
# Activity Report 2017

# Team ALMANACH

## Automatic Language Modelling and ANAlysis & Computational Humanities

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

RESEARCH CENTER
**Paris**

THEME
**Language, Speech and Audio**

# Table of contents

# Team ALMANACH

*Creation of the Team: 2017 January 01*

**Keywords:**

**Computer Science and Digital Science:**

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

**Other Research Topics and Application Domains:**

B1.2.2. - Cognitive science
B9.1.1. - E-learning, MOOC
B9.4.5. - Data science
B9.5.6. - Archeology, History
B9.5.8. - Linguistics
B9.5.10. - Digital humanities
B9.6. - Reproducibility
B9.7. - Knowledge dissemination
B9.7.1. - Open access
B9.7.2. - Open data

# 1. Personnel

**Research Scientists**

Benoît Sagot [Team leader, Inria, Researcher]
Pierre Boullier [Inria, Emeritus]
Laurent Romary [Inria, Senior Researcher, HDR]
Daniel Stökl Ben Ezra [EPHE, Senior Researcher]
Éric Villemonte de La Clergerie [Inria, Researcher]

**Faculty Members**

Marc Bui [Univ Vincennes-Saint Denis & EPHE, Professor]
Djamé Seddah [Univ Paris-Sorbonne, Associate Professor]

**PhD Students**

Jack Bowers [Vienna Academy of Sciences]
Loïc Grobol [Ministère de l'Education Nationale, from Oct 2017]
Axel Herold [Berlin-Brandenburg Academy of Sciences]
Mohamed Khemakhem [Inria]

Mathilde Regnault [Ecole Normale Supérieure Paris, from Oct 2017]

**Technical staff**

Wigdan Abbas Mekki Medeni [Inria, from Apr 2017]
Achraf Azhar [Inria, from Nov 2017]
Elias Benaissa [Inria, from Apr 2017]
Luca Foppiano [Inria]
Tanti Kristanti [Inria, from Nov 2017]
Alba Marina Malaga Sabogal [Inria, from Sep 2017]
Héctor Martínez Alonso [Inria, until Nov 2017]
Stefan Pernes [Inria, from Mar 2017 until Aug 2017]
Marie Puren [Inria]
Charles Riondet [Inria]
Dorian Seillier [Inria]
Lionel Tadonfouet [Inria, from May 2017]
Émilia Verzeni [Inria, from Apr 2017]

**Interns**

Raphael Avoustin [Inria, from May 2017 until Aug 2017]
Siham Feredj [Inria, from Apr 2017 until Sep 2017]
Florian Gouret [Inria, from Jun 2017 until Jul 2017]
Tanti Kristanti [Inria, from May 2017 until Oct 2017]
Yann-Alan Pilatte [Inria, from Jun 2017 until Jul 2017]
Maram Romdhane [Univ de Lorraine, from Apr 2017 until Sep 2017]
Roua Torjmen [Inria, from Apr 2017 until May 2017]
Julie Tytgat [Inria, from Jun 2017 until Aug 2017]

**Administrative Assistant**

Christelle Guiziou [Inria]

**Visiting Scientists**

Basant Agarwal [ERCIM, from Aug 2017 until Sep 2017]
Daniel Dakota [Indiana University, until Jan 2017]

# 2. Overall Objectives

## 2.1. Overall Objectives

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

ALMAnaCH's scientific positioning therefore extends ALPAGE's. We remain committed to developing state-of-the-art natural language processing 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 now reinforced and addressed in the broader context of computational humanities, with an emphasis on language evolution and, as a result, on ancient languages.

This new scientific orientation has motivated the creation of a new project-team with a new partner, namely the École Pratique des Hautes Études (EPHE). The EPHE is a leading institution in France in human sciences in general and in digital and computational humanities in particular. Two EPHE research directors, who have already been working together for some time in computational humanities, will be permanent members of the project-team: a philologist and a computer scientist, both specialists of computational approaches to philology and ancient language studies, in line with the above-mentioned scientific positioning.

# 3. Research Program

## 3.1. Overview and research strands

One of the main challenges in computational linguistics is **modelling and coping 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) and other dimensions (disabilities, for instance). But language is also in constant evolution 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 huge amounts of annotated data. They are still struggling with the high level of variability found for instance in **user-generated content** or in **ancient texts**.

ALMAnaCH tackles the challenge of language variation in two complementary directions.

### *3.1.1. Research strand 1*

We focus on linguistic representations that are less affected by language variation. This first requires improving the **production of semantic representations (semantic parsing)**. This also involves investigating 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 the field of natural language processing (hereafter NLP). We have to identify, model and take advantage of each type of contextual information available. Addressing these issues enables the development of new lines of research related to conversational content. Applications thereof include chatbot-based systems and improved information and knowledge extraction algorithms. We especially focus our work on challenging datasets such as domain-specific texts and historical documents, in the larger context of the development of digital humanities.

### *3.1.2. Research strand 2*

Language variation must be better understood and modelled in all its forms. In this regard, we put a strong emphasis on **three types** of language variation and their mutual interaction: **sociolinguistic variation** in synchrony (including non-canonical spelling and syntax in user-generated content), **complexity-based variation** in relation with language-related disabilities, and **diachronic variation** (computational exploration of language change and language history, with a focus on Old to all forms of Modern French, as well as Indo-European and Semitic languages in general). In addition, the noise introduced by OCR and HTR systems, especially in the context of historical documents, bears similarities with those brought by non-canonical input in user-generated content. This noise constitutes a more transverse kind of variation stemming from the way language is graphically encoded, which we call **language-encoding variation**. Dealing with diachronic and language-encoding variation, as well as their interaction, is the main motivations behind the creation of a joint project-team between Inria and EPHE.

### *3.1.3. Research strand 3*

These two first research strands rely on the availability of **language resources** (corpora, lexicons). The development of **raw corpora from original sources** is a domain of expertise of ALMAnaCH's EPHE members. The (manual, semi-automatic and automatic) development of **lexical resources** and **annotated corpora** is a domain of expertise of ALMAnaCH's Inria and Paris 4 members. This complementary expertise in language resource development (research strand 3) benefits to the whole team and beyond, and both feeds and benefits from the work of the other research strands.

## 3.2. Automatic Context-augmented Linguistic Analysis

This first research strand is centered 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: hybridation not only couples symbolic and statistical approaches, but neural approaches as well. AI applications are twofold, notwithstanding the impact of language variation (for which see the next strand): (i) information and knowledge extraction, whatever the type of input text (from financial documents to ancient, historical texts and from Twitter data to wikipedia) and (ii) chatbots and natural language generation. In many cases, our work on these AI applications is carried out in collaboration with industrial partners (for which cf. Section 8.1). The specificities and issues caused by language variation (a text in Old French, a contemporary financial document and tweets with a non-canonical spelling cannot be processed in the same way) are addressed in the next research strand.

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

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

Traditionally, a full natural language process (NLP) chain is organized 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 [55], [79] and one of the best for robust multilingual parsing as shown by our results in various shared tasks over the years [77], [28], [29]. 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 [28].

In particular, we continue to explore the hybridization 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 [1] and to Extrinsic Parsing Evaluation Shared Task [2].

