inria participate to deploy the GROBID tool suite for the automatic structuring of content from books available as plain PDF files.

8.2. European Initiatives

8.2.1. FP7 & H2020 Projects

- **H2020 Parthenos** (2015-2019, PI Franco Niccolucci [University of Florence]; LR is a work package coordinator) Topic: strengthening the cohesion of research in the broad sector of Linguistic Studies, Humanities, Cultural Heritage, History, Archaeology and related fields through a thematic cluster of European Research Infrastructures, integrating initiatives, e-infrastructures and other world-class infrastructures, and building bridges between different, although tightly interrelated, fields.

- **H2020 EHRI** "European Holocaust Research Infrastructure" (2015-2019, PI Conny Kristel [NIOD-KNAW, NL]; LR is task leader) Topic: transform archival research on the Holocaust, by providing methods and tools to integrate and provide access to a wide variety of archival content.
- **H2020 Iperion CH** (2015-2019, PI Luca Pezzati [CNR, IT], LR is task leader) Topic: coordinating infrastructural activities in the cultural heritage domain.
- **H2020 HIRMEOS**: HIRMEOS objective is to improve five important publishing platforms for the open access monographs in the humanities and enhance their technical capacities and services and rendering technologies, while making their content interoperable. Inria is responsable for improving integrating the entity-fishing component deplyed as an infrastructural service for the five platforms.
- **H2020 DESIR**: The DESIR project aims at contributing to the sustainability of the DARIAH infrastructure along all its dimensions: dissemination, growth, technology, robustness, trust and education. Inria is responsable for providing of a portfolio of text analytics services based on GROBID and entity-fishing.

8.2.2. Collaborations in European Programs, Except FP7 & H2020

- **ERIC DARIAH "Digital Research Infrastructure for the Arts and Humanities"** (set up as a consortium of states, 2014-2034; LR served president of the board of director until August 2018) Topic: coordinating Digital Humanities infrastructure activities in Europe (17 partners, 5 associated partners).
- **COST enCollect** (2017-2020, PI Lionel Nicolas [European Academy of Bozen/Bolzano]) Topic: combining language learning and crowdsourcing for developing language teaching materials and more generic language resources for NLP

8.2.3. Collaborations with Major European Organizations

Collaborations with institutions not cited above (for the SPMRL initiative, see below):

- Universität Zürich, Switzerland (Géraldine Walther) [computational morphology, lexicons]
- Berlin-Brandenburgische Akademie der Wissenschaften [Berlin-Brandenburg Academy of Sciences and Humanities], Berlin, Germany (Alexander Geyken) [lexicology]
- Österreichische Akademie der Wissenschaften [Austrian Academy of Sciences], Vienna, Austria (Karlheinz Moerth) [lexicology]
- University of Cambridge, United Kingdom (Ekaterina Kochmar) [text simplification]
- Univerza v Ljubljani [University of Ljubljana], Ljubljana, Slovenia (Darja Fišer) [wordnet development]

8.3. International Initiatives

8.3.1. Participation in International Programs

PHC Maimonide (2018-2019, PI Djamé Seddah, co-PI Yoav Goldberg (Bar Ilan University)). Topics: Building NLP resources for analyzing reactions to major events in Hebrew and French social media.

8.4. International Research Visitors

8.4.1. Visits of International Scientists

- Dr. Ekaterina Kochmar (University of Cambridge), 3 days in June
- Dr. Teresa Lynn (Dublin City University), 2 stays of 1 week each.

9. Dissemination

9.1. Promoting Scientific Activities

- LR was invited to present an overview of information extraction methods in the humanities in the context of the conference cycle: Ringvorlesung "Open Technology for an Open Society", Jan 2018, Berlin, Germany

9.1.1. Scientific Events Organisation

9.1.1.1. General Chair, Scientific Chair

- LR: Co-chair of the Lexical Data Masterclass, Berlin, 3-7 December https://digilex.hypotheses.org/551
- Mohamed Khemakhem: Chair of the GROBID-Camp: Inria de Paris 27th March 2018

9.1.2. Scientific Events Selection

9.1.2.1. Member of the Conference Program/Scientific/Reviewing Committees

- BS: Member of the Program, Scientific or Reviewing Committee of the following conferences and workshops: ACL 2018, NAACL 2018, International Morphology Meeting 2018, Int'l Colloquium on Loanwords and Substrata 2018
- LR: Member of the Program, Scientific or Reviewing Committee of the following conferences and workshops: Fourteenth Joint ACL - ISO Workshop on Interoperable Semantic Annotation, COLING 2018, TPDL 2018, ACL 2018, NAACL-HLT 2018, TOTh 2018, ELPUB 2018, DHd2018, LDL-2018, DH 2018
- DS: Member of the Program, Scientific or Reviewing Committee of the following conferences and workshops: ACL 2018, EMNLP 2018,CoNLL 2018, COLING 2018, EthicNLP 2018, LREC 2018, WNUT 2018, LAW-MWE-CxG 2018.
- EVdLC: Program Committee member and reviewer for LREC, ACL, COLING, NAACL, ToTH, EMNLP

9.1.3. Journal

9.1.3.1. Member of the Editorial Boards

- LR: Member of the JTEI advisory board
- LR: Member of the scientific board of the Revue Humanités numériques

9.1.3.2. Reviewer - Reviewing Activities

- BS: Reviewer for the following journals: *Language Resource and Evaluation*, *Traitement Automatique des Langues*
- LR: Reviewer for the following journals: Language Resource and Evaluation Journal, Journal of the TEI
- DS: Reviewer for the following journals: TALLIP, LRE, NLE, Poznan Studies in Contemporary Linguistics, Computational Linguistics

