Activity Report 2019

Project-Team ALMANACH

Automatic Language Modelling and Analysis & Computational Humanities

RESEARCH CENTER
Paris

THEME
Language, Speech and Audio

Table of contents

Project-Team ALMANACH

Creation of the Project-Team: 2019 July 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

Benoît Sagot [Team leader, Inria, Researcher, HDR]
Pierre Boullier [Inria, Emeritus, until Apr 2019]
Laurent Romary [Inria, Senior Researcher, HDR]
Djamé Seddah [Inria, Researcher, “détachement” at Inria from Sorbonne Univ.]
Éric Villemonte de La Clergerie [Inria, Researcher]

Faculty Member

Kim Gerdes [Univ Sorbonne Nouvelle, Associate Professor, from Mar 2019 (“délégation”), HDR]

Post-Doctoral Fellows

Yoann Dupont [Univ d’Orléans, Post-Doctoral Fellow, from Jun 2019]
Murielle Fabre [Inria, Post-Doctoral Fellow]

Gaël Guibon [Univ Denis Diderot, Post-Doctoral Fellow, from Jun 2019]
Ilia Markov [Inria, Post-Doctoral Fellow, until May 2019]

PhD Students

Jack Bowers [Austrian Academy of Sciences, PhD Student]
Marine Courtin [Sorbonne Université, PhD Student, from Mar 2019]
Yixuan Li [Sorbonne Université, PhD Student, from Mar 2019]
Clémentine Fourrier [Inria, PhD Student, from Oct 2019]
Loïc Grobol [École Normale Supérieure de Paris, PhD Student]
Mohamed Khemakhem [Inria, PhD Student]
Louis Martin [Facebook, PhD Student]
Benjamin Muller [Inria, PhD Student]
Pedro Ortiz Suárez [Inria, PhD Student]
Mathilde Regnault [École Normale Supérieure de Paris, PhD Student]
Jose Rosales Nuñez [CNRS, PhD Student]

Technical staff

Achraf Azhar [Inria, Engineer, until Apr 2019]
Alix Chagué [Inria, Engineer, from Feb 2019]
Farah Essaidi [Inria, Engineer, from Oct 2019]
Clémentine Fourrier [Inria, Engineer, from Mar 2019 until Sep 2019]
Ganesh Jawahar [Inria, Engineer, until Sep 2019]
Tanti Kristanti [Inria, Engineer]
Alba Marina Malaga Sabogal [Inria, Engineer, until Aug 2019]
Charles Riondet [Inria, Engineer, until May 2019]
Dorian Seillier [Inria, Engineer, until May 2019]
Lionel Tadonfouet [Inria, Engineer]

Interns and Apprentices

Damien Biabiany [Inria, Apprentice]
Matthieu Futeral [Inria, from Jul 2019]
Hafida Le Cloirec [Univ Denis Diderot, Jun 2019]
Victoria Le Fourner [Inria, from Mar 2019 until Jul 2019]
Abhishek Srivastava [Inria, from May 2019 until Jul 2019]

Administrative Assistant

Meriem Guemair [Inria, Administrative Assistant]

2. Overall Objectives

2.1. Overall Objectives

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

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 and 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**.

[1] ALMAnaCH was created as an Inria team ("équipe") on the 1st January, 2017 and as a project-team on the 1st July 2019.

Digital Humanities and social sciences (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** and computational social sciences aim 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 involved computational linguists from Inria as well as computational linguists from Paris-Diderot University with a strong background in linguistics, and proved successful. However, the context has changed since then, 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, as it provides them with new types of data (on which their tools, resources and algorithms can be used, thereby leading to new results in human sciences), as well as with new and challenging research problems, which, if solved, provide new ways of studying human sciences.

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. Finally, we remain dedicated to having an impact on the industrial world and more generally on society, via multiple types of collaboration with companies and other institutions (startup creation, industrial contracts, expertise, etc.).

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 at all time scales. 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 very large amounts of annotated data. They still suffer from the high level of variability found for instance in **user-generated content**, **non-contemporary texts**, as well as in **domain-specific documents** (e.g. financial, legal).

ALMAnaCH tackles the challenge of language variation in two complementary directions, supported by a third, transverse research axis on language resources. These three research axes do not reflect an internal organisation of eparate teams. They are meant to structure our scientific agenda, and most members of the project-team are involved in two or all of them.

ALMAnaCH's research axes, themselves structured in sub-axis, are the following:

1. Automatic Context-augmented Linguistic Analysis
    1. Processing of natural language at all levels: morphology, syntax, semantics
    2. Integrating context in NLP systems
    3. Information and knowledge extraction
2. Computational Modelling of Linguistic Variation
    1. Theoretical and empirical synchronic linguistics
    2. Sociolinguistic variation
    3. Diachronic variation
    4. Accessibility-related variation
3. Modelling and development of Language Resources
    1. Construction, management and automatic annotation of text corpora
    2. Development of lexical resources
    3. Development of annotated corpora

3. Research Program

3.1. Research strands

As described above, ALMAnaCH's scientific programme is organised around three research axes. The first two aim to tackle the challenge of language variation in two complementary directions. They are supported by a third, transverse research axis on language resources. Our four-year objectives are described in much greater detail in the project-team proposal, whose very recent final validation in June 2019 resulted in the upgrade of ALMAnaCH to the "project-team" status in July 2019. They can be summarised as follows:

3.1.1. Research axis 1

Our first objective is to **stay at a state-of-the-art level in key NLP tasks** such as shallow processing, part-of-speech tagging and (syntactic) parsing, which are core expertise domains of ALMAnaCH members. This will also require us to improve the **generation of semantic representations (semantic parsing)**, and to begin to explore tasks such as machine translation, which now relies on neural architectures also used for some of the above-mentioned tasks. Given the generalisation of neural models in NLP, we will also be involved in better understanding how such models work and what they learn, something that is directly related to the investigation of language variation (Research axis 2). We will also work on 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 will have to identify, model and take advantage of each type of contextual information available. Addressing these issues will enable the development of new lines of research related to conversational content. Applications include improved information and knowledge extraction algorithms. We will especially focus on challenging datasets such as domain-specific texts (e.g. financial, legal) as well as historical documents, in the larger context of the development of digital humanities. We currently also explore the even more challenging new direction of a cognitively inspired NLP, in order to tackle the possibility to enrich the architecture of state-of-the-art algorithms, such as RNNGs, based on human neuroimaging-driven data.

3.1.2. Research axis 2

Language variation must be better understood and modelled in all its forms. In this regard, we will put a strong emphasis on **four 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 to language-related disabilities, and **diachronic variation** (computational exploration of language change and language history, with a focus on Old to all forms of Modern French, as well as Indo-European languages in general). In addition, the noise introduced by Optical Character Recognition and Handwritten Text Recognition systems, especially in the context of historical documents, bears some similarities to that of non-canonical input in user-generated content (e.g. erroneous characters). This noise constitutes a more transverse kind of variation stemming from the way language is graphically encoded, which we call **language-encoding variation**. Other types of language variation will also become important research topics for ALMAnaCH in the future. This includes dialectal variation (e.g. work on Arabic varieties, something on which we have already started working, producing the first annotated data set on Maghrebi Arabizi, the Arabic variants used on social media by people from North-African countries, written using a non-fixed Latin-script transcription) as well as the study and exploitation of paraphrases in a broader context than the above-mentioned complexity-based variation.

Both research axes above rely on the availability of language resources (corpora, lexicons), which is the focus of our third, transverse research axis.

3.1.3. Research axis 3

Language resource development (raw and annotated corpora, lexical resources) is not just a necessary preliminary step to create both evaluation datasets for NLP systems and training datasets for NLP systems based on machine learning. When dealing with datasets of interest to researchers from the humanities (e.g. large archives), it is also a goal *per se* and a preliminary step before making such datasets available and exploitable online. It involves a number of scientific challenges, among which (i) tackling issues related to the digitalisation of non-electronic datasets, (ii) tackling issues related to the fact that many DH-related datasets are domain-specific and/or not written in contemporary languages; (iii) the development of semi-automatic and automatic algorithms to speed up the work (e.g. automatic extraction of lexical information, low-resource learning for the development of pre-annotation algorithms, transfer methods to leverage existing tools and/or resources for other languages, etc.) and (iv) the development of formal models to represent linguistic information in the best possible way, thus requiring expertise at least in NLP and in typological and formal linguistics. Such endeavours are domains of expertise of the ALMAnaCH team, and a large part of our research activities will be dedicated to language resource development. In this regard, we aim to retain our leading role in the representation and management of lexical resource and treebank development and also to develop a complete processing line for the transcription, analysis and processing of complex documents of interest to the humanities, in particular archival documents. This research axis 3 will benefit the whole team and beyond, and will benefit from and feed the work of the other research axes.