Fundamentally, we want to build tool less sensitive to variation, more easily configurable, and self-adapting. Our short-terms goals is to explore techniques such multi-task learning (cite refs in soogard 2016-2017) to propose a joint model of tokenization, normalization, morphological analysis and syntactic analysis. We also explore adversarial learning, considering drastic variation we face in user generated content parsing and historical text processing, as noisy input that needs to be handled at training and decoding time.

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 challenge in NLP now: handling the context where a speech act takes place.

Indeed, there is a strong tendency in NLP to assume that each sentence is independent from both other sentences and its context of enunciation, in order 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

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

a complete representation of the post's context, and in turn its accurate machine translation. Even for the presumably simpler task of POS tagging dialogue sequences, the addition of context-based features (namely the speakers information and the dialogue moves) was beneficial [59]. In the case of UGC, working accross sentence boundaries was explored for instance, with limited success, by [54] for document-wise parsing and by [68] 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 point carried out 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 datasets, 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 metadata about some situation (such as discussions between gamers in relation with the state of the game) or multiple points of views (pictures and captions, movies and subtitles). Even if the relationship between a language production and its context is imprecise and indirect, it is still a valuable source of information, notwithstanding the need for less supervised machine learning techniques (cf. the use of LSTM neural networks by Google to automatically suggest replies to emails).

### *3.2.2. Information and knowledge extraction*

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

However, each specialised domain (economy, law, medicine…) exhibits its own set of concepts with associated terms. This is also true of communities (e.g. on social media), and it is even possible to find communities discussing the same topics (e.g. immigration) with very distinct vocabularies. Global LODs weakly related to language may be too general and not sufficient for a specific language variant. Following and extending previous work in ALPAGE, we put an emphasis on information acquisition from corpora, including error mining techniques in parsed corpora (to detect specific usages of a word that are missing in the resources used), 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 joint work with Patrice Lopez on named entity detection and linking in scientific texts. Through the co-supervision of a PhD funded by the LabEx EFL (on which see below), we are also involved in pronominal coreference resolution (finding

the referent of pronouns). Finally, we plan to continue working on deeper syntactic representations (as initiated with the Deep Sequoia Treebank), thus paving the way towards deeper, semantic representations. Such information is instrumental when looking for more precise and complete information about who does what, to whom, when and where in a document. These lines of research are motivated by the need to extract useful contextual information, but it is also worth noting their strong potential in industrial applications.

### *3.2.3. Chatbots and text generation*

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

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

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

## 3.3. Computational Modelling of Linguistic Variation

NLP and DH tools and resources are very often developed for contemporary, edited, non-specialised texts, often based on journalistic corpora. However, such corpora are not representative of the variety of existing textual data. As a result, the performance of most NLP systems decrease, sometimes dramatically, when faced with non-contemporary, non-edited or specialised texts. Despite the existence of domain-adaptation techniques and robust tools, for instance for trying to process social media texts, 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, and also regarding intertextuality, on which see below). 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 accesibility 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 sequencial nature of the problem at hand. We 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-resources languages (e.g. Old French texts), and OCR/HTR system outputs.

Nevertheless, the different types of language variation 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*

We plan to explore computational models to deal with language variation. But it is important to start by getting more insights about language in general and about the way humans apprehend it. We 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. In this regard, ALMAnaCH memebrs 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 (for which 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 is information about processing evolutions (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 on our daily life through social media and such 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 linguistics behaviours. In particular, social medias 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 standardized 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 unstandardized historical sources. To define appropriate tools, descriptions of these varieties are needed. Yet such descriptions need tools to be validated. We address this circularity interdisciplinarily, by working both on linguistics descriptions and on NLP tool development. Recently, sociodemographic variables have been shown to bear a strong impact on NLP processing tools. This is why, in a first step, jointly with researchers involved in the ANR project SoSweet (ENS Lyon and Inria's Dante), we 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 instance of 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 [62] uses such models to automatically generate forms attested in (classical) Indo-European languages but is based on a idiosyncrasic and inacceptable 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 [53] and [58]. Such models rely on a good understanding of the phonetic evolution of the languages at hand.

In ALMAnaCH, we focus 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 are in direct interaction with sub-strand 3b (development of lexical resources). We 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, 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 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 are to be paired with semantic resources (wordnets, word embeddings, and other corpus-based statistical information). This makes it possible 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 the other (e.g. for studying the non-inherited layers in the Ancient Greek lexicon). In the longer term, we intend to 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 rely on etymological datasets and standards for representing etymological information, for which see Section 3.4.2.

### *3.3.4. Accessibility-related variation*

Language variation does not always constitute an additional complexity in 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 [78]). 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 [63]. 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. [3]

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 [57], [73], our goal is 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 therefore investigate hybrid techniques involving rule-based, statistical and neural approaches based on parsing results (for an example of previous parsing-based work, see [51]). Lexical simplification, another aspect of text simplification [60], [64], is also to be investigted. [4]

[3] http://www.unapei.org/IMG/pdf/GuidePathways.pdf

[4] We have started a collaboration with Facebook's Parisian FAIR laboratory, 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-extraxction suite 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.

### *3.3.5. Intertextual variation*

Language variation is not restricted to language-internal dimensions such as the effects of sociolinguistic and diachronic factors. It also involves variation in the way a same content can be expressed. Detecting, analysing and qualifying this type of variation is a challenge that can be applied in different settings, such as the automatic study of intertextuality in ancient documents (different versions of a same myth, for instance), automatic comparison of documents dealing with the same facts and citations (e.g. journalistic articles and news wires), assessment of textual entailment, and automatic detection of plagiarism. In ALMAnaCH, we put an emphasis on the first two of these examples.

Intertextual comparison of close witnesses of the same text produces valuable data on orthographic, morphological or semantic equivalences and variance (textual criticism). Automatic parallel detection not only informs about the positive intertextuality between two sources (e.g. the use of Biblical quotations among Church Fathers or Rabbinic authors) but also reveal the differences in their use and transformation of the same textual material, and therefore the authorial strategies and politics.

In automatic language processing, it is customary to focus on similarities when dealing with distinct documents. Instead, we can focus on modelling what is idiosyncratic to a certain text, given a reference. This can allow, for instance, to identify whether an elided passage is relevant or not. Identifying such relevant omissions was one of the goals of the VerDi Project (on which see below).

## 3.4. Modelling and Development of Language Resources

### *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 is involved). It is therefore necessary to design a workflow that makes it possible to deal with data collections, even if they are initially available as photos, scans, wikipedia dumps, etc.

These challenges are particularly relevant when dealing with ancient languages or scripts where fonts, OCR techniques, language models may be not extant or of inferior quality, as a result, among others, of the variety of writing systems and the lack of textual data. This project-team will therefore work on improving print OCR for some of these languages for this very aim (e.g. Syriac, Ge'ez, Armenian). When an ancient source is still unpublished (book, manuscript, stele, tablet...), and therefore available in raw (image) form, we intend to develop OCR / HTR techniques, at least for certain scripts (Hebrew, Coptic and Greek Uncials, Ge'ez), and construct a pipeline for historical manuscripts. Initial success for Hebrew and Latin manuscripts has been very comforting (ca. 3% CER). On the one hand, access to existing electronic corpora, especially epigraphic corpora (e.g. Aramaic, North and South Arabic), still have to be negotiated. On the other hand, data that has been produced directly in electronic form (e.g. on social media) is readily usable, but far from normalised. Of course, contemporary texts can be often gathered in very large volumes, as we already do within the ANR project SoSweet, but this results in specific issues.