9.1.4. Invited Talks

- BS was invited to give a talk to Master 2 computational linguistics students and University staff at the Université Grenoble Alpes (November)
- LR was invited to give talks at Open-Access-Tage, Sep 2018, Graz; Workshop DARIAH-CH, University of Neuchâtel, November 2018; "Stay tuned to the future", an international conference on the impact of research infrastructures for social sciences and humanities – bologna, January 2018; NIMS, Tskuba, Japan, September 2018; Rétro-numérisation de documents historiques et partage dans le Web sémantique : l'exemple de la lexicographie – Atelier de formation annuel du consortium Cahier – Montpellier – 26-29 juin 2018; "Serving Learning and Scholarship", Fiesole retreat, Barcelona, April 2018

- DS was invited to give a talk at the Indiana University's department of linguistics (October), at Bar Ilan University (November) respectively on Noisy User-Generated Content Treebanking and on Tackling language variability via diachronic word embeddings.

9.1.5. Training

- Mohamed Khemakhem chaired and tutored the GROBID-Dictionaries series:
  - BBAW & Praxiling joint workshop - Berlin: February 2018
  - Atelier de formation annuel du consortium Cahier – Montpellier – 26-29 June 2018
  - SADiLaR GROBID-Dictionaries Workshop (Pretoria) : October 26, 2018
  - SADiLaR GROBID-Dictionaries Workshop (Potchefstroom) : October 30, 2018 from
  - SADiLaR GROBID-Dictionaries Workshop (StellenBosch) : November 2, 2018
  - Lexical Data Masterclass 2018 - Berlin 3-7 December 2018
- Mathilde Reignault attended the ESSLLI 2018 Summer School in Language and Information as part of her doctoral studies training.

9.1.6. Leadership within the Scientific Community

- LR: President of the board of directors of DARIAH (until August 2018)
- LR: Member of the board of directors or the TEI consortium
- LR: President of ISO committee TC 37 (Language and terminology)
- LR: Member of the ELEXIS Interoperability and Sustainability Committee (ISC) — ELEXIS is the European Lexicographic Infrastructure (https://elex.is)
- EVdLC: Chairman of the ACL special interest group SIGPARSE
- BS: Member, Deputy Treasurer and Member of the Board of the Société de Linguistique de Paris
- DS: Board member of the French NLP society (Atala, 2017-2020), Vice-President of the Atala and program chair of the "journée d'études".
- DS: Member of the ACL's BIG (Broad Interest Group) Diversity group.
- : Charles Riondet: Co-chair of the DARIAH Guidelines and Standards Working Group.
- : Marie Puren: Co-chair of the DARIAH Guidelines and Standards Working Group.

9.1.7. Scientific Expertise

- BS: member of the recruitment committee for the new "ingénieur d'études" position in Inria Paris's communication department
- LR: has carried out various project assesment expertises for: City University Honk Kong, the go!digital programm at the Austrian Academy of Sciences, the Haifa-Technion Joint Research Submission to Milgrom Foundation, teh Swiss National Science Foundation
- DS: Project evaluation for the Flanders Research Agency.
- EVdLC: Evaluator for a European COST proprosal
- EVdLC: Evaluator for the Program Call of DGLFLF on "Langue et Numérique"

9.1.8. Research Administration

- BS: Member of the Board of Inria Paris's Scientific Committee ("Comité des Projets")
- BS: Member of the International Relations Working Group of Inria's Scientific and Technological Orientation Council (COST-GTRI)
- BS: Deputy Head of the research strand on Language Resources of the LabEx EFL (Empirical Foundations of Linguistics), and is therefore a deputy member of the Governing Board of the LabEx; BS and DS are in charge of several research operations in the LabEx
- LR: President of the board of directors of DARIAH

- LR: President of the scientific committee of ABES (Agence Bibliographique de l'Enseignement Supérieur)
- LR: President of ISO committee TC 37 (Language and Terminology)
- Mohamed Khemakhem and LR: Project leaders of the ISO 24613-4 LMF "TEI Serialisation"
- LR: Convener of ISO working group TC 37/SC 4/WG 4 (lexical resources)
- LR: Member of the Text Encoding Initiative board
- LR: advisor for scientific information to the director for science at Inria

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

- Master: Benoît Sagot (with Emmanuel Dupoux), "Speech and Language Processing", 20h, M2, Master "Mathématiques, Vision, Apprentissage", ENS Paris-Saclay, France
- Licence: Djamé Seddah, "Certificat Informatique et Internet", 30h, L1-L2-L3, Université Paris Sorbonne, France
- Master: Djamé Seddah, "Modèles pour la linguistique computationnelle",36h, M1, Université Paris Sorbonne, France
- Master: Djamé Seddah, "Traduction automatique", 30h, M2, Université Paris Sorbonne, France
- Master: Loïc Grobol, "Introduction à la fouille de textes", 39h, M1, Université Sorbonne Nouvelle, France
- Master: Yoann Dupont and Loïc Grobol, "Langages de script", 39h, M2, INALCO, France

9.2.2. Supervision

HdR: Benoît Sagot, "Informatiser le lexique — Modélisation, développement et exploitation de lexiques morphologiques, syntaxiques et sémantiques", 28th June 2018, mentored by Laurent Romary

PhD in progress: Mohamed Khemakhem, "Structuration automatique de dictionnaires à partir de modèles lexicaux standardisés", September 2016, Paris Diderot, supervised by Laurent Romary