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. 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. 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 [74], [112] and one of the best for robust multilingual parsing as shown by our results in various shared tasks over the years [108], [105], [111], [82]. 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 [105].

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 [2] and to Extrinsic Parsing Evaluation Shared Task [3].

[2] We ranked 3 for UPOS tagging and 6 for dependency parsing out of 33 participants.

[3] Semantic graph parsing, evaluated on biomedical data, speech and opinion. We ranked 1 in a joint effort with the Stanford NLP team

Fundamentally, we want to build tools that are less sensitive to variation, more easily configurable, and self-adapting. Our short-term goal is to explore techniques such as multi-task learning (cf. already [110]) 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.

3.2.2. Integrating context in NLP systems

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 [85]. In the case of UGC, working across sentence boundaries was explored for instance, with limited success, by [73] for document-wise parsing and by [96] 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).

3.2.3. 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.

3.3. 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 [70]. 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 [80] 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 [90] 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 [71] and [81]. 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 [109]). 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 [91]. 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. [4]

[4] 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).

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 [79], [103], 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 [68]). Lexical simplification, another aspect of text simplification [86], [92], 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 [5] 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.

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.

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

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 [6] 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 Inria predecessor 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 [101], [97], [111] 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 languages (see already [113]), 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 [100], [83], [102], while using and developing (semi)automatic lexical information extraction techniques based on existing corpora [98], [104]. A new line of research will be to integrate

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

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 [83], [102].

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) [94] 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 [107], [93], [106], [88]) 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 [76]. 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. Highlights of the Year

5.1. Highlights of the Year

The main highlight for ALMAnaCH in 2019 is the publication of CamemBERT, a French neural language model trained on the French section of OSCAR, our very large multilingual web-based raw corpus. The publication of this first Transformer-based language model for French, which allowed us to improve the state of the art in several classical NLP tasks, met a large success both in the academic and industrial worlds. It is the topic of an article in the major French daily newspaper Le Monde and of a broadcast on the national radio France Culture.

6. New Software and Platforms

6.1. Enqi

- Author: Benoît Sagot
- Contact: Benoît Sagot

6.2. 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/

6.3. 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: Eric de La Clergerie
- Contact: Éric De La Clergerie
- URL: http://mgkit.gforge.inria.fr/

6.4. 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/

6.5. 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

6.6. FSMB

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

6.7. 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: Eric de La Clergerie
- Contact: Eric de La Clergerie
- URL: http://dyalog.gforge.inria.fr/

6.8. 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 Eric de La Clergerie
- Contact: Benoît Sagot
- URL: http://lingwb.gforge.inria.fr/

6.9. 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: Eric de La Clergerie
- Contact: Eric de La Clergerie
- URL: http://alpage.inria.fr/frmgwiki/

6.10. 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

6.11. 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

6.12. 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/

6.13. 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

6.14. 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

7. New Results

7.1. New results on text simplification

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

Text simplification (TS) aims at making a text easier to read and understand by simplifying grammar and structure while keeping the underlying meaning and information identical. It is therefore an instance of language variation, based on language complexity. It can benefit numerous audiences, such as people with disabilities, language learners or even everyone, for instance when dealing with intrinsically complex texts such as legal documents.

We have initiated in 2017 a collaboration with the Facebook Artificial Intelligence Research (FAIR) lab in Paris and with the UNAPEI, the federation of French associations helping people with mental disabilities and their families. The objective of this collaboration is to develop tools for helping the simplification of texts aimed at mentally disabled people. More precisely, the is to develop a computer-assisted text simplification platform (as opposed to an automatic TS system). In this context, a CIFRE PhD thesis has started in collaboration with the FAIR on the TS task. We have first dedicated important efforts to the problem of the evaluation of 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 judgement of grammaticality and meaning preservation, whereas simplicity is best evaluated by basic length-based metrics [87]. Our implementations of several metrics have been made this year easily accessible and described in a demo paper in collaboration with the University of Sheffield [16].

In 2019, we have also investigated an important issue inherent to the TS task. Although it is often considered an all-purpose generic task where the same simplification is suitable for all, multiple audiences can benefit from simplified text in different ways. We have therefore introduced a discrete parametrisation mechanism that provides explicit control on TS systems based on Seq2Seq neural models. As a result, users can condition the simplifications returned by a model on parameters such as length and lexical complexity. We also show that carefully chosen values of these parameters allow out-of-the-box Seq2Seq neural models to outperform their standard counterparts on simplification benchmarks. Our best parametrised model improves over the previous state of the art performance [61].

Finally, we are involved in the development of a new text simplification corpus. In order to simplify a sentence, human editors perform multiple rewriting transformations: splitting it into several shorter sentences, paraphrasing (i.e. replacing complex words or phrases by simpler synonyms), reordering components, and/or deleting information deemed unnecessary. Despite the vast range of possible text alterations, current models for automatic sentence simplification are evaluated using datasets that are focused on single transformations, such as paraphrasing or splitting. This makes it impossible to understand the ability of simplification models in more abstractive and realistic settings. This is what motivated the development of ASSET, a new dataset for assessing sentence simplification in English, in collaboration with the University of Sheffield (United Kingdom). ASSET is a crowdsourced multi-reference corpus where each simplification was produced by executing several rewriting transformations. Through quantitative and qualitative experiments, we have shown that simplifications in ASSET are better at capturing characteristics of simplicity when compared to other standard evaluation datasets for the task. Furthermore, we have motivated the need for developing better methods for automatic evaluation using ASSET, since we show that current popular metrics may not be suitable for assessment when multiple simplification transformations were performed.

7.2. NLP and computational neurolinguistics

Participants: Éric Villemonte de La Clergerie, Murielle Fabre.

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 brain imagery (fMRI) data resulting from listening of the novel Le Petit Prince. Following the availability of an increasing amount of fMRI datasets in French and English, the project has investigated the correlations between fMRI observations and an increasing number of parser-based features based on several parsers representing a number of architecture types (LSTM, RNN, Dyalog-SR [statistical], FRMG [hybrid symbolic/statistical]) [20].

While pursuing the purely computation goal of developing a method of variable beam size inference for Recurrent Neural Network Grammar (RNNG) the project investigated how different beam search methods can show different goodness of fit with fMRI signal recorded during naturalistic story listening [58]. This approach is part of a new trend that is now emerging under the name of cognitively inspired NLP, where the effort to leverage from what we know of human cognition to increase machine processing of language data. Drawing inspiration from sequential Monte-Carlo methods such as particle filtering, we illustrated the relevance of our new method for speeding up the computations of direct generative parsing for RNNG, and revealing the potential cognitive interpretation of the underlying representations built by the search method and its beam activity through the analysis of neuro-imaging signal.

A second focus of the project is on compositionality, memory retrieval and syntactic composition during language comprehension. By using quantifications of these hypothesised processes as obtained from computational linguistics we seek to highlight their neural substrates and better understand or model human cognitition.

While linguistic expressions have been binarised as compositional and non-compositional given the lack of compositional linguistic analysis, the so-called Multi-word Expressions (MWEs) demonstrate finer-grained degrees of conventionalisation and predictability in psycho-linguisitcs, which can be quantified through computational Association Measures, like Point-wise Mutual Information and Dice's Coefficient [57]. An fMRI analysis was conducted to investigate to what extent these computational measures and the underlying cognitive processes they reflect are observable during on-line naturalistic sentence processing. Our results show that predictability, as quantified through Dice's Coefficent, is a better predictor of neural activation for processing MWEs and the more cognitively plausible computational metric. Computational results (1348) were obtained on MWE identification in French based on new method searching for frequent dependency-patterns [13]. These identifications in the Little Prince are contrasted with the ones published for English [69] and will yield an fMRI analysis comparing the two languages and the possible typological differences that the two languages may reflect in terms of morphological strategies to achieve lexical conventionalisation.

7.3. Large-scale raw corpus development

Participants: Benoît Sagot, Éric Villemonte de La Clergerie, Laurent Romary, Pedro Ortiz Suárez, Murielle Fabre, Louis Martin, Benjamin Muller, Yoann Dupont.

In order to be in phase (and comparable) with the US partners of the "Petit-Prince" ANR project, Murielle Fabre assembled 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 and 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, called CaBerNET, represents a balanced counterpart to the American COCA corpus. [7]

We have also developed a general, highly parallel, multi-threaded pipeline to clean and classify Common Crawl by language. Common Crawl is a huge (over 20TB), heterogeneous multilingual corpus comprised of documents crawled from the internet, not sorted per language. We designed our pipeline, called goclassy, so that it runs efficiently on medium to low resource infrastructures where I/O speeds are the main constraint. We

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

have created and we distribute a 6.3TB version of Common Crawl, called OSCAR, which is filtered, classified by language, shuffled at line level in order to avoid copyright issues, and ready to be used for NLP applications [29]. OSCAR corpora served as input data to train a variety of neural language models, including the French BERT model CamemBERT (see relevant module for more information). Bridging corpus development, NLP and computational neurolinguistics on of our next step is to train BERT model with the above cited French balanced corpus CaBerNet to create CaBERTnet and extract form it parsing metrics that will be correlated with brain activity as measured by French fMRI recording while listening *Le Petit Prince* in French.