An inventory of already available resources developed or used by ALMAnaCH members has been developed.

The team pays a specific attention to the re-usability [5] of all resources produced and maintained within its various projects and research activities. To this end, we want to ensure maximum compatibilities with available

[5] From a larger point of view we intend to be conformant to the s-called FAIR principles (http://force11.org/group/fairgroup/fairprinciples)

international standards for representing textual sources and their annotations. More precisely we consider TEI 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 workflows for exploiting corpora, starting from a 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 process may be seen as progressivily enriching the documents with new layers of annotations produced by various NLP modules and possibility validated by users, preferably in a collaborative way. It relies on the use of clearly identified representation formats for the annotations, as advocated by ISO TC 37/SC 4 and TEI, but also on the existence of well-designed collaborative interfaces for browsing and validation. ALMAnaCH has been or is working on several of the NLP bricks needed for setting such a workflow, and has a solid expertise in the issues related to standardisation (of documents and annotations). However, putting all these elements in a unified workflow that is simple to deploy and configure remains to be done.

It should be noted that such workflows have also a large potential besides DH, for instance for valorising internal documentation (for a company) or exploring existing relationships between entities. [6]

### *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 [4], [1]). Such resources play a crucial role in all NLP tools, as has been proven among other tasks for POS tagging [71], [25], [29] and parsing, and some of the lexical resource development are targeted towards the improvement of NLP tools. They 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 are also one of the primary sources of linguistic information for augmenting language models used in OCR systems for ancient scripts, and allow us to develop automatic annotation tools (e.g. POS taggers) for low-resourced languages (see already [30]), especially ancient languages. Finally, semantic lexicons such as wordnets play a crucial role in assessing lexical similarity and automating etymological research.

Therefore, an important effort towards the development of new morphological lexicons is intended, with a focus on ancient languages of interest. Following previous work by ALMAnaCH members, we try and leverage all existing resources whenever possible such as electronic dictionaries, OCRised dictionaries, both modern and ancient [70], [19], [26], [27], while using and developing (semi)automatic lexical information extraction techniques based on existing corpora [69], [74]. A new line of research consists in the integration of the diachronic axis by linking lexicons that are in diachronic relation with the one another thanks to phonetic and morphological change laws (e.g. XIIth century French with XVth century French and contemporary French). Another novelty is 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 [19], [26], [27].

An underlying effort for this research is 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 dicitonaries such as the Dictionnaire Universel de Basnage (preparation of an ANR project with the University of Grenoble-Alpes and Paris Sorbonne Nouvelle).

[6] In this regard, we have started preliminary discussions with Fujitsu Lab and with the International Consortium of Investigative Journalists.

In the same way as we signalled the importance of standards for the representation of interoperable corpora and their annotations, we intend to keep making the best of the existing standardisation background for the representation of our various lexical resources. There again, the TEI guidelines play a central role, and we have recently participated in the "TEI Lex 0" initiative to provide a reference subset for the "Dictionary" chapter of the guidelines. We are also responsible, as project leader, of the edition of the new part 4 of the ISO standard 24613 (LMF - Lexical Markup Framework) dedicated to the definition of the TEI serialisation of the LMF model. [7] We consider that contributing to standards allows us to stabilize 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 are either be only morphosyntactic or 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 [76], [65], [75], [61]) and now participate to the creation of valuable resources originating from the historical domain genre.

# 4. Application Domains

## 4.1. Application domains of NLP and Computational Humanities

ALMAnaCH's research areas cover Natural Language Processing (nowadays recognised 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 include:

- Information extraction, information retreival, 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

- ALMAnaCH's submission to the 2017 CoNLL multilingual parsing shared task was ranked 3rd (out of 33) in part-of-speech tagging, and 6th (out of 33) in dependency parsing.
- Joint submissions of ALMAnaCH and Stanford University to the Extrinsic Parsing Evaluation campaign ranked 1st and 3rd.

# 6. New Software and Platforms

## 6.1. Enqi

- Author: Benoît Sagot
- Contact: Benoît Sagot

## 6.2. SYNTAX

KEYWORD: Parsing

[7] Defined in ISO 24613 part 1 (core model), 2 (Machine Readable Dictionaries) and 3 (Etymology).

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: Éric Villemonte 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. Crapbank

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

- Contact: Djamé Seddah

## 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: Éric Villemonte De La Clergerie
- Contact: Éric Villemonte 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 Éric Villemonte 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: Éric Villemonte De La Clergerie
- Contact: Éric Villemonte 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. Standardisation of Natural Language data

**Participants:** Loïc Grobol, Laurent Romary, Stefan Pernes, Jack Bowers, Charles Riondet, Mohamed Khemakhem.

One essential aspect of working with human traces as they occur in digital humanities at large and in natural language processing in particular, is to be able to re-use any kind of primary content and further enrichments thereof. The central aspect of re-using such content is the development and applications of reference standards that reflect the best state of the art in the corresponding domains. In this respect, our team is particularly attentive to the existing standardisation background when both producing language resources or devlopping NLP components. Furthermore, our specific leading roles in the domain of standardisation in both the Parthenos [41] and EHRI [40] projects as well as in related initiatives (TEI consortium, ISO committee TC 37, COST action ENeL (European Network in e-Lexicography), DARIAH lexical working group) has allowed to make progress along the following lines:

- Contribution to the improvement of the TEI guidelines [15], [20] and in particular to the definition of an extension for stand-off annotation in the continuity of [52] [8]
- Editing an ISO standard on the annotation of reference phenomena in discourse [9] that intends to be feature complete from a linguistic point of view (from simple co-reference to complex bridging anaphora phenomena) and compliant with the TEI stand-off annotation module from the point of view of its implementation [18]
- Editing the draft for the future project ISO 24613-4, which, on the basis of the proposals made in [67], intends to provide a reference TEI based serialisation for the LMF model (comprising core model (ISO 24613-1), machine readable dictionary (ISO 24613-2) and etymology (ISO 24613-3, cf. below) modules). This work is also the basis for the output format of Grobid-dictionary [19]
- Editing the draft for the future project ISO 24613-4, which will provide the model for representing etymological information in dictionaries and lexical resources, on the basis of [11]. Preliminary experiments have been carried out in [26], [27] (see also section 7.10)
- Proposal of a modular specification of the TBX standard (ISO 30642) by means of a TEI ODD specification [24]
- Participation to a call for contribution to the future evolution of the archival standard EAC-CPF (Encoded Archival Context for Corporate Bodies, Persons, and Families), proposing to use the TEI ODD specification language [47]

## 7.2. Digital Humanities and Cultural Heritage

**Participants:** Stefan Pernes, Marie Puren, Charles Riondet, Laurent Romary, Dorian Seillier, Lionel Tadonfouet.

[8] https://github.com/laurentromary/stdfSpec
[9] https://www.iso.org/standard/69658.html

The very broad scope of Digital Humanities and Cultural Heritage is well represented in the latest works of the ALMAnaCH team, undertaken in various contexts (European and national research infrastructures and bilateral partnerships). However, the issues tackled always deal with interoperability, reusability and standardization:

- The "Data Reuse Charter"[33] project is carried by a large consortium of European infrastructures and institutions
- The “Standardization Survival Kit” (or SSK) [66] developed within the PARTHENOS project intends to show that proper data modelling and corresponding standards make digital content more sustainable and reusable. Arts and Humanities would be well-suited to taking up the technological prerequisites of standardization [41], as most technological domains have already done.
- A concrete application of what offers the SSK has been developed within the EHRI project, where we built a methodology for the management of heterogeneous archival sources—expressed in the EAD Encoded archival description format—in one single environment, namely a federated portal [40], [48]. This method is based on a specification and customisation method inspired from the TEI, i.e. the definition of project-specific subsets of the standard and the maintenance of both technical and editorial specifications within a single framework.
- the Time-US project aims to reconstruct the remuneration and time budgets of women and men working in the textile trades in four French industrial regions (Lille, Paris, Lyon, Marseille) in a long-term perspective. During the launch phase, the team has been active in the following domains:
  - Collection of primary sources. The Time-Us team works on a heterogenous corpus of French handwritten and printed sources spanning from the seventeenth to the twentieth century; it includes court decisions, petitions, police reports and files, and sociological surveys on living conditions of the working class.
  - Evaluation of technical solutions for image visualization, transcription and collaboration, such as Transkribus (https://transkribus.eu/Transkribus/). The Transkribus interface enables Humanities scholars to transcribe handwritten and printed historical sources, and offers a very powerful Handwritten Text Recognition engine.
  - Creation of an annotation schema in XML/TEI. As the corpus gathers together diverse historical sources, the definition of a light and flexible annotation schema is a major step to create data to train parsing models. This data take the form of annotated texts encoded in TEI (Text Encoding Initiative). The annotation process starts as a collaborative effort, in order to get a first dataset that will later be used to train and configure NLP tools. The current step also helps designing a precise annotation guide between the NLP people and historians, in particular to clarify their expectations.
  - Installation of a customized MediaWiki. Several digital projects have already taken into account the specific needs of historians in terms of image visualization, transcription and collaboration. But they do not address all the requirements of Humanities scholars working on primary sources, and the need of comprehensive Digital Humanities-based publishing systems is emerging. We have chosen to setup a specific digital workflow enabling historians and NLP experts to work together, namely a wiki under Mediawiki (http://timeusage.paris.inria.fr/mediawiki/index.php/Accueil) with the Transcribe Bentham transcription desk, adapted to our needs, and a TEI toolbar, specifically customized for tagging named entities and measures.
- Archives nationales
- In a complex of projects (eRabbinica, LAKME, NEH/DFG Mishna-Tosefta Synopsis) with different partners dealing with classical rabbinic literature in Middle Hebrew we thrive to create a critical edition with translation, linguistic annotation and lexicon of the Mishna (200k tokens, the hypotext of the Talmud). Hebrew, a script written from right to left and a highly agglutinative language, poses great challenges to encoding standards and demands the development of new technical solutions. No open source corpora exist for linguistically annotated texts in rabbinic Hebrew.

- Building on ocropus HTR capacities, we have added our own layout analysis algorithms for column and line segmentation [35] that have proven very succesful for literary manuscripts for the tasks of aligning existing transcriptions of manuscripts with the word and character ROIs and for new transcriptions reaching similar results to transkribus but with a much easier complete control of the layout analysis.
- With our partners at the University of Maryland we have produced a preliminary TEI transcription of the most important manuscript Kaufman A50 (https://raw.githubusercontent.com/umd-mith/mishnah/master/data/tei/S07326.xml). Further improvements are currently undertaken. We have been able to use this transcription to realign it with the manuscript glyphs.
- We have produced preliminary transcriptions of two further manuscripts (Cambridge 450.2 and Parma A) that are in the process of TEIzation. A fourth manuscript (Munich Cod. Ebr. 95) is currently in treatment.
- Our partners at Dicta, have produced a preliminary automatical linguistic annotation of a vulgate text of the Mishna with HMMs with data for lemma, POS and morphological analysis. In the LAKME project, we have now manually corrected 25k tokens (ca. 12 percent of the whole text) that will be used to train RNN to improve the current transcription of the remaining text and enter a human-machine dialogue to fully annotate the whole Mishna. The annotation will not only be the first open source annotation. It will also be considerably more detailed than the excellent but closed annotation of the Israel Academy of the Hebrew Language (http://maagarim.hebrew-academy.org.il/). The resulting system will enable us to annotate other texts such as Tosefta and Halakhic Midrashim for the upcoming Sofer Mahir (tachygraph) project.

## 7.3. Computational Humanities and ancient texts

**Participants:** Daniel Stökl Ben Ezra, Marc Bui.

In collaboration with Jérémie Bosom and Dogu Kaan Eraslan (PhD students (co-)supervised by Marc Bui at EPHE).

Ancient languages of interest: ancient Egyptian (hieroglyphics, hieratic, demotic) , ancient Greek, Aramaic, Elamite, biblical Hebrew, classical Arabic, Hán Nôm (ancient vietnamese), old Persian

Computational approaches in humanities makes it possible to address the problems encountered by philologists such as reading, analyze and archiving old texts in a systematic way. We based our research on algorithms, their implementations, and human expertise on ancient languages to automate these difficult tasks.

The research scope of 2017 was the work around historical document or manuscripts available in images. Our work program (or work in progress) includes:

- Document layout analysis for ancient manuscripts using computer vision techniques and machine learning
- Script identification taking into account the environment where the trace is located: image, artefact, noise due to deterioration of the medium of writing. By stacking auto-encoding neural networks in order our approach provides an alternative representation of the input data received.
- Text recognition (handwritten text recognition) by enhancing it with LSTM
- Palaeographic classification of manuscripts and ancient inscriptions. Classification of historical document images can be addressed through script identification, in that case, our proposed method is based on the use of Convolutional Auto-Encoders (CAE) stacked in several layers in order to obtain fine-grained features and automatically learn representations of the line of writing or drawing of script
- Cross language Information Retrieval and Information Retrieval applied to ancient languages.

## 7.4. Information Extraction with GROBID

**Participants:** Luca Foppiano, Mohamed Khemakhem, Laurent Romary.

GROBID is an open source software suite initiated in 2007 by Patrice Lopez with the purpose of extracting metadata automatically from scholarly papers available in PDF. Over the years, it has developped into a rich information extration environment, and deployed in many Inria projects, but also national and international services, among which we can quote HAL. It is a central piece for our information extraction activities and we have been particularly active in 2017 in the following domains:

- General contributions to GROBID (https://github.com/kermitt2/grobid):
  - Major refactoring and design improvements
  - fixes, tests, documentation and update of the pdf2xml fork for Windows
  - added and improved several models in collaboration with CERN (e.g. for the recognition of arXiv identifier)
- Contribution to entity-fishing (https://github.com/kermitt2/nerd):
  - integration into the main open-access platform: EKT/OMP, OAPEN, OpenEdition, Gottinghen University Library Press, Ubiquity press
  - deployment in the DARIAH infrastructure via Huma-NUM
  - adding supported languages for Italian and Spanish
  - various fixes and refactoring
  - Creation of a specific client for Historical documents, combined with a POS-tagger that connect the found entities between them and with their structural context[34]
- Contribution to GROBID-Dictionaries [10]: the lexical GROBID extension has been implemented and tested on modern and multilingual dictionaries [19]. The architecture has been further developed and an extension for etymology has been plugged-in on the top of the existing models. First experiments on etymological samples have been carried out and more work is required on the features selection. In parallel, the output of the system is actively synchronised with the Standardisation initiatives such as TEI Lex0 and ISO 24613 (LMF). Usability has been enhanced as well by lightening the annotation process and simplifying the setup process of the tool. Such measures are going to unlock the workforce potential of different interested research partners to generate more annotated data required for feature engineering. A first user experiment has been carried out during a dedicated workshop at the Lexical Masterclass, where the new features have been tested

## 7.5. Multilingual POS-tagging and Parsing

**Participants:** Éric Villemonte de La Clergerie, Djamé Seddah, Benoît Sagot, Héctor Martínez Alonso.

Our participation in 2017 to two international shared tasks (CONLL UD and EPE—the latter in collaboration with Stanford University) led us to develop a new generation of statistical multilingual NLP tools, in particular for POS-tagging and for Parsing [29]. In particular, the CoNLL shared task involved 80+ datasets covering 50+ languages (including low-resource and no-resource languages) and, for some languages, various genres.