PhD in progress: Loïc Grobol, ""Reconnaissance automatique de chaînes de coréférences en français par combinaison d'apprentissage automatique et de connaissances linguistiques", "Université Sorbonne Nouvelle", started in Oct. 2016, supervised by Frédéric Landragin (main superviser), Isabelle Tellier† (main superviser), Éric de La Clergerie and Marco Dinarelli

PhD in progress: Jack Bowers, "Technology, description and theory in language documentation: creating a comprehensive body of multi-media resources for Mixtepec-Mixtec using standards, ontology and Cognitive Linguistics", started in Oct. 2016, EPHE, supervised by Laurent Romary

PhD in progress: Axel Herold, "Automatic identification and modeling of etymological information from retro-digitized dictionaries", October 2016, EPHE, Laurent Romary

PhD in progress: Mathilde Regnault,"Annotation et analyse de corpus hétérogènes", "Université Sorbonne Nouvelle", started in Oct. 2017, supervised by Sophie Prévost (main superviser), Isabelle Tellier†, and Éric de la Clergerie

PhD in progress: Pedro Ortiz, "Automatic Enrichment of Ancient Dictionaries", October 2018, Sorbonne Université, supervised by Laurent Romary and Benoît Sagot

PhD in progress: Benjamin Muller, "Multi-task learning for text normalisation, parsing and machine translation", October 2018, Sorbonne Université, supervised by Benoît Sagot and Djamé Seddah

PhD in progress: José Carlos Rosales, supervised by Guillaume Wisnewski (Limsi) and Djamé Seddah

9.2.3. Juries

- BS: president of the Habilitation committee for Kim Gerdes at Université Paris Nanterre on November 29th (Title: “Same Same but Different: Paradigms in Syntax”; Mentor: Sylvain Kahane)
- BS: reviewer (“rapporteur”) in the PhD committee for Sébastien Delecraz at Aix-Marseille Université on December 10th (Title: “Approches jointes texte/image pour la compréhension multimodale de documents”; Supervisor:)
- LR: member of the PhD committee for Cyrille Suire, University of La Rochelle, September 2019 (Title: "Recherche d’information et humanités numériques : une approche et des outils pour l’historien")
- BS: member of the recruiting committee for a communication officer at Inria Paris (Aug–Oct 2018)
- LR: member of the selection committee for the assistant professor position on linguistics and NLP at University of Orléans (May 2018)

9.3. Popularization

9.3.1. Interventions

- Welcoming of schoolchildren at Inria Paris (half a day with ALMAnaCH members within an one-week-long stay; December 2018)
- ALMAnaCH members were involved in the Profiterole ANR project’s presentation at the *Salon de l’Innovation* of the conference TALN 2018 (the “SiTAL” show).
- Presentation in Education Network ISN "Informatique et Science du Numérique" (March)
- Invited speaker in a citizen debate on Artificial Intelligence (association "Les coteaux en Seine", Bougival, November 21st 2018)

10. Bibliography

Major publications by the team in recent years

[1] D. FIŠER, B. SAGOT. *Constructing a poor man’s wordnet in a resource-rich world*, in "Language Resources and Evaluation", 2015, vol. 49, n° 3, pp. 601-635 [*DOI :* 10.1007/s10579-015-9295-6], https://hal.inria.fr/hal-01174492

[2] P. LOPEZ, L. ROMARY. *HUMB: Automatic Key Term Extraction from Scientific Articles in GROBID*, in "SemEval 2010 Workshop", Uppsala, Sweden, ACL SigLex event, July 2010, 4 p. , https://hal.inria.fr/inria-00493437

[3] C. RIBEYRE, É. VILLEMONTE DE LA CLERGERIE, D. SEDDAH. *Because Syntax does Matter: Improving Predicate-Argument Structures Parsing Using Syntactic Features*, in "Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies", Denver, USA, United States, Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, June 2015, https://hal.archives-ouvertes.fr/hal-01174533

[4] L. ROMARY. *TEI and LMF crosswalks*, in "JLCL - Journal for Language Technology and Computational Linguistics", 2015, vol. 30, n° 1, https://hal.inria.fr/hal-00762664

[5] B. SAGOT. *The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French*, in "7th international conference on Language Resources and Evaluation (LREC 2010)", Valletta, Malta, May 2010, https://hal.inria.fr/inria-00521242

[6] B. SAGOT, É. VILLEMONTE DE LA CLERGERIE. *Error Mining in Parsing Results*, in "Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics", Sydney, Australia, Association for Computational Linguistics, July 2006, pp. 329–336

[7] D. SEDDAH, B. SAGOT, M. CANDITO, V. MOUILLERON, V. COMBET. *The French Social Media Bank: a Treebank of Noisy User Generated Content*, in "COLING 2012 - 24th International Conference on Computational Linguistics", Mumbai, Inde, Kay, Martin and Boitet, Christian, December 2012, http://hal.inria.fr/hal-00780895

[8] R. TSARFATY, D. SEDDAH, Y. GOLDBERG, S. KÜBLER, Y. VERSLEY, M. CANDITO, J. FOSTER, I. REHBEIN, L. TOUNSI. *Statistical Parsing of Morphologically Rich Languages (SPMRL) What, How and Whither*, in "Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages", États-Unis Los Angeles, Association for Computational Linguistics, 2010, pp. 1–12

[9] R. TSARFATY, D. SEDDAH, S. KÜBLER, J. NIVRE. *Parsing Morphologically Rich Languages: Introduction to the Special Issue*, in "Computational Linguistics", March 2013, vol. 39, n^{o} 1, 8 p. [*DOI :* 10.1162/COLI_A_00133], https://hal.inria.fr/hal-00780897