7.4. Neural language modelling

Participants: Benoît Sagot, Djamé Seddah, Éric Villemonte de La Clergerie, Laurent Romary, Louis Martin, Benjamin Muller, Pedro Ortiz Suárez, Yoann Dupont, Ganesh Jawahar.

Pretrained language models are now ubiquitous in Natural Language Processing. Despite their success, most available models have either been trained on English data or on the concatenation of data in multiple languages. This makes practical use of such models—in all languages except English—very limited. In 2019, one of the most visible achievements of the ALMAnaCH team was the training and release of CamemBERT, a BERT-like [75] (rather, RoBERTa-like) neural language model for French trained on the French section of our large-scale web-based OSCAR corpus, together with CamemBERT variants [60]. Our goal was to investigate the feasibility of training monolingual Transformer-based language models for other languages, taking French as an example and evaluating our language models on part-of-speech tagging, dependency parsing, named entity recognition and natural language inference tasks. We have shown that the use of web-crawled data such as found in OSCAR to train such language models is preferable to the use of Wikipedia data, because of the homogeneity of Wikipedia data. More surprisingly, we have also shown that a relatively small web crawled dataset (4GB randomly extracted from the French section of OSCAR) leads to results that are as good as those obtained using larger datasets (130+GB, i.e. the whole French section of OSCAR). CamemBERT allowed us to reach or improve the state of the art in all four downstream tasks.

Beyond training neural language models, we have reinforced the exploration of an active question, that of their interpretability. With the emergence of contextual vector representations of words, such as the ELMo [89] and BERT language models and word embeddings, the interpretability of neural models becomes a key research topic. It is a way to understand what such neural networks actually learn in an unsupervised way from (huge amounts of) textual data, and in which circumstances they manage to do so. The work carried out in the team this year to identify where morphological vs. syntactic vs. semantic information is stored in a BERT language model [26] was part of a more general trend (see for example [78]). And our work on training ELMo models for five mid-resourced languages has shown that such LSTM-based models, when trained on large scale although non edited dataset such as our web-based corpora OSCAR, can lead to outperforming state-of-the-art performance on an number of downstream tasks such as part-of-speech tagging and parsing. Finally, we have carried out comparative evaluations of the performance of CamemBERT and of ELMo models trained on the same French section of OSCAR on a number of downstream task, with an emphasis on named-entity recognition—a work that led us to publish a new version of the named-entity-annotated version of the French TreeBank [67] that we published in 2012 [99].

We have also investigated how word embeddings can capture the evolution of word usage and meaning over time, at a fine-grained scale. 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 since 2018 into studying language variation within user-generated content (UGC), taking into account two main interrelated dimensions: how language variation is related to socio-demographic and dynamic network variables, and how UGC language evolves over time. Taking advantage of the SoSweet corpus (600 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 via the use of dynamic word embeddings, that is embeddings evolving over time. We devised a novel attention model, based on Bernouilli word embeddings, that are conditioned on contextual extra-linguistic features such as network, spatial and socio-economic variables, which can be inferred from Twitter users metadata, 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. We published these ideas and results in [41].

A deep understanding of what is learned, and, beyond that, of how it is learned by neural language models, both synchronic and diachronic, will be a crucial step towards the improvement of such architectures (e.g. targeting low-resource languages or scenarios) and the design and deployment of new generations of neural networks for NLP. Particularly important is to assess the role of the training corpus size and heterogeneity, as well as the impact of the properties of the language at hand (e.g. morphological richness, token-type ratio, etc.). This line of research will also have an impact on our understanding of language variation and on our ability to improve the robustness of neural-network-based NLP tools to such variation.

7.5. Processing non-standard language: user-generated content and code-mixed language

Participants: Djamé Seddah, Benoît Sagot, Éric Villemonte de La Clergerie, Benjamin Muller, Ganesh Jawahar, Abhishek Srivastava, Jose Rosales Nuñez, Hafida Le Cloirec, Farah Essaidi, Matthieu Futeral.

In 2019, we have resumed our long-lasting efforts towards increasing the robustness of our language analysis tools to the variation found in user-generated content (UGC). We have done this in two directions, in the context of the SoSweet and Parsiti projects.

Firstly, we have investigated how our state-of-the-art hybrid (symbolic and statistical) parsing architecture for French, based on SxPipe, FRMG and the Lefff, behaves on French UGC data, namely on 20 millions tweets from the SoSweet corpus. A first observation was that the current level of pre-parsing normalization was not sufficient to ensure a good parsing coverage with FRMG (around 67%, to be compared with around 93% on journalistic texts such as the French TreeBank), also leading to high parsing times because of correction strategies. However, we applied our error mining strategy [6] to identify a first set of easy errors. Clustering and word embedding were also tried for lemmas relying on the dependency parse trees, again leading to semi-successful results due to the poor quality of the pre-parsing phases.

Secondly, we have investigated the normalisation task, whose goal is to transform possibly noisy UGC into less noisy inputs that are more adapted to our standard neural analysis models (e.g. taggers and parsers). More precisely, we have investigated how useful a language model such as BERT [75], trained on standard data, can be in handling non-canonical text. We study the ability of BERT to perform lexical normalisation in a realistic, and therefore low-resource, English UGC scenario [28]. By framing lexical normalisation as a token prediction task, by enhancing its architecture and by carefully fine-tuning it, we have shown that BERT can be a competitive lexical normalisation model without the need of any UGC resources aside from 3,000 training sentences. To the best of our knowledge, it is the first work done in adapting and analysing the ability of this model to handle noisy UGC data.

Thirdly, we have compared the performances achieved by Phrase-Based Statistical Machine Translation systems (PBSMT) and attention-based Neural Machine Translation systems (NMT) when translating UGC from French to English [44]. We have shown that, contrary to what could have been expected, PBSMT outperforms NMT when translating non-canonical inputs. Our error analysis uncovers the specificities of UGC that are problematic for sequential NMT architectures and suggests new avenue for improving NMT models.

Finally, building natural language processing systems for highly variable and low resource languages is a hard challenge. The recent success of large-scale multilingual pretrained neural language models (including our CamemBERT language model for French) provides us with new modeling tools to tackle it. We have studied the ability of the multilingual version of BERT to model an unseen dialect, namely the Latin-script user-generated North African Arabic dialect called Arabizi. We have shown in different scenarios that multilingual language models are able to transfer to such an unseen dialect, specifically in two extreme cases: across script

(Arabic to Latin) and from Maltese, a related language written in the Arabic script, unseen during pretraining. Preliminary results have already been published [66].

7.6. Long-range diachronic variation

Participants: Benoît Sagot, Laurent Romary, Éric Villemonte de La Clergerie, Clémentine Fourrier, Gaël Guibon, Mathilde Regnault, Kim Gerdes.

ALMAnaCH members have resumed their work on longer-range diachronic variation, in two distinct directions:

- Firstly, we have been working on resources and tools for Old French, using contemporary French as a starting point for which resources and tools are available. This work is carried out within the ANR project “Profiterole”, whose goal is 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 as found in such a corpus. Indeed, Old French does not only involve diachronic variation when contrasted with contemporary French. It also involve large internal variation, notably because of diachronic (within Old French), dialectal, geographic, stylistic and genre-based variation. We have carried out experiments on morphosyntactic 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 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 [31], [30]. This work relies on the new morphological and syntactic lexicon for Old French, OFrLex, developed at ALMAnaCH [34]. On the other hand, we conducted parsing experiments with neural models (DyALog’s SRNN models).
- Secondly, we have started experiments to investigate whether and under which conditions neural networks can be used for learning sound correspondences between two related languages, i.e. for predicting cognates of source language words in a related target language. In order to obtain suitably large homogeneously phonetised data, we extracted bilingual lexicons and cognate sets from available resources, including our EtymDB etymological database, of which a new, extended version was created in 2019. This data was then used to train and evaluate several neural architectures (seq2seq, Siamese). Preliminary results are promising, but further investigation is required.