For POS tagging, we have developed a new feature-based POS tagger, following our previous work on MElt [56], [72]. This new tagger, named alVWTagger, uses the Vowpal Wabbit system for training linear POS models, resulting in an important drop in training times. This has allowed us to better explore the feature set space based on development data for each and numerous ways to encode the information provided by external morphological lexicons, resulting in better tagging results. We also developed a derivative of this tagger for performing tokenisation and sentence segmentation. Experiments on the development sets of the CoNLL shared task allowed us to chose the best setting for each corpus between several configurations, by using the UDPipe baseline (provided by the shared task organisers) or alVWtagger for each of the 3 subtasks

[10] https://github.com/MedKhem/grobid-dictionaries

(tokenisation, segmentation in sentences, UPOS tagging). As a result, we ranked 3rd (out of 33 participants) in the UPOS tagging ranking of the CoNLL shared task, and 5th for the tokenisation subtask and 6th for the sentence segmentation substask. Moreover, later improvements in the parsing models resulted in alVWtagger being more often used than for the official run, with improved results (unofficial post-campaign ranking on UPOS tagging: 2nd/33).

In parallel, we have developed a neural POS tagger based on Barbara Plank's LSTM tagger, by exploring the impact of integrating lexical information extracted from morphological lexicons within the neural architecture. We showed that such information improves POS tagging on average [25]. A careful comparison of this neural tagger, alNNtagger, w.r.t. alVWtagger is yet to be carried out, but preliminary experiments tend to show that both taggers perform similarly on average. This is likely because POS tagging is a relatively easy task for which the manual design of adequate features is relatively easy. As a result, using a neural architecture, which has the advantage of learning the optimal features rather than relying on manually crafted ones, does not result in massive improvements as observed in many other NLP tasks and beyond.

For Parsing, DyALog-SR, a feature-based parser on top of DyALog system, was extended (into DyALog-SRNN) to integrate predictions proposed by deep neuronal layers, based on a global char LSTM and a word bi-LSTM. Based on the results of the CONLL UD shared task, further extensions were added to DyALog-SRNN, namely an adaptation of Stanford's winner system (based on a bi-affine prediction of word governors) and a version of the Maximum-Spanning Tree (MST) algorithm, allowing us to move from the 6th place (for parsing) to an unofficial post-campaign 4th place.

The new version DyALog-SRNN has preserved the functionality of DyALog-SR to produce (deep) dependency graphs rather than standard shallow dependency trees. This functionality was used during the EPE (Extrinsic Parsing Evaluation) shared task to test several dependency tree and graph representations for several downstream application tasks [28].

The goal of that collaboration with the Stanford NLP team was to evaluate the usability of several representations derived from English Universal Dependencies (UD), as well as the Stanford Dependencies (SD), Predicate Argument Structure (PAS), and DM representations. We further compared two parsing strategies: Directly parsing to graph-based dependency representations and a two-stage process of first parsing to surface syntax trees and then applying rule-based augmentations to obtain the final graphs. Our systems used advanced deep learning techniques on top of state-of-the-art preprocessing and part-of-speech tagging. Overall, our systems performed very well and our results were ranked first and third on that shared task (over more than 20 submitted systems). The main advantage of that shared task was to provide an extrinsic evaluation scenario which consisted in extracting relevant information for information retrieval from speech and biomedical data, as well as opinion mining.This showed the relevance of our approach and the interest of producing graph-based representations to downstream applications that were developed for tree-based structures.

In particular, it showed the interest of deeper syntactic representation instead of shallow ones. In parallel with these efforts, work was also carried out on the issues related to polylexical units in parsing [17]. Moreover, the *International Journal of Lexicography* has accepted a paper written in collaboration with three other European research centres on the interactions between NLP and lexicography on polylexical units (to appear in 2018).

## 7.6. Tweet processing

### *7.6.1.*

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

In the context of the SoSweet and Parsiti ANR actions, we run various experiments on large amounts of tweets.

In a first experiment, around 20 millions tweets were normalized, and then parsed with FRMG. 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 FTB journalistic texts), also leading to high parsing times because of correction strategies. However, error mining was tried to identify a first set of easy errors and further developments are planned to track errors more related to segmentation and

normalization. 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.

In a second experiment, we adapted our two clustering (DepCluster) and word embeddings (DepGlove) algorithms to take into account non-linguistic relations, such as the author-word relation (between an author and the words of her tweets). The algorithms were applied on raw tweets with only a basic tokenisation, and results produced on a month basis over 18 months (2016/02 to 2017/08). Several tools, with a special focus on Cytoscape, were tried to visualize the results as networks, in order to identify and explain communities.

## 7.7. Syntax modelling and treebank development

**Participants:** Djamé Seddah, Héctor Martínez Alonso, Benoît Sagot, Elias Benaissa, Wigdan Abbas Mekki Medeni, Émilia Verzeni.

In 2017, ALMAnaCH members have contributed to the *Universal Dependency* initiative [44]:

- Héctor Martínez Alonso has resumed his contribution to the *Universal Dependencies* (UD) initiative, with annotations and data evaluations for Catalan, Danish and Spanish datasets.
- Several ALMAnaCH members have worked on converting the French TreeBank into the UD model and format (paper to be presented in 2018) and on the automatic identification of syntactic structures in UD.

As part of the ANR Parsiti project (2016-2020), whose goal is to build the next generation of context-enhanced NLP tools, we are currently developing a parallel data set of user-generated content language pairs, French-English and North-African dialect Arabic-French. Each of those pairs contains highly non-canonical text, heavily contextualized. We built the translation pairs and are currently carrying out annotations at the morpho-syntactic level. None of these data set already exist, they will be first used for the evaluation of our current processing chains and then to bootstrap state-of-the-art models as part of their training data. 3 annotators are involved over a year long period (18 man.month, end in June 2018).

## 7.8. Context-Enhanced NLP tools building

**Participants:** Djamé Seddah, Julie Tytgat, Florian Gouret, Yann-Alan Pilatte.

The ANR Parsiti project also aims 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. These information can be of spatial temporal nature for example, and have been shown to improve Entity Linking over social media streams [11]. In our case, we decided to focus on a closed world scenario in order to study context and speech acts interaction. We built 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 organized and enable the modelling of the extra-linguistics context with different level of granularity. Recorded over many games sessions, we 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.