[10] É. VILLEMONTE DE LA CLERGERIE. *Improving a symbolic parser through partially supervised learning*, in "The 13th International Conference on Parsing Technologies (IWPT)", Naria, Japan, November 2013, https://hal.inria.fr/hal-00879358

Publications of the year

Doctoral Dissertations and Habilitation Theses

[11] B. SAGOT. *Computerising the lexicon : Modelling, development and use of morphological, syntactic and semantic lexicons*, Sorbonne Université, June 2018, Habilitation à diriger des recherches, https://hal.inria.fr/tel-01895229

Articles in International Peer-Reviewed Journals

[12] S. BENIAMINE, O. BONAMI, B. SAGOT. *Inferring inflection classes with description length*, in "Journal of Language Modelling", February 2018, vol. 5, n^{o} 3, pp. 465–525, https://hal.inria.fr/hal-01718879

[13] J. BOWERS, L. ROMARY. *Bridging the Gaps between Digital Humanities, Lexicography, and Linguistics: A TEI Dictionary for the Documentation of Mixtepec-Mixtec*, in "Dictionaries: Journal of the Dictionary Society of North America", 2018, vol. 39, n^{o} 2, pp. 79-106, https://hal.inria.fr/hal-01968871

[14] L. ROMARY, C. RIONDET. *EAD-ODD: A solution for project-specific EAD schemes*, in "Archival Science", April 2018, Special thanks to Annelies van Nispen (NIOD) and Hector Martinez Alonso (ALMAnaCH) for their help, and to Lou Burnard (TEI) for his wise comments [*DOI :* 10.1007/s10502-018-9290-y], https://hal.inria.fr/hal-01737568

Invited Conferences

[15] A. BERTINO, L. FOPPIANO, L. ROMARY, P. MOUNIER. *Leveraging Concepts in Open Access Publications*, in "PUBMET 2018 - 5th Conference on Scholarly Publishing in the Context of Open Science", Zadar, Croatia, September 2018, https://hal.inria.fr/hal-01900303

[16] M. KHEMAKHEM, L. ROMARY, S. GABAY, H. BOHBOT, F. FRONTINI, G. LUXARDO. *Automatically Encoding Encyclopedic-like Resources in TEI*, in "The annual TEI Conference and Members Meeting", Tokyo, Japan, September 2018, https://hal.inria.fr/hal-01819505

International Conferences with Proceedings

[17] J. BOWERS, L. ROMARY. *Encoding Mixtepec-Mixtec Etymology in TEI*, in "TEI Conference and Members' Meeting", Tokyo, Japan, September 2018, https://hal.inria.fr/hal-02003975

[18] L. FOPPIANO, L. ROMARY. *entity-fishing: a DARIAH entity recognition and disambiguation service*, in "Digital Scholarship in the Humanities", Tokyo, Japan, September 2018, https://hal.inria.fr/hal-01812100

[19] L. GROBOL, F. LANDRAGIN, S. HEIDEN. *XML-TEI-URS: using a TEI format for annotated linguistic resources*, in "CLARIN Annual Conference 2018", Pisa, Italy, October 2018, https://hal.archives-ouvertes.fr/hal-01827563

[20] L. GROBOL, I. TELLIER, É. VILLEMONTE DE LA CLERGERIE, M. DINARELLI, F. LANDRAGIN. *ANCOR-AS: Enriching the ANCOR Corpus with Syntactic Annotations*, in "LREC 2018 - 11th edition of the Language Resources and Evaluation Conference", Miyazaki, Japan, May 2018, https://hal.inria.fr/hal-01744572

[21] G. JAWAHAR, B. MULLER, A. FETHI, L. MARTIN, É. VILLEMONTE DE LA CLERGERIE, B. SAGOT, D. SEDDAH. *ELMoLex: Connecting ELMo and Lexicon features for Dependency Parsing*, in "CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies", Brussels, Belgium, October 2018 [*DOI :* 10.18653/v1/K18-2023], https://hal.inria.fr/hal-01959045

[22] M. KHEMAKHEM, C. BRANDO, L. ROMARY, F. MÉLANIE-BECQUET, J.-L. PINOL. *Fueling Time Machine: Information Extraction from Retro-Digitised Address Directories*, in "JADH2018 "Leveraging Open Data"", Tokyo, Japan, September 2018, https://hal.archives-ouvertes.fr/hal-01814189

[23] M. KHEMAKHEM, A. HEROLD, L. ROMARY. *Enhancing Usability for Automatically Structuring Digitised Dictionaries*, in "GLOBALEX workshop at LREC 2018", Miyazaki, Japan, May 2018, https://hal.archives-ouvertes.fr/hal-01708137

[24] L. MARTIN, S. HUMEAU, P.-E. MAZARÉ, A. BORDES, É. VILLEMONTE DE LA CLERGERIE, B. SAGOT. *Reference-less Quality Estimation of Text Simplification Systems*, in "1st Workshop on Automatic Text Adaptation (ATA)", Tilburg, Netherlands, November 2018, https://arxiv.org/abs/1901.10746 , https://hal.inria.fr/hal-01959054

[25] A. MORE, Ö. ÇETINOĞLU, Ç. ÇÖLTEKIN, N. HABASH, B. SAGOT, D. SEDDAH, D. TAJI, R. TSARFATY. *CoNLL-UL: Universal Morphological Lattices for Universal Dependency Parsing*, in "11th Language Resources and Evaluation Conference", Miyazaki, Japan, May 2018, https://hal.inria.fr/hal-01786125