These two research directions will find a common ground now that we have begun to investigate, in the context of the Profiterole ANR project, how we can model the diachronic evolution of the lexicon from Old French to contemporary French. Moreover, our work on Basnage’s 1701 *Dictionnaire Universel*, in the context of the BASNUM ANR project might draw some inspiration from the Profiterole project. But since 1700’s French is much closer from contemporary French than Old French, another source of inspiration for BASNAGE might come from our work on sociolinguistic variation in contemporary French and more generally on our work on User-Generated Content (UCG).

7.7. Syntax and treebanking

Participants: Djamé Seddah, Benoît Sagot, Kim Gerdes, Benjamin Muller, Pedro Ortiz Suárez, Marine Courtin.

In 2019 we have introduced the first treebank for a romanized user-generated content of Algerian, a North-African Arabic dialect called Arabizi. It contains 1500 sentences, fully annotated in morpho-syntax and universal dependencies, and is freely available. We complement it with 50k unlabeled sentences that were collected using intensive data-mining techniques from Common Crawl and web-crawled data. Preliminary results show its usefulness for POS tagging and dependency parsing.

We have also developed the first syntactic treebank for spoken Naija, an English pidgincreole, which is rapidly spreading across Nigeria. The syntactic annotation is developed in the Surface-Syntactic Universal Dependency annotation scheme (SUD) [77] and automatically converted into Universal Dependencies (UD). A crucial step in the syntactic analysis of a spoken language consists in manually adding a markup onto the transcription, indicating the segmentation into major syntactic units and their internal structure. We have shown that this so-called "macrosyntactic" markup improves parsing results. We have also studied some iconic syntactic phenomena that clearly distinguish Naija from English. This work is published in [36].

We have carried out two pilot studies in empirical syntax based on UD treebanks. In a first study [38], we investigate the relationship between dependency distance and frequency based on the analysis of an English dependency treebank. The preliminary result shows that there is a non-linear relation between dependency distance and frequency. This relation between them can be further formalised as a power law function which can be used to predict the distribution of dependency distance in a treebank. In a second study [40], we discussed an empirical refoundation of selected Greenbergian word order univer-sals based on a data analysis of the Universal Dependencies project. The nature of the data we worked on allows us to extract rich details for testing well-known typological universals and constitutes therefore a valuable basis for validating Greenberg's universals. Our results show that we can refine some Greenbergian universals in a more empirical and accurate way by means of a data-driven typological analysis.

Finally, we have introduced a new schema to annotate Chinese Treebanks on the character level. The original UD and SUD projects provide token-level resources with rich morphosyntactic language details. However, without any commonly accepted word definition for Chinese, the dependency parsing always faces the dilemma of word segmentation. Therefore we have presented a character-level annotation schema integrated into the existing Universal Dependencies schema as an extension [39]. The different SUD projects were also presented at the Journées scientifiques "Linguistique informatique, formelle et de terrain" (LIFT 2019), Nov 28-29, 2019 at the University of Orléans.

7.8. Analysing and enriching legacy dictionaries

Participants: Laurent Romary, Benoît Sagot, Mohamed Khemakhem, Pedro Ortiz Suárez, Achraf Azhar.

2019 has been a year of deployment and large-scale experiment of the work initiated in 2016 on the analysis and enrichment of legacy dictionaries and implemented in the GROBID-dictionary framework [84]. GROBID-dictionary is an extension of the generic GROBID Suite [95] and implements an architecture of cascading CRF models with the purpose to parse and categorize components of a pdf documents, whether born-digital or resulting from an OCR. It is developped as part of the doctoral work of Mohamed Khemakhem. GROBID dictionaries produces an output that is conformant to the Text Encoding Initiative guideline and thus easy to distribute and further process in an open science context. We have had the opportunity the show the performances and robustness of the architecture on a variety of dictionaries and contexts resulting both from internal and external collaborations:

- In the context of the language documentation project of Jack Bowers dealing with Mixtepec-Mixtec (ISO 639-3: mix, [72], we have been successful in completely parsing a new edition of an historical lexical resource of Colonial Mixtec 'Voces del Dzaha Dzahui' published by the Dominican fray Francisco Alvarado in the year 1593, published by Jansen and Perez Jiménez (2009). The result is now integrated into the reference lexical description maintained by Jack). See [18];
- Within the Nénufar project, a collaboration with the Praxiling laboratory in Montpellier, we have been contributing to the analyses and encoding of several editions of the Petit Larousse Illustré, a central legacy publication for the French language. [17], [27];
- For the ANR funded project BASNUM, we are deeply involved in understanding how a complex, semi-structured dictionary, for which we do not necessarily have a high quality digitized primary source, can be properly segmented in lexical entries and subfields from which we expect being able to extract fine-grained linguistic content (e.g. named entities for literary sources). In [42], we have shown for instant how the GROBID-dictionary framework could be robust to variations in scanning and thus OCR quality;

- In the same context of the BASNUM project, we have also started to explore the possibility of deploying deep learning components. As shown in [43], the main challenges is the lack of available annotated data in order to train machine learning models, decreased accuracy when using modern pre-trained models due to the differences between present-day and 18th century French, and even unreliable or low quality OCRisation;
- These various experiments have been accompanied by an intense training and hand-on activity in the context in particular of the Lexical Data Master Class and collaboration within the ELEXIS project, which has opted for using the system for building a dictionary matrix from legacy dictionaries [8].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.
- Finally, and as a nice example of the kind of DH collaborations that our researches can lead to, we should mention here the targeted experiments that we carried out on extending the GROBID-dictionary framework to deal with objects which, although analogous with dictionary entries from a distance, appear to have a highly specific structure. This is the case Manuscript Sales Catalogues, which are highly important for authenticating documents and studying the reception of authors. Their regular publication throughout Europe since the beginning of the 19th c. has raised the interest around scaling up the means for automatically structuring their contents. [33] presents the results of advanced tests of the system's capacity to handle a large corpus with MSC of different dealers, and therefore multiple layouts.

7.9. Coreference resolution

Participants: Loïc Grobol, Éric Villemonte de La Clergerie.

In 2019 we have resumed our work on coreference resolution for French with the release in [25] of the first end-to-end automatic coreference resolution system for spoken French by adapting state-of-the art neural network system to the case of noisy non-standard inputs.

This first release uses no external knowledge beyond pretrained non-contextual word embeddings, making it suitable for applications to languages with less pre-existing resources. We also investigated the integration of further knowledge, both in the form of contextual embedding techniques such as CamemBERT and syntactic parsers developped at ALMAnaCH (works to be published in 2020).

8. Bilateral Contracts and Grants with Industry

8.1. Bilateral Contracts with Industry

Ongoing contracts:

Verbatim Analysis Verbatim Analysis is an Inria start-up co-created in 2009 by Benoît Sagot. It uses some of ALMAnaCH's free NLP software (SxPipe) as well as a data mining solution co-developed by Benoît Sagot, VERA, for processing employee surveys with a focus on answers to open-ended questions.

opensquare was co-created in December 2016 by Benoît Sagot with 2 senior specialists of HR (human resources) consulting. It is dedicated to designing, carrying out and analysing employee surveys as well as HR consulting based on these results. It uses a new employee survey analysis tool, *enqi*, which is still under development. This tool being co-owned by opensquare and Inria, both parties have signed a Software Licence Agreement in exchange for a yearly fee paid by opensquare to ALMAnaCH based on its turnover. Benoît Sagot currently contributes to opensquare, under the "Concours scientifique" scheme.

[8] https://grobid.elex.is

Facebook A collaboration on text simplification ("français Facile À Lire et à Comprendre", FALC) is ongoing with Facebook's Parisian FAIR laboratory. It involves a co-supervised (CIFRE) PhD thesis in collaboration with UNAPEI, the largest French federation of associations defending and supporting people with special needs and their families. This collaboration, is part of a larger initiative called Cap'FALC involving (at least) these three partners as well as the relevant ministries. Funding received as a consequence of the CIFRE PhD thesis: 60,000 euros

Bluenove A contract with this company has been signed in 2018, which initiated a collaboration for the integration of NLP tools within Bluenove's platform Assembl, dedicated to online employee and citizen debating forums. It involved 12 months of fixed-term contracts (a post-doc, who worked at ALMAnaCH in 2018-2019). Funding received: 77,137 euros

Active collaborations without a contract:

Science Miner ALMAnaCH (following ALPAGE) has collaborated since 2014 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 for 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 to provide a scholarly dashboard on scientific papers available from the HAL national publication repository.

Software Heritage , whose goal is to collect and preserve software in source code form. ALMAnaCH's collaboration with Software Heritage, on large-scale programming language identification, also involves Qwant, who provided some funding to Software Heritage.

Fortia Financial Solutions ALMAnaCH members led a proposal for the creation of an ANR LabCom with this French FinTech company on the analysis of (raw, PDF) financial documents from investment funds. The proposal was rejected, but future collaboration is still planned.