## 7.9. Quantitative and computational morphology

**Participant:** Benoît Sagot.

[11] fang2014entity

In 2017 we have resumed our work on empirical and computational morphology, although at a slower pace than during the previous years. Apart from the preparation of an issue of the *Morphology* journal on computational morphology as a guest editor, together with Olivier Bonami (LLF) [10], our work in this regard was threefold:

- Contribution to the development of a morphological lexicon, a small-scale POS-annotated corpus and a POS tagger (based on MElt) for Romansh Tuatschin, a variety of the Sursilvan dialect of Romansh (a Romance language spoken in Switzerland); this work is a collaboration with Géraldine Walther and Claudia Cathomas (University of Zurich) [30];
- Formal and quantitative work on the verbal morphological system of Khaling, a Kiranti (Sino-Tibetan) language from Nepal, following earlier work of ours [80], [81]; this is a collaboration with Géraldine Walther (University of Zurich) and Guillaume Jacques (CRLAO, CNRS);
- Preliminary work on the diachronic modelling of lexical information at the morphological and phonetic levels.

## 7.10. Creation, Extraction and Standardisation of Etymological Information

**Participants:** Jack Bowers, Mohamed Khemakhem, Laurent Romary, Benoît Sagot.

A new, important line of research in 2017 was the work around etymological information and resources. This work can be divided into three main dimensions:

- Standards for the representation of etymological information.
- Extraction of etymological resources from existing datasets. Two main resource types were exploited:
  - Digitalised legacy etymological dictionaries, using GROBID-dictionaries, in collaboration with the Berlin-Brandenburg Academy of Sciences. The output of the process is a TEI-structured dictionary (see module 7.4 for more details).
  - The English Wiktionary, from which structured, formalised etymological information was extracted and published (open-source) in the form of a database of lexemes (i.e. language/lemma/meaning triples) and an associated database of etymological relations (input lexeme(s)/output lexeme/type of relation) [26], [27].
- Etymological research (i.e. producing novel etymological hypotheses), in collaboration with Romain Garnier (Université de Limoges & Institut Universitaire de France) and, although to a lesser extent, Laurent Sagart (CRLAO, CNRS) [12], [37]. Although limited (for now), the contribution of computational models in our research is real; it allowed us to check the validity of the diachronic phonetic evolution model we have postulated for a new, hypothetical Indo-European language we suggest could have served as a source of borrowings for the ancestors of both Greek and Italic languages [12].

## 7.11. Automatic Detection of Coreference

**Participants:** Éric Villemonte de La Clergerie, Loïc Grobol.

In 2017, ALMAnaCH members have investigated coreference detection for French using machine learning and existing linguistic knowledge. Our efforts consisted in using insight gathered from deep and shallow parsers and standard machine learning approaches to detect entity mentions [31], adapting knowledge-poor deep-learning techniques for end-to-end coreference resolution to the case of oral French and researching new ways of exploiting structured such as parse trees in deep neural models.

## 7.12. Detecting omissions in journalistic texts

**Participants:** Héctor Martínez Alonso, Benoît Sagot.

In the journalistic genre that is characteristic of online news, editors make frequent use of citations as prominent information; yet these citations are not always in full. The reasons for leaving information out are often motivated by the political leaning of the news platform.

Existing approaches to the detection of political bias rely on bag-of-words models that examine the words present in the writings. In the context of the VerDI project (see below), we have resumed our work aimed at going beyond such approaches, which focus on what is said, by instead focusing on what is *ommited*. Thus, this method requires a pair of statements; an original one, and a shortened version with some deleted words or spans. The task is then to determine whether the information left out in the second statement conveys *substantial* additional information. If so, we consider that a certain statement pair presents an omission. To tackle this question, we used a supervised classification framework, for which we require a dataset of sentence pairs, each pair manually annotated for omission.

We had developed last year a small reference corpus for evaluation purposes, using and comparing both crowd and expert annotation. This corpus has allowed us to examine which features help automaticallly identify cases of omission. In 2017, we have finalized the annotation tools for the VerDI project [23], and published them online as free software (see below).

# 8. Bilateral Contracts and Grants with Industry

## 8.1. Industrial Collaborations

- **Verbatim Analysis**: this Inria start-up was co-created in 2009 by BS. It uses some of ALPAGE/ALMAnaCH's free NLP software (SxPipe) as well as a data mining solution co-developed by BS, VERA, for processing employee surveys with a focus on answers to open-ended questions. A new Inria startup, **opensquare**, was co-created in December 2016 by BS with 2 senior specialists of HR 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.
- **Facebook**: A collaboration on text simplification ("français Facile À Lire et à Comprendre", FALC) is starting with Facebook's Parisian FAIR laboratory. It should start with 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 (the CIFRE application has just been submitted). This collaboration is expected to be part of a larger initiative involving (at least) these three partners as well as the relevant ministries.
- **Bluenove**: A contract with this company has been signed, which initiates a collaboration in the integration of NLP tools (e.g. chatbot-related modules) within Bluenove's plateform Assembl, dedicated to online citizen debating forums. It involves a total of 24 months of fixed-term contracts (12 months for an engineer and 12 months for a research ingineer).
- **Science Miner**: ALMAnaCH (following ALPAGE) has been collaborating since 2014 years with this company founded by Patrice Lopez, a specialist in machine learning techniques and initiator of the GROBID and NERD (now entity-fishing) suites. Patrice Lopez provides scientific support on the corresponding software components in the context of the Parthenos, EHRI and Iperion projects, as well as in the context of the Inria anHALytics initiative, aiming at providing a scholarly dashboard on the scientific papers available from the HAL national publication repository.
- **Konverso**: A collaboration with this start-up is starting, focused on chatbots and text generation. One of our objectives with this collaborations is to initiate a larger initiative involving ALMAnaCH and several small companies, whose goal will be the development of open-source, NLP-enhanced chatbot modules. This is because such developments are complex and would benefit from such a mutualisation initiative. In turn, an open-source chatbot engine would allow startups and ALMAnaCH to more rapidly develop and deploy high-performance application-specific chatbots. The first concrete outcome of this collaboration is our joint submission to the call for projects published by the DILA (French government agency) for exploring the relevance of deploying a chatbot on the public information plateform service-public.fr.

- There exists at least one formal collaboration between a company and EPHE involving future ALMAnaCH members. It involves **Insight-Signals**, an EPHE start-up that "designs data analytics and decision support systems that integrate the complexity of humans' behaviour and their interactions".
- **Trooclick**: A direct and active collaboration with this company is now strengthened by the "RAPID" ANR project VerDI on the automatic detection of omissions in news reports and other types of texts. This project will come to an end in February 2018.
- ALMAnaCH members have recently initiated discussions with other companies (Fujitsu, HyperLex, Fortia Financial Solutions...), so that additional collaborations might start in the near future. They have also presented their work to companies interested in knowing more about the activities of Inria Paris in AI and NLP (Google, Toyota, Samsung...).