[26] B. SAGOT. *A multilingual collection of CoNLL-U-compatible morphological lexicons*, in "Eleventh International Conference on Language Resources and Evaluation (LREC 2018)", Miyazaki, Japan, May 2018, https://hal.inria.fr/hal-01798798

[27] D. SEDDAH, É. VILLEMONTE DE LA CLERGERIE, B. SAGOT, H. MARTINEZ ALONSO, M. CANDITO. *Cheating a Parser to Death: Data-driven Cross-Treebank Annotation Transfer*, in "Eleventh International Conference on Language Resources and Evaluation (LREC 2018)", Miyazaki, Japan, May 2018, https://hal.inria.fr/hal-01798801

National Conferences with Proceedings

[28] M. BRASSIER, A. PURET, A. VOISIN-MARRAS, L. GROBOL. *Classification par paires de mention pour la résolution des coréférences en français parlé interactif*, in "Conférence jointe CORIA-TALN-RJC 2018", Rennes, France, ATALA and ARIA, May 2018, https://hal.inria.fr/hal-01821213

Conferences without Proceedings

[29] H. BOHBOT, F. FRONTINI, G. LUXARDO, M. KHEMAKHEM, L. ROMARY. *Presenting the Nénufar Project: a Diachronic Digital Edition of the Petit Larousse Illustré*, in "GLOBALEX 2018 - Globalex workshop at LREC2018", Miyazaki, Japan, May 2018, pp. 1-6, https://hal.archives-ouvertes.fr/hal-01728328

[30] M. DINARELLI, L. GROBOL. *Modeling a label global context for sequence tagging in recurrent neural networks*, in "Journée commune AFIA-ATALA sur le Traitement Automatique des Langues et l'Intelligence Artificielle pendant la onzième édition de la plate-forme Intelligence Artificielle (PFIA 2018)", Nancy, France, July 2018, https://hal.archives-ouvertes.fr/hal-02002111

[31] R. GARNIER, B. SAGOT. *New results on a centum substratum in Greek: the Lydian connection*, in "International Colloquium on Loanwords and Substrata in Indo-European languages", Limoges, France, June 2018, https://hal.inria.fr/hal-01798979

[32] D. LINDEMANN, M. KHEMAKHEM, L. ROMARY. *Retro-digitizing and Automatically Structuring a Large Bibliography Collection*, in "European Association for Digital Humanities (EADH) Conference", Galway, Ireland, EADH, December 2018, https://hal.archives-ouvertes.fr/hal-01941534

[33] H. MARAOUI, K. HADDAR, L. ROMARY. *Segmentation tool for hadith corpus to generate TEI encoding*, in "4th International Conference on Advanced Intelligent Systems and Informatics (AISI'18)", Cairo, Egypt, September 2018, https://hal.archives-ouvertes.fr/hal-01794105

[34] M. PUREN, A. CHAGUÉ, M. MARTINI, É. VILLEMONTE DE LA CLERGERIE, C. RIONDET. *Creating gold data to understand the gender gap in the French textile trades (17th–20th century). Time-Us project*, in "Digital Humanities 2018: "Puentes/ Bridges"", Mexico, Mexico, June 2018, https://hal.archives-ouvertes.fr/hal-01850080

[35] M. PUREN, C. RIONDET, L. ROMARY, D. SEILLIER, L. TADJOU. *The Standardization Survival Kit (SSK): Bringing best practices to research communities in the Humanities*, in "Digital Humanities Benelux 2018", Amsterdam, Netherlands, June 2018, https://hal.archives-ouvertes.fr/hal-01850075

[36] M. PUREN, D. SEILLIER, C. RIONDET, L. TADJOU. *Le Standardization Survival Kit (SSK): Faciliter l'usage des standards dans les Humanités*, in "Rencontres de la TGIR Huma-Num", Ecully, France, June 2018, https://hal.archives-ouvertes.fr/hal-01850078

[37] B. SAGOT. *A new PIE root *h1er '(to be) dark red, dusk red': drawing the line between inherited and borrowed words for 'red(ish)', 'pea', 'ore', 'dusk' and 'love' in daughter languages*, in "International Colloquium on Loanwords and Substrata in Indo-European languages", Limoges, France, June 2018, https://hal.inria.fr/hal-01798976

Scientific Books (or Scientific Book chapters)

[38] T. BLANKE, C. KRISTEL, L. ROMARY. *Crowds for Clouds: Recent Trends in Humanities Research Infrastructures*, in "Cultural Heritage Digital Tools and Infrastructures", A. BENARDOU, E. CHAMPIO, C. DALLAS, L. HUGHES (editors), Routledge, 2018, https://arxiv.org/abs/1601.00533 , https://hal.inria.fr/hal-01248562

Research Reports

[39] M.-L. BONHOMME. *Répertoire des Notaires parisiens Segmentation automatique et reconnaissance d'écriture : Rapport exploratoire*, Inria, October 2018, pp. 1-10, https://hal.inria.fr/hal-01949198

[40] C. RIONDET, D. SEILLIER, L. TADJOU, L. ROMARY. *Standardization Survival Kit (Final)*, Inria Paris, October 2018, https://hal.inria.fr/hal-01925144

Scientific Popularization

[41] M. PUREN, C. RIONDET, D. SEILLIER, L. TADJOU. *The Standardization Survival Kit : For a wider use of standards within Arts and Humanities*, April 2018, Journée de formation : "Gérer et explorer les données textuelles", https://hal.inria.fr/hal-01763688