Hyperlex A collaboration was initiated in 2018 on NLP and information extraction from raw legal documents (mostly PDF format), involving especially Éric de La Clergerie, who is now a part-time employee of the company.

Ongoing discussions that should/could be formalised in the form of a contract in 2020:

Winespace : information extraction from wine descriptions to develop a wine recommendation system

Newsbridge : automatic extraction of short summaries of filmed events (e.g. sport events) based on social media coverage analysis

INPI : Patent classification

Cour de cassation (in the context of the LabIA): retrieval of relevant jurisprudence

DGCCRF (in the context of the LabIA): automatic identification of illicit clauses in B-to-C commercial contracts

9. Partnerships and Cooperations

9.1. National Initiatives

9.1.1. 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.

9.1.2. Competitivity Clusters and Thematic Institutes

- **PRAIRIE institute** (2019-2024, Dir.: Isabelle Ryl). Benoît Sagot was granted a Chair in this newly created research institute dedicated to Artificial Intelligence.
- **GDR LiFT** (2019-): LiFT is a CNRS-funded national coordination structure (GDR) involving many French teams involved in computational, formal and descriptive linguistics, in order to facilitate the emergence of fruitful collaborations. ALMAnaCH is involved in the GDR.
- **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]).

9.1.3. Other National Initiatives

- **LECTAUREP project** (2017-2019): A preliminary 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 digitised 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** (2017-2020): 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.

9.2. European Initiatives

9.2.1. 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.

9.2.2. 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

9.2.3. Collaborations with Major European Organizations

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

- 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]
- Bar Ilan University (Yoav Goldberg, Hila Gonen) [non-canonical text processing]
- Dublin City University, Ireland (Teresa Lynn) [low-resource languages, user-generated content]
- University of Sheffield, United Kingdom (Lucia Specia, Carolina Scarton, Fernando Alva-Manchego) [text simplification]
- Univerza v Ljubljani [University of Ljubljana], Ljubljana, Slovenia (Darja Fišer) [wordnet development]

9.3. International Initiatives

9.3.1. Participation in Other International Programs

ANR-NSF project MCM-NL "Petit Prince" (2016-2020, PI John Hale [Cornell University, USA], resp. for Inria Paris/ALMAnaCH: Éric de La Clergerie) Topic: exploring correlations between data from neuro-imagery (fMRI, EEG) and data from NLP tools (mostly parsers). The data will come from "Le Petit Prince" read in French and English, and parsed with different parsers. Other partners: Cornell Univ., Univ. Michigan, Paris Saclay/Neurospin, Univ. Paris 8. Grant for ALMAnaCH: 108,500 euros

PHC Maïmonide (2018-2019, PI Djamé Seddah, co-PI Yoav Goldberg [Bar Ilan University]). Topics: Building NLP resources for analysing reactions to major events in Hebrew and French social media. Amount of the grant for the French side: 59,000 euros (89,000 euros for the whole project).

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific Events: Organisation

10.1.1.1. General Chair, Scientific Chair

- BS organised with Annie Rialland and Catherine Schnedecker the 2019 “Journée de la Société de Linguistique de Paris”, dedicated this year to “Corpus, analyses quantitatives et modèles linguistiques” (January 2019)
- DS was the co-chair of the SyntaxFest (Paris, 26-30 August 2019), a one-week event colocating a number of previously independent conferences and workshops on topics ranging from syntax to parsing and treebank development (including the UD workshop), as well as the co-chair of one of these conferences (Treebank and Linguistic Theories, TLT).
- KG was the co-chair of the International Conference on Dependency Linguistics, Depling 2019, another event of the SyntaxFest.

10.1.2. Scientific Events: Selection

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

- BS: Member of the Program, Scientific or Reviewing Committee of the following conferences and workshops: EMNLP-IJCNLP 2019, ACL 2019, NAACL 2019
- LR: Member of the Program, Scientific or Reviewing Committee of the following conferences and workshops: CMLC 2019, eLex 2019, ISA-15, TOTh 2019, LDK 2019, DATeCH 2019, ElPub 2019
- KG: Member of the Program, Scientific or Reviewing Committee of the following conferences and workshops: ACL 2019, SyntaxFest 2019 including Depling and Workshop on Universal Dependencies, Workshop on Multilingual Surface Realization, 2019, CoNLL 2019, EMNLP-IJCNLP 2019, International Conference on Natural Language Generation (INLG) 2019, LAW XIII 2019 (The 13th Linguistic Annotation Workshop), Rencontre des Jeunes Chercheurs en Parole

10.1.3. Journal

10.1.3.1. Member of the Editorial Boards

- LR: Member of the scientific committe of the French speaking DH journal *Humanités numériques*
- KG: Member of the scientific committe of the *Journal of Linguistics*

10.1.3.2. Reviewer - Reviewing Activities

- BS: Reviewer for *Traitement Automatique des Langues*, *wék*^{W}*os*
- AC: Reviewer for *Digital Humanities 2020 - Intersections/Carrefours*

10.1.4. Invited Talks

- DS, invited talk, ”Faire face au syndrome du Jabberwocky: Analyses morpho-syntaxiques en environnement hostile”, at the Societé Linguistique de Paris (SLP) (Janvier 2019)
- BS, “Représentation et exploitation des informations lexicales”, CENTAL, Louvain-la-Neuve, Belgium (April 2019)
- BS, “Morphological complexities”, joint invited talk (with Géraldine Walther) at ACL 2019’s SIGMORPHON workshop, Florence, Italy (July 2019) [15]

- LR, “The TEI as a modeling infrastructure: TEI beyond the TEI realms”. Ringvorlesung Digital Humanities, Paderborn, Germany (July 2019).
- LR, “Where is the place of scientific lexicography in science?”, Lorentz workshop: The Future of Academic Lexicography, Leiden, The Netherlands (Nov. 2019)
- LR, “Digital transition of SSH”, LERU Information and Open Access (INFO) Policy Group meeting, Strasbourg, France (Dec. 2019)
- EVdlC, “Acquisition de connaissances à partir de corpus analysés syntaxiquement” Seminar at Nokia Bell Labs, France (June 2019)
- EVdlC, “Évolutions et pertinence des plongements lexicaux”, workshop on “Machine learning, données textuelles et recherche en sciences humaines et sociales”, ENS Lyon, France (Nov. 2019)
- Benjamin Muller, “Transfer Learning on an unseen North African Arabic Dialect”, Bar Ilan University, Israel (24 Nov 2019)
- Mohamed Khemakhem was invited to the “ELEXIS Observers” event to give a talk [9] about structuring historical dictionaries and the use of GROBID-Dictionaries, Austrian Academy of Sciences, Austria (February 2019)
- KG, “Instant treebanks: Tools and methods to quickly build syntactically annotated corpora from scratch”, Workshop “Annotation of non-standard corpora (2019)”, University of Bamberg, Germany in September.

10.1.5. Training

- Mohamed Khemakhem chaired and tutored the GROBID-Dictionaries workshop series:
  - ELEXIS workshop - Berlin (March 2019)
  - Atelier de formation CollEx-Persée - Paris (September 2019)
- Following the aforementioned ‘Symposium on Naija’, a Master Class of one week was organised in June 2019 at the ARCIS institute of the University of Ibadan, Nigeria, on “Crowd-sourcing Web Corpora of Nigerian Languages” taught by KG and Slavomír Čéplö (from the Austrian Academy of Sciences).

10.1.6. Leadership within the Scientific Community

An important aspect of ALMAnaCH’s work relates to standardisation initiatives, especially within the ISO TC37 committee on language and terminology. Laurent Romary is the President of this committee and the convenor of its working group TC37/SC4 on lexical resources, within which he and a number of other ALMAnaCH members (Jack Bowers, Mohamed Khemakhem, Benoît Sagot, Éric de La Clergerie) have responsibilities (as project leaders or co-leaders). Most of them are related to the revision as a multi-part standard of ISO 24613 (Lexical Markup Framework, [14]), the first part of which was published in June 2019.

ALMAnaCH members have also played a key role in developing the Standardisation Survival Kit, an online tool hosted by Huma-num which focuses on giving researchers access to standards in a meaningful way by using research scenarios which cover all the domains of the Humanities, from literature to heritage science, including history, social sciences, linguistics, etc. We have published one publications on this topic in 2019 [64], following numerous ones over the last few years.

Other examples of ALMAnaCH’s leadership within the scientific community are the following:

- 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 (ended in 2019)
- 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.
- Mohamed Khemakhem: Member of the DARIAH Working Group “Bibliographical Data”

[9] http://videolectures.net/elexisobserver2019_khemakhem_dictionaries/

10.1.7. Scientific Expertise

- LR is an advisor for scientific information to Inria's deputy CEO for Science.