# 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

### *9.1.2. Competitivity Clusters*

- **LabEx EFL** (2010-2019, PI Christian Puech [HTL, Paris 3], Sorbonne Paris Cité). Topic: empirical foundations of linguistics, including computational linguistics and natural language processing. ALPAGE was one of the partner teams of this LabEx, which gathers a dozen of teams within and around Paris whose research interests include one aspects of linguistics or more. BS serves as deputy head (and former head) of one of the scientific strands of the LabEx, namely strand 6 dedicated to language resources. BS and DS are in charge of a number of scientific "operations" within strands 6, 5 ("computational semantic analysis") and 2 ("experimental grammar"). BS, EVdLC and DS are now individual members of the LabEx EFL since 1st January 2017, and BS still serves as the deputy head of strand 6. Main collaborations are on language resource development (strands 5 and 6), syntactic and semantic parsing (strand 5, especially with LIPN [CNRS and U.Paris 13]) and computational morphology (strands 2 and 6, especially with CRLAO [CNRS and Inalco]).
- **PSL project LAKME** (2015-2017, PI Thierry Poibeau [LATTICE]). Topic: language resource development for morphologically rich languages, especially Rabbinic Hebrew (syntactic level), Medieval French (morphological level) and some Finno-Ugric languages (to a lesser extent).

- **PSL Iris project SCRIPTA** This project emanates from the history and philology department of the EPHE (DSBE). It is directed by Andreas Stauder (EPHE) with Philip Huyse (EPHE) and Charlotte Schmid (EFEO). It unites the forces of a great number of researchers in PSL (EPHE, ENS, EHESS, ENC, Collège de France and in addition the IRHT) working on written texts in all its forms, on all kinds of material, from all periods and regions and has important digital and computational ambitions especially with regard to epigraphy, palaeography, digital editions and NLP.

### *9.1.3. Other National Initiatives*

- **TGIR Huma-Num** ALPAGE was a member of the CORLI consortium on "corpora, languages and interactions" (BS is a member of the consortium's board), and ALMAnaCH is in the process of joining this consortium. With a joint funding of Huma-Num and the H2020 project Parthenos (on which see below), ALMAnaCH members have also co-organised a workshop on 3D techniques for Humanities in Bordeaux (December 2016).
- **Institut de Linguistique Française (ILF)**: ALPAGE was a member of this CNRS "federation". ALMAnaCH is in the process of joining this federation if possible, especially as BS is the scientific head of the "Corpus de Référence du Français" initiative, an ILF project whose other head is Franck Neveu and whose goal is to develop a French National Corpus, a resource that has been awaited for a long time.
- **Notary registers project** (2017-2018): An explorative study has been launched in collaboration with the National Archives in France, in the context of the framework agreement between Inria and the Ministry of Culture, to explore the possibility of extracting various components from gigitized 19th Century notary registers.
- **Nénufar (DGLFLF - Délégation générale à la langue française et aux langues de France)**: The projects is intended to digitize and exploit the early editions (beginning of the 20th Century) of the Petit Larousse dictionary. ALMAnaCH is involve to contribute to the automatic extraction of the dictionary content by means of GROBID-dictionaries and define a TEI compliant interchange format for all results.
- **PIA Opaline**: The objective of the project is to provide a better access to published French literature and reference material for visually impaired persons. Financed by the Programme d'Investissement d'Avenir, it will integrate technologies related to document analysis and re-publishing, textual content enrichment and dedicated presentational interfaces. Inria participate to deploy the GROBID tool suite for the automatic structuring of content from books available as plain PDF files.

## 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 is president of the board of director) 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*

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

- University of Ljubljana (Darja Fišer) [wordnet development]
- University of Zürich, Switzerland (Géraldine Walther) [computational morphology, lexicons]
- Academy of Sciences, Berlin, Germany (Karl-Heinz Moerth) [lexicology]
- University of Fribourg, Switzerland [historical document analysis]
- University of Valencia, Spain [historical document analysis]
- University of Groningen, Netherlands [historical document analysis]
- University of Innsbruck, Austria [historical document analysis]

## 9.3. International Initiatives

### *9.3.1. International Partners*

- **ANR-NSF project MCM-NL** (2016-2020, PI John Hale [Cornell University, USA], resp. for Inria Paris / ALMAnaCH: EVdLC) 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. Informal collaborations:
- **The SPMRL initiative** (Statistical Parsing of Morphologically Rich Languages): a worldwide network of internationally renowned teams that was initiated during the IWPT'09 conference ALPAGE organised in Paris, DS playing a leading role since then. Other institutions involved include the University of Heidelberg (Germany), Bar Ilan University (Israel), Potsdam University (Germany) and Indiana University (USA). The outcomes of this initiative include the successful SPMRL Workshop and Shared Task series hosted successively by NAACL-HLT (2010), IWPT (2011), ACL (2012), EMNLP (2013), CoLing (2014) and IWPT (2015), in which DS as well as other ALPAGE/ALMAnaCH members played an active role. DS also served as a co-editor of a special issue of Computational Linguistics on this topic.
- **Sofer Mahir ("fast scribe") project**. Joint work on the computational processing of Rabbinic Hebrew manuscripts involving DSBE: Nachum Dershowitz (Tel Aviv University, Israel), Moshe Koppel (DICTA, Bar Ilan University, Israel), Meni Adler (DICTA, Ben Gurion University, Israel), Michael Elhadad (Ben Gurion University, Israel) on the NLP side and Hayim Lapin (University of Maryland, USA), Tal Ilan (FU Berlin, Germany) Shamma Friedmann (Bar Ilan University, Israel) on morphological analysis of Rabbinic Hebrew, alignment of manuscript witnesses (textual criticism), finding parallels, aligning related but different texts (like the Gospels). This work is also connected to the LAKME project mentioned above.

## 9.4. International Research Visitors

### *9.4.1. Visits of International Scientists*

- Daniel Dakota (Indiana University, 4 months, until Jan 2017)
- Theresa Lynn (Dublin City University, 10 days in January 2017)
- Amir More (Open University of Israel, 10 days in April 2017)

*9.4.1.1. Internships*

- Basant Agarwal (ERCIM, Aug-Sep 2017)

# 10. Dissemination

## 10.1. Promoting Scientific Activities

### *10.1.1. Scientific Events Organisation*

*10.1.1.1. General Chair, Scientific Chair*

- DSBE: Co-Chair and coorganizer of the summerschool manuSciences '17, Fréjus, France, 10-15 September 2017.
- LR, DSBE: Coorganizers of the France/Israel DH conference, Jerusalem, 27 February-1 March 2017.
- DSBE: Co-chair and coorganizer of the workshop Cambridge-PSL-Lausanne, Paris, France, 11-12 March 2017.
- DSBE: Co-chair and coorganizer of the workshop FSP-Patrima-EPHE on Scientific Approaches to Inscribed Objects, Ministry of Culture and Communication, Paris, France, 24 January, 2017.

*10.1.1.2. Member of the Organizing Committees*

- LR, BS: Members of the Organizing Committee of the Lexical Data Masterclass, Berlin, 4-8 December 2017
- CR: Member of the Organizing Committee of the conference "La Part de l'ombre, Action clandestine et imaginaire du complot, XXe-XXIe siècles", Paris, France, 18-19 May 2017
- MP: Member of the Organizing Committee of the conference Text as a resource. Text mining in Historical Science, Paris, France, 29-30 June 2017.
- MP: Member of the Organizing Committee of the masterclass Penser/ Utiliser les données de la recherche, Paris, France, 25-29 September 2017.
- DSBE: Member of Organizing Committee for the Poster Session on Digital Humanities at the World Congress of Jewish Studies, Jerusalem, 6-10 August 2017.

### *10.1.2. Scientific Events Selection*

*10.1.2.1. Chair of Conference Program Committees*

- MB, Program chair of the 8th ACM-SoICT, 2017, Nha Trang, Vietnam.