[42] C. RIONDET. *TEI: de l'image au texte : Décrire son corpus grâce aux métadonnées*, February 2018, pp. 1-69, TEI: de l'image au texte, https://hal.inria.fr/hal-01708839

Other Publications

[43] A. ADLI, E. ENGEL, L. ROMARY, F. SAME. *A stand-off XML-TEI representation of reference annotation*, March 2018, DGfS 2018: 40. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft, Poster, https://hal.inria.fr/hal-01876327

[44] A. BERTINO, L. FOPPIANO, L. ROMARY, P. MOUNIER. *Leveraging Concepts in Open Access Publications*, January 2019, working paper or preprint, https://hal.inria.fr/hal-01981922

[45] H. BOHBOT, A. FAUCHERE, F. FRONTINI, A. JACKIEWICZ, G. LUXARDO, A. STEUCKARDT, M. KHEMAKHEM, L. ROMARY. *A Diachronic Digital Edition of the Petit Larousse illustré*, May 2018, Journée d'étude CORLI : Traitements et standardisation des corpus multimodaux et web 2.0., Poster, https://hal.archives-ouvertes.fr/hal-01873805

[46] J. BOWERS. *Language Documentation and Standards in Digital Humanities: TEI and the documentation of Mixtepec-Mixtec*, February 2019, working paper or preprint, https://hal.inria.fr/hal-02004005

[47] S. GABAY, M. KHEMAKHEM, L. ROMARY. *GROBID and catalogues*, November 2018, Lecture, https://hal.archives-ouvertes.fr/cel-01951107

[48] K. ILLMAYER, M. PUREN. *How to work together successfully with e-Humanities and e-Heritage Research Infrastructures (PARTHENOS Webinar)*, February 2018, Lecture, https://hal.archives-ouvertes.fr/cel-01731455

[49] F. LANDRAGIN, M. DELABORDE, Y. DUPONT, L. GROBOL. *Description et modélisation des chaînes de référence. Le projet ANR Democrat (2016-2020) et ses avancées à mi-parcours*, May 2018, Cinquième édition du Salon de l'Innovation en TAL (Traitement Automatique des Langues) et RI (Recherche d'Informations), Poster, https://hal.archives-ouvertes.fr/hal-01797982

[50] M. PUREN, C. RIONDET, L. ROMARY, D. SEILLIER, L. TADJOU. *SSK by example. Make your Arts and Humanities research go standard*, June 2018, Digital Humanities 2018 : "Bridges/Puentes", Poster, https://hal.archives-ouvertes.fr/hal-01848882

[51] M. PUREN, C. RIONDET, L. ROMARY, D. SEILLIER, L. TADJOU. *The SSK. Make your Arts and Humanities research go standard. TEI inside !*, September 2018, TEI2018 - Annual TEI Conference and Members Meeting, Poster, https://hal.inria.fr/hal-01902702

[52] D. REINEKE, L. ROMARY. *Reference SKOS and TBX vocabularies used for comparing the two standards*, September 2018, technical document associated to a journal submission, https://hal.inria.fr/hal-01883377

[53] C. RIONDET. *Stewardship of Cultural Heritage Data. In the shoes of a researcher*, April 2018, Cultural Heritage Data Re-use Charter Feedback workshop hosted by the LIBER Digital Humanities & Digital Cultural Heritage Working group, https://hal.inria.fr/hal-01762295

[54] C. RIONDET. *Traces de l'héroïsme. Le programme mémoriel de la résistance parisienne*, February 2018, À paraître en 2018 dans "La clandestinité politique, des anarchistes au djihadisme (XXe-XXIe siècle)" (sous la dir. de Cirefice, France, Le Quang, Riondet), https://hal.inria.fr/hal-01715006

[55] C. RIONDET. *À la recherche de l'archive clandestine*, February 2018, À paraître en 2018 dans "La clandestinité politique, des anarchistes au djihadisme (XXe-XXIe siècle)" (sous la dir. de Cirefice, France, Le Quang, Riondet), https://hal.inria.fr/hal-01715002

[56] L. ROMARY. *Data Mining Technologies at the service of Open Knowledge*, January 2018, pp. 1-65, Ringvorlesung "Open Technology for an Open Society", https://hal.inria.fr/hal-01708771

[57] L. ROMARY. *Open Access in France: how the call of Jussieu reflects our social, technical and political landscape*, September 2018, Open-Access-Tage, https://hal.inria.fr/hal-01881469

References in notes

[58] M. J. ARANZABE, A. D. DE ILARRAZA, I. GONZALEZ-DIOS. *Transforming complex sentences using dependency trees for automatic text simplification in Basque*, in "Procesamiento del lenguaje natural", 2013, vol. 50, pp. 61–68

[59] P. BANSKI, B. GAIFFE, P. LOPEZ, S. MEONI, L. ROMARY, T. SCHMIDT, P. STADLER, A. WITT. *Wake up, standOff!*, September 2016, TEI Conference 2016, https://hal.inria.fr/hal-01374102

[60] O. BONAMI, B. SAGOT. *Computational methods for descriptive and theoretical morphology: a brief introduction*, in "Morphology", 2017, vol. 27, n^{o} 4, pp. 1-7 [*DOI :* 10.1017/CBO9781139248860], https://hal.inria.fr/hal-01628253

[61] A. BOUCHARD-CÔTÉ, D. HALL, T. GRIFFITHS, D. KLEIN. *Automated Reconstruction of Ancient Languages using Probabilistic Models of Sound Change*, in "Proceedings of the National Academy of Sciences", 2013, n^{o} 110, pp. 4224–4229