10.1.8. Research Administration

- BS is a member of Inria Paris's Scientific Committee ("Comité des Projets") and of its Board ("Bureau du Comité des Projets"), and a member of the International Relations Working Group of Inria's Scientific and Technological Orientation Council (COST-GTRI)
- BS is the Deputy Head (and former 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
- LR is the President of the scientific committee of ABES (Agence Bibliographique de l'Enseignement Supérieur), and a Member of the Text Encoding Initiative board

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

- Master: Benoît Sagot (with Emmanuel Dupoux), "Speech and Language Processing", 20h, M2, Master "Mathématiques, Vision, Apprentissage", ENS Paris-Saclay, France
- Master: Loïc Grobol, "Introduction à la fouille de textes", 39h, M1, Université Sorbonne Nouvelle, France
- Master: Loïc Grobol, "Langages de script", 39h, M2, INaLCO, France
- Eq. Master: Clémentine Fourrier, "Module TDLog - Techniques de Developpement Logiciel" (computer science opt-in course in Python), 18h, M1 and M2, École des Ponts ParisTech, France
- Eq. Bachelor second year: Gaël Guibon, "Introduction to Python and Programming" (for business school students and english speakers) 25h, International BBA 2, ESSEC Business School, France
- Master: Alix Chagué, "Humanités Numériques", 18h, M1, Institut d'études culturelles et internationales (IECI), Université Versailles-Saint-Quentin-en-Yvelines, France.
- Master: Alix Chagué, "Introduction à LaTeX", 5h, M1, Institut d'études culturelles et internationales (IECI), Université Versailles-Saint-Quentin-en-Yvelines, France.
- Master: Alix Chagué, "Introduction à Linux", 5h, M1, Institut d'études culturelles et internationales (IECI), Université Versailles-Saint-Quentin-en-Yvelines, France.
- Licence: Pedro Ortiz, "Mathématiques discrètes", 40h, L2, Licence de Sciences et Technologies, Faculté de Sciences et Ingénierie, Sorbonne Université, Paris, France.
- Licence: Mathilde Regnault, "Informatique et industries de la langue", 20h, L2, Licence Sciences du langage, Université Sorbonne nouvelle - Paris 3, Paris, France.

10.2.2. Supervision

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, "Coreference resolution for spoken French", "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: Louis Martin, “Text Simplification”, June 2018, Facebook & Sorbonne Université (“CIFRE” PhD), supervised by Benoît Sagot 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, October 2018

PhD in progress: Clémentine Fourrier, “Neural approaches to the modelling of phonetic evolution”, October 2019, EPHE (Inria fellowship), Benoît Sagot

10.2.3. Juries

- BS: reviewer (“rapporteur”) of the PhD committee for Nourredine Alliane at Université Paris 8 on 17th May (Title: “Évaluation des représentations vectorielles de mots”; Supervisor: Gilles Bernard)
- BS: reviewer of the PhD committee for Tamara Álvarez López at Université de Vigo, Spain, on 29th November (Title: “Sentiment analysis in social media contents using natural language processing techniques”; Supervisors: Enrique Costa Montenegro and Milagros Fernández Gavilanes)
- BS: member of the PhD committee for Hazem Al Saied at Université de Lorraine on 20th December (Title: “Analyse automatique par transitions pour l’identification des expressions polylexicales”; Supervisors: Matthieu Constant and Marie Candito)
- LR: Reviewer of the PhD committee for Elina Leblanc, Université de Grenoble Alpes (title: “Bibliothèques numériques enrichies et participatives : utilisateurs, services, interfaces”, Supervisors: Elena Pierazzo and Hervé Blanchon)
- EVdlC: Reviewer of the PhD committee for Diana Nicoleta Popa, Université de Grenoble Alpes (title: “Vers des représentations contextualisées de mots”, Supervisor: Éric Gaussier)
- KG: reviewer of the PhD committee for Mohammed Galal at the University of Ain Shams, Egypt, on January 8 (Title: “Les constructions exceptives du français et de l’arabe : syntaxe et interface sémantique-syntaxe”; Supervisor: Sylvain Kahane in codirection with Dina El-Kassas)
- KG: President of the PhD committee for Marie-Amélie Botalla at the Sorbonne Nouvelle on May 14 (Title: “Modélisation de la production des énoncés averbaux : le cas des compléments différés”, Supervisors: Jeanne-Marie Debaisieux and Sylvain Kahane)

10.3. Popularization

10.3.1. Articles and contents

- BS was interviewed by David Larousserie, from the French daily newspaper Le Monde, about the release of CamemBERT, in October and December 2019. This resulted in an article in the newspaper, accessible here (subscription required).
- LR was interviewed the radio France Culture about the release of CamemBERT, in October 2019. The interview was aired on the radio.
- AC has taken in charge the content management of the Time Us project’s research blog since April 2019, which aims at presenting sources, good practices and results used and elaborated by all the project members.

10.3.2. Interventions

- Welcoming of schoolchildren at Inria Paris (half a day with ALMAnaCH members within an one-week-long stay; December 2019)
- Clémentine Fourrier co-organised the Young Women Mathematician Days (Inria Paris edition): 2 days of talks, mathematical animations and speed meetings with women scientists or engineers at the Inria Paris center, for high school girls
- Clémentine Fourrier was a speaker at the ESPCI highschool students summer school, to introduce the different jobs of computer science engineer to high school girls, in research and the industry, and animated an introductory software development workshop in Python over an afternoon
- Alix Chagué was the speaker for 6th workshop organised by the Alumni Association of the École nationale des chartes's Masters (ADEMEC), to introduce the software Transkribus and guidelines for text extraction (March 2019).

10.3.3. Internal action

- BS presented he January 2019 edition of the "demi-heure de science" at Inria Paris, whose target audience is researchers from all Inria Paris teams
- BS presented he June 2019 edition of Inria Paris' "my team in 180 seconds" event, whose target audience is both researchers and non-scientific Inria personnel
- CF presented softwares and good practices (Docker, GitLab and version control, Diverting Gitlab to create a small website) at the Inria developper meetup, whose target audience is both researchers and engineers.

11. 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^o^ 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, pp. 248-251, 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^o^ 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 "The 21st International Conference of the Association for Computational Linguistics (ACL 2006)", Sydney, Australia, July 2006, pp. 329-336, https://hal.inria.fr/hal-02270412

[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

Articles in International Peer-Reviewed Journals

[11] D. Reineke, L. Romary. *Bridging the gap between SKOS and TBX*, in "edition - Die Fachzeitschrift für Terminologie", November 2019, vol. 19, n^{O} 2, https://hal.inria.fr/hal-02398820

[12] L. Romary, C. Riondet. *Towards multiscale archival digital data*, in "Umanistica digitale", 2019 [*DOI :* 10.6092/ISSN.2532-8816/9045], https://hal.inria.fr/hal-01586389

Invited Conferences

[13] M. Fabre, Y. Dupont, É. Villemonte de La Clergerie. *Syntactic Parsing versus MWEs: What can fMRI signal tell us*, in "PARSEME-FR 2019 consortium meeting", Blois, France, PARSEME-FR 2019, June 2019, https://hal.inria.fr/hal-02272288

[14] L. Romary, M. Khemakhem, F. Khan, J. Bowers, N. Calzolari, M. George, M. Pet, P. Bański. *LMF Reloaded*, in "AsiaLex 2019: Past, Present and Future", Istanbul, Turkey, June 2019, https://arxiv.org/abs/1906.02136 , https://hal.inria.fr/hal-02118319

[15] G. Walther, B. Sagot. *Morphological complexities*, in "16th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology", Florence, Italy, August 2019, https://hal.inria.fr/hal-02266999

International Conferences with Proceedings

[16] F. Alva-Manchego, L. Martin, C. Scarton, L. Specia. *EASSE: Easier Automatic Sentence Simplification Evaluation*, in "EMNLP-IJCNLP 2019 - Conference on Empirical Methods in Natural Language

Processing and 9th International Joint Conference on Natural Language Processing (demo session)", Hong Kong, China, November 2019, pp. 49-54, https://hal.inria.fr/hal-02272950

[17] H. BOHBOT, F. FRONTINI, F. KHAN, M. KHEMAKHEM, L. ROMARY. *Nénufar: Modelling a Diachronic Collection of Dictionary Editions as a Computational Lexical Resource*, in "ELEX 2019: smart lexicography", Sintra, Portugal, October 2019, https://hal.inria.fr/hal-02272978

[18] J. BOWERS, M. KHEMAKHEM, L. ROMARY. *TEI Encoding of a Classical Mixtec Dictionary Using GROBID-Dictionaries*, in "ELEX 2019: Smart Lexicography", Sintra, Portugal, October 2019, https://hal.inria.fr/hal-02264033

[19] J. BOWERS, L. ROMARY. *TEI and the Mixtepec-Mixtec corpus: data integration, annotation and normalization of heterogeneous data for an under-resourced language*, in "6th International Conference on Language Documentation and Conservation (ICLDC)", Honolulu, United States, February 2019, https://hal.inria.fr/hal-02075475

[20] B. CRABBÉ, M. FABRE, C. PALLIER. *Variable beam search for generative neural parsing and its relevance for the analysis of neuro-imaging signal*, in "EMNLP-IJCNLP 2019 - Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing", Hong-Kong, China, November 2019, https://hal.inria.fr/hal-02272303

[21] M. DINARELLI, L. GROBOL. *Hybrid Neural Networks for Sequence Modelling : The Best of Three Worlds*, in "TALN-RECITAL 2019 - 26ème Conférence sur le Traitement Automatique des Langues Naturelles", Toulouse, France, Conférence sur le Traitement Automatique des Langues Naturelles (TALN-RECITAL), ATALA, July 2019, https://hal.archives-ouvertes.fr/hal-02157160

[22] M. DINARELLI, L. GROBOL. *Seq2Biseq: Bidirectional Output-wise Recurrent Neural Networks for Sequence Modelling*, in "CICLing 2019 - 20th International Conference on Computational Linguistics and Intelligent Text Processing", La Rochelle, France, April 2019, https://hal.inria.fr/hal-02085093

[23] L. FOPPIANO, L. ROMARY, M. ISHII, M. TANIFUJI. *Automatic Identification and Normalisation of Physical Measurements in Scientific Literature*, in "DocEng '19 - ACM Symposium on Document Engineering 2019", Berlin, Germany, ACM Press, September 2019, pp. 1-4 [*DOI :* 10.1145/3342558.3345411], https://hal.inria.fr/hal-02294424

[24] S. GABAY, L. RONDEAU DU NOYER, M. KHEMAKHEM. *Selling autograph manuscripts in 19th c. Paris: digitising the Revue des Autographes*, in "IX Convegno AIUCD", Milan, Italy, AIUCD, January 2020, https://hal.archives-ouvertes.fr/hal-02388407

[25] L. GROBOL. *Neural Coreference Resolution with Limited Lexical Context and Explicit Mention Detection for Oral French*, in "Second Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC19)", Minneapolis, United States, June 2019, https://hal.inria.fr/hal-02151569

[26] G. JAWAHAR, B. SAGOT, D. SEDDAH. *What does BERT learn about the structure of language?*, in "ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics", Florence, Italy, July 2019, https://hal.inria.fr/hal-02131630

[27] A. F. KHAN, H. BOHBOT, F. FRONTINI, M. KHEMAKHEM, L. ROMARY. *Historical Dictionaries as Digital Editions and Connected Graphs: the Example of Le Petit Larousse Illustré*, in "Digital Humanities 2019", Utrech, Netherlands, July 2019, https://hal.inria.fr/hal-02111199

[28] B. MULLER, B. SAGOT, D. SEDDAH. *Enhancing BERT for Lexical Normalization*, in "The 5th Workshop on Noisy User-generated Text (W-NUT)", Hong Kong, China, November 2019, https://hal.inria.fr/hal-02294316

[29] P. J. ORTIZ SUÁREZ, B. SAGOT, L. ROMARY. *Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures*, in "7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)", Cardiff, United Kingdom, P. BAŃSKI, A. BARBARESI, H. BIBER, E. BREITENEDER, S. CLEMATIDE, M. KUPIETZ, H. LÜNGEN, C. ILIADI (editors), Leibniz-Institut für Deutsche Sprache, July 2019, https://hal.inria.fr/hal-02148693

[30] M. REGNAULT, S. PRÉVOST, É. VILLEMONTE DE LA CLERGERIE. *Challenges of language change and variation: towards an extended treebank of Medieval French*, in "TLT 2019 - 18th International Workshop on Treebanks and Linguistic Theories", Paris, France, August 2019, https://hal.inria.fr/hal-02272560

[31] M. REGNAULT. *Adapting a Metagrammar for Contemporary French to Medieval French*, in "TALN-RECITAL 2019 - 26e édition de la conférence TALN (Traitement Automatique des Langues Naturelles) et 21e édition de la conférence jeunes chercheur·euse·s RECITAL", Toulouse, France, July 2019, https://hal.inria.fr/hal-02147686

[32] L. ROMARY. *The place of lexicography in (computer) science*, in "The Future of Academic Lexicography: Linguistic Knowledge Codification in the Era of Big Data and AI", Leiden, Netherlands, Frieda Steurs and Dirk Geeraerts and Niels Schiller and Marian Klamer and Iztok Kosem, November 2019, https://hal.inria.fr/hal-02358218

[33] L. RONDEAU DU NOYER, S. GABAY, M. KHEMAKHEM, L. ROMARY. *Scaling up Automatic Structuring of Manuscript Sales Catalogues*, in "TEI 2019: What is text, really? TEI and beyond", Graz, Austria, September 2019, https://hal.inria.fr/hal-02272962

[34] B. SAGOT. *Development of a morphological and syntactic lexicon of Old French*, in "26ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN)", Toulouse, France, July 2019, https://hal.inria.fr/hal-02148701

Conferences without Proceedings

[35] S. BASSETT, L. WESSELS, S. KRAUWER, B. MAEGAARD, H. HOLLANDER, F. ADMIRAAL, L. ROMARY, F. UITERWAAL. *Connecting the Humanities through Research Infrastructures*, in "4th Digital Humanities in the Nordic Countries (DHN 2019)", Copenhagen, Denmark, March 2019, https://hal.inria.fr/hal-02047512

[36] B. CARON, M. COURTIN, K. GERDES, S. KAHANE. *A Surface-Syntactic UD Treebank for Naija*, in "TLT 2019, Treebanks and Linguistic Theories, Syntaxfest", Paris, France, August 2019, https://hal.archives-ouvertes.fr/hal-02270530

[37] A. CHAGUÉ, V. LE FOURNER, M. MARTINI, É. VILLEMONTE DE LA CLERGERIE. *Deux siècles de sources disparates sur l'industrie textile en France : comment automatiser les traitements d'un corpus non-uniforme ?*, in "Colloque DHNord 2019 "Corpus et archives numériques"", Lille, France, MESHS Lille Nord de France, October 2019, https://hal.inria.fr/hal-02448921

[38] X. Chen, K. Gerdes. *The relation between dependency distance and frequency*, in "Quasy 2019, Quantitative Syntax 2019, Syntaxfest", Paris, France, August 2019, https://hal.archives-ouvertes.fr/hal-02270528

[39] C. Dong, Y. Li, K. Gerdes. *Character-level Annotation for Chinese Surface-Syntactic Universal Dependencies*, in "Depling 2019 - International Conference on Dependency Linguistics", Paris, France, August 2019, https://hal.archives-ouvertes.fr/hal-02270535

[40] K. Gerdes, S. Kahane, X. Chen. *Rediscovering Greenberg's Word Order Universals in UD*, in "UDW, Universal Dependencies Workshop 2019, Syntaxfest", Paris, France, August 2019, https://hal.archives-ouvertes.fr/hal-02270531

[41] G. Jawahar, D. Seddah. *Contextualized Diachronic Word Representations*, in "1st International Workshop on Computational Approaches to Historical Language Change 2019 (colocated with ACL 2019)", Florence, Italy, August 2019, https://hal.archives-ouvertes.fr/hal-02194763

[42] M. Khemakhem, I. Galleron, G. Williams, L. Romary, P. J. Ortiz Suárez. *How OCR Performance can Impact on the Automatic Extraction of Dictionary Content Structures*, in "19th annual Conference and Members' Meeting of the Text Encoding Initiative Consortium (TEI) - What is text, really? TEI and beyond", Graz, Austria, September 2019, https://hal.archives-ouvertes.fr/hal-02263276

[43] P. J. Ortiz Suárez, L. Romary, B. Sagot. *Preparing the Dictionnaire Universel for Automatic Enrichment*, in "10th International Conference on Historical Lexicography and Lexicology (ICHLL)", Leeuwarden, Netherlands, June 2019, https://hal.inria.fr/hal-02131598

[44] J. C. Rosales Nunez, D. Seddah, G. Wisniewski. *A Comparison between NMT and PBSMT Performance for Translating Noisy User-Generated Content*, in "The 22nd Nordic Conference on Computational Linguistics (NoDaLiDa'19)", Turku, Finland, September 2019, https://hal.archives-ouvertes.fr/hal-02270524

Scientific Books (or Scientific Book chapters)

[45] *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, August 2019, https://hal.inria.fr/hal-02450315