[62] J. C. K. CHEUNG, G. PENN. *Utilizing Extra-sentential Context for Parsing*, in "Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing", Cambridge, Massachusetts, EMNLP '10, 2010, pp. 23–33

[63] M. CONSTANT, M. CANDITO, D. SEDDAH. *The LIGM-Alpage Architecture for the SPMRL 2013 Shared Task: Multiword Expression Analysis and Dependency Parsing*, in "Fourth Workshop on Statistical Parsing of Morphologically Rich Languages", Seattle, United States, October 2013, pp. 46-52, https://hal.archives-ouvertes.fr/hal-00932372

[64] T. DOZAT, C. D. MANNING. *Deep Biaffine Attention for Neural Dependency Parsing*, in "CoRR", 2016, vol. abs/1611.01734, http://arxiv.org/abs/1611.01734

[65] Y. FANG, M. CHANG. *Entity Linking on Microblogs with Spatial and Temporal Signals*, in "TACL", 2014, vol. 2, pp. 259–272, https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/323

[66] L. GROBOL, F. LANDRAGIN, S. HEIDEN. *Interoperable annotation of (co)references in the Democrat project*, in "Thirteenth Joint ISO-ACL Workshop on Interoperable Semantic Annotation", Montpellier, France, H. BUNT (editor), ACL Special Interest Group on Computational Semantics (SIGSEM) and ISO TC 37/SC 4 (Language Resources) WG 2, September 2017, https://hal.archives-ouvertes.fr/hal-01583527

[67] J. E. HOARD, R. WOJCIK, K. HOLZHAUSER. *An automated grammar and style checker for writers of Simplified English*, in "Computers and Writing: State of the Art", 1992, pp. 278–296

[68] D. HOVY, T. FORNACIARI. *Increasing In-Class Similarity by Retrofitting Embeddings with Demographic Information*, in "Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing", Association for Computational Linguistics, 2018, pp. 671–677, http://aclweb.org/anthology/D18-1070

[69] D. HRUSCHKA, S. BRANFORD, E. SMITH, J. WILKINS, A. MEADE, M. PAGEL, T. BHATTACHARYA. *Detecting Regular Sound Changes in Linguistics as Events of Concerted Evolution*, in "Current Biology", 2015, vol. 1, n^{o} 25, pp. 1–9

[70] M. KHEMAKHEM, L. FOPPIANO, L. ROMARY. *Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields*, in "electronic lexicography, eLex 2017", Leiden, Netherlands, September 2017, https://hal.archives-ouvertes.fr/hal-01508868

[71] M. KHEMAKHEM, L. FOPPIANO, L. ROMARY. *Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields*, in "electronic lexicography, eLex 2017", Leiden, Netherlands, September 2017, https://hal.archives-ouvertes.fr/hal-01508868

[72] S. KÜBLER, M. SCHEUTZ, E. BAUCOM, R. ISRAEL. *Adding Context Information to Part Of Speech Tagging for Dialogues*, in "NEALT Proceedings Series", M. DICKINSON, K. MUURISEP, M. PASSAROTTI (editors), 2010, vol. 9, pp. 115-126

[73] A.-L. LIGOZAT, C. GROUIN, A. GARCIA-FERNANDEZ, D. BERNHARD. *Approches à base de fréquences pour la simplification lexicale*, in "TALN-RÉCITAL 2013", 2013, 493 p.

[74] P. LOPEZ. *entity-fishing*, in "WikiDATA Conf", September 2017, https://grobid.s3.amazonaws.com/presentations/29-10-2017.pdf

[75] P. LOPEZ, A. MEYER, L. ROMARY. *CENDARI Virtual Research Environment & Named Entity Recognition techniques*, February 2014, Grenzen überschreiten – Digitale Geisteswissenschaft heute und morgen, Poster, https://hal.inria.fr/hal-01577975

[76] H. MARTINEZ ALONSO, D. SEDDAH, B. SAGOT. *From Noisy Questions to Minecraft Texts: Annotation Challenges in Extreme Syntax Scenarios*, in "2nd Workshop on Noisy User-generated Text (W-NUT) at CoLing 2016", Osaka, Japan, December 2016, https://hal.inria.fr/hal-01584054

[77] M. E. PETERS, M. NEUMANN, M. IYYER, M. GARDNER, C. CLARK, K. LEE, L. ZETTLEMOYER. *Deep contextualized word representations*, in "Proc. of NAACL", 2018

[78] J. PYSSALO. *System PIE: the Primary Phoneme Inventory and Sound Law System for Proto-Indo-European*, University of Helsinki, 2013

[79] L. RELLO, R. BAEZA-YATES, S. BOTT, H. SAGGION. *Simplify or help?: text simplification strategies for people with dyslexia*, in "Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility", ACM, 2013, 15 p.