[46] PARTHENOS (editor). *Share - Publish - Store - Preserve. Methodologies, Tools and Challenges for 3D Use in Social Sciences and Humanities*, PARTHENOS and consortium 3D-SHS and LIA MAP-ISTI, Marseille, France, May 2019, https://hal.archives-ouvertes.fr/hal-02155055

[47] J. Edmond, F. Fischer, L. Romary, T. Tasovac. *9. Springing the Floor for a Different Kind of Dance : Building DARIAH as a Twenty-First-Century Research Infrastructure for the Arts and Humanities*, in "Digital Technology and the Practices of Humanities Research", Open Book Publishers, February 2020, pp. 207-234 [*DOI :* 10.11647/OBP.0192.09], https://hal.inria.fr/hal-02464622

[48] J. Edmond, L. Romary. *3. Academic Publishing*, in "Digital Technology and the Practices of Humanities Research", Open Book Publishers, February 2020, pp. 49-80 [*DOI :* 10.11647/OBP.0192.03], https://hal.inria.fr/hal-02464616

[49] K. Gerdes, S. Kahane, R. Bawden, J. Beliao, É. Villemonte de La Clergerie, I. Wang. *Annotation tools for syntax*, in "Rhapsodie: A Prosodic and Syntactic Treebank for Spoken French", June 2019, https://hal.inria.fr/hal-02450311

[50] S. Kahane, K. Gerdes, R. Bawden. *The microsyntactic annotation*, in "Rhapsodie: A Prosodic and Syntactic Treebank for Spoken French", June 2019, https://hal.inria.fr/hal-02450018

[51] S. Kahane, P. Pietrandrea, K. Gerdes. *The annotation of list structures*, in "Rhapsodie: A Prosodic and Syntactic Treebank for Spoken French", June 2019, https://hal.inria.fr/hal-02450034

[52] L. Romary, J. Edmond. *A Tangential View on Impact for the Arts and Humanities through the Lens of the DARIAH-ERIC*, in "Stay Tuned To The Future - Impact of the Research Infrastructures for Social Sciences and Humanities", B. Maegaard, R. Pozzo (editors), Leo S. Olschki Editore, 2019, https://hal.inria.fr/hal-02094713

Other Publications

[53] A. Bertino, L. Foppiano, L. Romary, P. Mounier. *Leveraging Concepts in Open Access Publications*, March 2019, working paper or preprint, https://hal.inria.fr/hal-01981922

[54] 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

[55] J. Bowers. *Pathways and patterns of metaphor & metonymy in Mixtepec-Mixtec body-part terms*, February 2020, working paper or preprint, https://hal.inria.fr/hal-02075731

[56] Y. Dupont. *Un corpus libre, évolutif et versionné en entités nommées du français*, July 2019, TALN 2019 - Traitement Automatique des Langues Naturelles, Poster, https://hal.archives-ouvertes.fr/hal-02448590

[57] M. Fabre, S. Bhattasali, C. Pallier, J. Hale. *Modeling Conventionalization and Predictability in Multi-Word Expressions at Brain-level*, September 2019, CRCNS 2019, Poster, https://hal.inria.fr/hal-02272435

[58] M. Fabre, B. Crabbe, C. Pallier. *Variable beam search for generative neural parsing and its fit with neuro-imaging signal*, September 2019, CRCNS 2019, Poster, https://hal.inria.fr/hal-02272475

[59] K. Gerdes, B. Guillaume, S. Kahane, G. Perrier. *Pourquoi se tourner vers le SUD : L'importance de choisir un schéma d'annotation en dépendance surface-syntaxique*, November 2019, LIFT 2019 - Journées scientifiques "Linguistique informatique, formelle & de terrain", https://hal.inria.fr/hal-02449922

[60] L. Martin, B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, É. Villemonte de La Clergerie, D. Seddah, B. Sagot. *CamemBERT: a Tasty French Language Model*, October 2019, https://arxiv.org/abs/1911.03894 - Web site: https://camembert-model.fr, https://hal.inria.fr/hal-02445946

[61] L. Martin, B. Sagot, É. Villemonte de La Clergerie, A. Bordes. *Controllable Sentence Simplification*, October 2019, https://arxiv.org/abs/1910.02677 - Code and models: https://github.com/facebookresearch/access [*DOI :* 10.02677], https://hal.inria.fr/hal-02445874

[62] A. Málaga Sabogal, S. Troubetzkoy. *Unique ergodicity for infinite area Translation Surfaces*, August 2019, https://arxiv.org/abs/1908.04019 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-02265283

[63] C. Rochereau, B. Sagot, E. Dupoux. *Modeling German Verb Argument Structures: LSTMs vs. Humans*, December 2019, https://arxiv.org/abs/1912.00239 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-02417640

[64] L. Romary, D. Biabiany, K. Illmayer, M. Puren, C. Riondet, D. Seillier, L. Tadjou. *SSK by example - Make your Arts and Humanities research go standard*, May 2019, DARIAH Annual Event, Poster, https://hal.inria.fr/hal-02151788

[65] L. Romary. *The TEI as a modeling infrastructure: TEI beyond the TEI realms*, July 2019, Ringvorlesung Digital Humanities, https://hal.inria.fr/hal-02265036

[66] A. Srivastava, B. Muller, D. Seddah. *Unsupervised Learning for Handling Code-Mixed Data: A Case Study on POS Tagging of North-African Arabizi Dialect*, October 2019, EurNLP - First annual EurNLP, Poster, https://hal.archives-ouvertes.fr/hal-02270527

References in notes

[67] A. Abeillé, L. Clément, F. Toussenel. *10*, in "Building a Treebank for French", Kluwer, Dordrecht, 2003, pp. 165-187

[68] 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

[69] S. Bhattasali, M. Fabre, W.-M. Luh, H. Al Saied, M. Constant, C. Pallier, J. R. Brennan, R. N. Spreng, J. Hale. *Localising Memory Retrieval and Syntactic Composition: An fMRI Study of Naturalistic Language Comprehension*, in "Language, Cognition and Neuroscience", 2018, vol. 34, n^{o} 4, pp. 1-20 [*DOI :* 10.1080/23273798.2018.1518533], https://hal.archives-ouvertes.fr/hal-01930201

[70] 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

[71] 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

[72] 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

[73] 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

[74] 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

[75] J. Devlin, M. Chang, K. Lee, K. Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, in "Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)", 2019, pp. 4171–4186, https://www.aclweb.org/anthology/N19-1423/

[76] 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

[77] K. Gerdes, B. Guillaume, S. Kahane, G. Perrier. *SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD*, in "Universal Dependencies Workshop 2018", Brussels, Belgium, November 2018, https://hal.inria.fr/hal-01930614

[78] J. Hewitt, C. D. Manning. *A Structural Probe for Finding Syntax in Word Representations*, in "Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies", Association for Computational Linguistics, 2019, https://nlp.stanford.edu/pubs/hewitt2019structural.pdf

[79] 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

[80] 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

[81] 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

[82] 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

[83] 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

[84] 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

[85] 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

[86] 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.

[87] 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://hal.inria.fr/hal-01959054

[88] H. MARTÍNEZ 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

[89] M. E. PETERS, M. NEUMANN, M. IYYER, M. GARDNER, C. CLARK, K. LEE, L. ZETTLEMOYER. *Deep Contextualized Word Representations*, in "Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)", 2018, pp. 2227–2237, https://www.aclweb.org/anthology/N18-1202/

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

[91] 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.

[92] 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

[93] 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

[94] L. ROMARY, M. KHEMAKHEM, F. KHAN, J. BOWERS, N. CALZOLARI, M. GEORGE, M. PET, P. BAŃSKI. *LMF Reloaded*, in "AsiaLex 2019: Past, Present and Future", Istanbul, Turkey, June 2019, https://hal.inria.fr/hal-02118319

[95] L. ROMARY, P. LOPEZ. *GROBID - Information Extraction from Scientific Publications*, in "ERCIM News", January 2015, vol. 100, https://hal.inria.fr/hal-01673305

[96] 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

[97] B. SAGOT, H. MARTÍNEZ 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

[98] 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

[99] B. SAGOT, M. RICHARD, R. STERN. *Annotation référentielle du Corpus Arboré de Paris 7 en entités nommées*, in "Traitement Automatique des Langues Naturelles (TALN)", Grenoble, France, G. ANTONIADIS, H. BLANCHON, G. SÉRASSET (editors), Actes de la conférence conjointe JEP-TALN-RECITAL 2012, June 2012, vol. 2 - TALN, https://hal.inria.fr/hal-00703108

[100] 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

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

[102] 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

[103] 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

[104] 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

[105] S. SCHUSTER, É. VILLEMONTE DE LA CLERGERIE, M. 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

[106] 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

[107] 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

[108] 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

[109] 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

[110] 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

[111] 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

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

[113] 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