[80] L. RELLO, R. BAEZA-YATES, L. DEMPERE-MARCO, H. SAGGION. *Frequent words improve readability and short words improve understandability for people with dyslexia*, in "IFIP Conference on Human-Computer Interaction", Springer, 2013, pp. 203–219

[81] C. RIBEYRE, M. CANDITO, D. SEDDAH. *Semi-Automatic Deep Syntactic Annotations of the French Treebank*, in "The 13th International Workshop on Treebanks and Linguistic Theories (TLT13)", Tübingen, Germany, Proceedings of TLT 13, Tübingen Universität, December 2014, https://hal.inria.fr/hal-01089198

[82] A. M. RUSH, R. REICHART, M. COLLINS, A. GLOBERSON. *Improved Parsing and POS Tagging Using Inter-sentence Consistency Constraints*, in "Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning", Jeju Island, Korea, EMNLP-CoNLL '12, 2012, pp. 1434–1444

[83] B. SAGOT, H. MARTINEZ ALONSO. *Improving neural tagging with lexical information*, in "15th International Conference on Parsing Technologies", Pisa, Italy, September 2017, pp. 25-31, https://hal.inria.fr/hal-01592055

[84] B. SAGOT, D. NOUVEL, V. MOUILLERON, M. BARANES. *Extension dynamique de lexiques morphologiques pour le français à partir d'un flux textuel*, in "TALN - Traitement Automatique du Langage Naturel", Les sables d'Olonne, France, June 2013, pp. 407-420, https://hal.inria.fr/hal-00832078

[85] B. SAGOT. *DeLex, a freely-avaible, large-scale and linguistically grounded morphological lexicon for German*, in "Language Resources and Evaluation Conference", Reykjavik, Iceland, European Language Resources Association, May 2014, https://hal.inria.fr/hal-01022288

[86] B. SAGOT. *External Lexical Information for Multilingual Part-of-Speech Tagging*, Inria Paris, June 2016, n° RR-8924, https://hal.inria.fr/hal-01330301

[87] B. SAGOT. *Extracting an Etymological Database from Wiktionary*, in "Electronic Lexicography in the 21st century (eLex 2017)", Leiden, Netherlands, September 2017, pp. 716-728, https://hal.inria.fr/hal-01592061

[88] C. SCARTON, M. DE OLIVEIRA, A. CANDIDO JR, C. GASPERIN, S. M. ALUÍSIO. *SIMPLIFICA: a tool for authoring simplified texts in Brazilian Portuguese guided by readability assessments*, in "Proceedings of the NAACL HLT 2010 Demonstration Session", Association for Computational Linguistics, 2010, pp. 41–44

[89] Y. SCHERRER, B. SAGOT. *A language-independent and fully unsupervised approach to lexicon induction and part-of-speech tagging for closely related languages*, in "Language Resources and Evaluation Conference", Reykjavik, Iceland, European Language Resources Association, May 2014, https://hal.inria.fr/hal-01022298

[90] S. SCHUSTER, É. VILLEMONTE DE LA CLERGERIE, M. D. CANDITO, B. SAGOT, C. D. MANNING, D. SEDDAH. *Paris and Stanford at EPE 2017: Downstream Evaluation of Graph-based Dependency Representations*, in " EPE 2017 - The First Shared Task on Extrinsic Parser Evaluation", Pisa, Italy, Proceedings of the 2017 Shared Task on Extrinsic Parser Evaluation, September 2017, pp. 47-59, https://hal.inria.fr/hal-01592051

[91] D. SEDDAH, M. CANDITO. *Hard Time Parsing Questions: Building a QuestionBank for French*, in "Tenth International Conference on Language Resources and Evaluation (LREC 2016)", Portorož, Slovenia, Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC 2016), May 2016, https://hal.archives-ouvertes.fr/hal-01457184

[92] D. SEDDAH, B. SAGOT, M. CANDITO, V. MOUILLERON, V. COMBET. *The French Social Media Bank: a Treebank of Noisy User Generated Content*, in "COLING 2012 - 24th International Conference on Computational Linguistics", Mumbai, India, Kay, Martin and Boitet, Christian, December 2012, https://hal.inria.fr/hal-00780895

[93] D. SEDDAH, B. SAGOT, M. CANDITO. *The Alpage Architecture at the SANCL 2012 Shared Task: Robust Pre-Processing and Lexical Bridging for User-Generated Content Parsing*, in "SANCL 2012 - First Workshop on Syntactic Analysis of Non-Canonical Language , an NAACL-HLT'12 workshop", Montréal, Canada, June 2012, https://hal.inria.fr/hal-00703124

[94] M. SHARDLOW. *A survey of automated text simplification*, in "International Journal of Advanced Computer Science and Applications", 2014, vol. 4, n^{o} 1, pp. 58–70

[95] A. SØGAARD, Y. GOLDBERG. *Deep multi-task learning with low level tasks supervised at lower layers*, in "Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)", Berlin, Germany, 2016, pp. 231–235

[96] É. VILLEMONTE DE LA CLERGERIE, B. SAGOT, D. SEDDAH. *The ParisNLP entry at the ConLL UD Shared Task 2017: A Tale of a #ParsingTragedy*, in "Conference on Computational Natural Language Learning", Vancouver, Canada, Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, August 2017, pp. 243-252 [*DOI :* 10.18653/v1/K17-3026], https://hal.inria.fr/hal-01584168

[97] É. VILLEMONTE DE LA CLERGERIE. *Jouer avec des analyseurs syntaxiques*, in "TALN 2014", Marseilles, France, ATALA, July 2014, https://hal.inria.fr/hal-01005477

[98] G. WALTHER, B. SAGOT. *Speeding up corpus development for linguistic research: language documentation and acquisition in Romansh Tuatschin*, in "Joint SIGHUM Workshop on Computational Linguistics for

Cultural Heritage, Social Sciences, Humanities and Literature", Vancouver, Canada, Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, August 2017, pp. 89 - 94 [*DOI :* 10.18653/v1/W17-2212], https://hal.inria.fr/hal-01570614

[99] G. WALTHER, B. SAGOT. *Speeding up corpus development for linguistic research: language documentation and acquisition in Romansh Tuatschin*, in "Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature", Vancouver, Canada, Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, August 2017, pp. 89 - 94 [*DOI :* 10.18653/v1/W17-2212], https://hal.inria.fr/hal-01570614