*10.1.2.2. Member of the Conference Program/Scientific/Reviewing Committee*

- LR: Member of the Reviewing Committee of the following conferences: DHd 2017, DATeCH2017, ACL 2017, TOTh 2017, TPDL 2017, MDQual 2017
- EVdLC: Member of the Reviewing Committee of ACL'17 (Tagging, Chunking, Syntax and Parsing area), DaTeCH'17 (co-reviewer), TMPA-2017 (co-reviewer)
- EVdLC: Member of the Scientific, Programm or Reviewing Committee of EMNLP'17, DepLing'17, EPIA'17, LATA'18, Games4NLP'17, ToTH'17, LREC'18.

- DS: Member of the Scientific, Programm or Reviewing Committee of ACL'17, EMNLP'17, CoNLL'17, TALN2017, LREC2018, TLT #15, LAW 2017, W-NUT 2017
- DSBE: Member of the Program Committeee for HIP'17, member of the Reviewing Committee of HIP 2017
- BS: Member of the Program, Scientific or Reviewing Committe of the following conferences and workshops: *SEM 2017, CoNLL 2017, DeriMo 2017, EACL 2017, LAW XI (2017), LREC 2018, WRDTM 2017
- CR: Member of the Reviewing Committee of DH Nord and DH 2018
- MP: Member of the Reviewing Committee of the following conferences: DH Nord, Digital Humanities in the Nordic Countries, DH 2018

### *10.1.3. Journal*

*10.1.3.1. Member of the Editorial Boards*

- LR: Member of the Editorial Board of the following journals: *JDMDH*, *ACM JOCCH*
- LR: Member of the Scientific/Advisory Committee of the following journals: *Revue Humanités Numériques*, Journal of the TEI, *Language Resources and Evaluation*
- DSBE: Member of the Advisory Committee of the following journals: *Jewish Studies Quarterly*, *Henoch*, *Biblical Annals*, *Digital Paleography and Book History*, *Cultural Heritage Digitization*, *Marginalia*
- DSBE: Member of the Advisory Committee of the book series "Humanités numériques et patrimoine" (U. Grenoble, E. Pierazzo)
- BS, together with Olivier Bonami (LLF), was Guest Editor for a thematic issue of the journal *Morphology* on computational methods for descriptive and theoretical morphology [10].

*10.1.3.2. Reviewer - Reviewing Activities*

- BS: Reviewer for the following journals: *Computational Linguistics*, *Journal of Language Modelling*, *Arabian Journal for Science and Engineering*, *wékʷos*
- BS: Reviewer for the MIT Press
- DS: Computational Linguistics Journal, Transactions on Asian and Low-Resource Language Information Processing Journal

### *10.1.4. Invited Talks*

- Héctor Martínez Alonso gave a talk at the Inria Paris Junior Seminar, Paris, France
- CR, "Traces de l'héroïsme, les plaques commémoratives de la Résistance parisienne", Plaques commémoratives dans l'espace parisien, Paris, France
- LF, CR, "Grobid for Humanities, when engineering meets History", DHIHA, Paris, France
- CR, "La Libération du Sud-Est parisien, un mouvement populaire, CLIO94, Créteil, France
- CR, "Traces de l'héroïsme, les plaques commémoratives de la Résistance parisienne", Plaques commémoratives dans l'espace parisien, Paris, France
- MP, "Fcailiter l'accès des chercheurs aux données patrimoniales. La Charte de réutilisation des données", séminaires Nouveaux champs d'étude en droit du patrimpine culturel, Le Mans, France, 14 December 2017.
- MP, "La TEI et les historiens. Le projet Time-Us : Travail, rémunération, textile et foyer (XVIIe-XXe siècles)", Edition électronique & TEI : enjeux, pratiques & perpectives, 4ème journée dayClic TEI, Le Mans, France
- DSBE: 'Automatic layout analysis and transcription of medieval manuscripts", Mondes anciens nouveaux regards, AnHIMA, Paris, 8 June 2017.
- DSBE: "Digital Humanities", University of Haifa, 10 June 2017.

- DSBE: “Automatic lectionary analysis with the database Thesaurus Antiquorum Lectionarium Ecclesiae Synagogaeque”, Hagiographico-homiletic collections in Greek, Latin and Oriental Manuscripts –Histories of Books and Text Transmission in a comparative perspective, Hamburg, 27 June 2017.
- DSBE: “Teaching DH and Jewish Studies”, Hamburg, 5 September 2017.
- DSBE: “Automatic layout analysis and transcription of medieval Hebrew manuscripts”, Jerusalem, 22 June 2017.
- DS: "Robust Data Driven Parsing of Deep Syntax Graph: Syntax Is Not Dead Yet", Naverlabs, ex-Xerox, Grenoble, December 8th, 2017.

### *10.1.5. Leadership within the Scientific Community*

- LR: President of the board of directors of DARIAH
- LR: Member of the board of directors or the TEI consortium
- LR: President of ISO committee TC 37 (Language and terminology)
- EVdLC: Chairman of the ACL special interest group SIGPARSE
- BS: Member, Deputy Treasurer and Member of the Board of the Société de Linguistique de Paris (since Dec. 2017)
- DS: Board member of the French NLP society (Atala, 2017-2020), program chair of the "journée d’études".
- Member of the ACL’s BIG (Broad Interest Group) Diversity group.

### *10.1.6. Scientific Expertise*

- EVdLC: Reviewer for an European ERC proposal and for an European COST proposal
- BS: Reviewer for the ANR (CE32 committee)
- EVdLC: Member of the selection committee for the call for projects “Langues et numérique 2017” organized by DGLFLF (Délégation générale à la langue française et aux langues de France)
- DSBE: Reviewer for a proposal at the SNF.

### *10.1.7. Research Administration*

- XY: [description]
- DSBE: Director of the DH Programme of the EPHE.
- BS: Member of the Board of Inria Paris’s Scientific Committee ("Comité des Projets")
- BS: Member of the International Relations Working Group of Inria’s Scientific and Technological Orientation Council (COST-GTRI)
- BS: Deputy Head of the research strand on Language Resources of the LabEx EFL (Empirical Foundations of Linguistics), and is therefore a deputy member of the Governing Board of the LabEx; BS and DS are in charge of several research operations in the LabEx

### *10.1.8. Teaching*

Licence: Djamé Seddah, “Certificat Informatique et Internet”, 30h, L1-L2-L3, Université Paris Sorbonne, France

Licence: Djamé Seddah, “Programmation et algorithmique en Java”, 50h, L2, Université Paris Sorbonne, France

Licence: Loïc Grobol, “Informatique et Industries de la Langue”, 22h, L2, Université Sorbonne Nouvelle, France

Licence: Loïc Grobol, “Introduction aux Humanités Numériques”, 15h, L2, Université Sorbonne Nouvelle, France

Master: Djamé Seddah, “Modèles pour la linguistique computationnelle”, 36h, M2, Université Paris Sorbonne, France

Master: Djamé Seddah, “Modèles pour la linguistique computationnelle”,36h, M1, Université Paris Sorbonne, France

Master: Djamé Seddah, “Programmation générique et C++”, 26h, M1, Université Paris Sorbonne, France

Master: Djamé Seddah, “Programmation réseau et Java”, 26h, M1, Université Paris Sorbonne, France

Master: Djamé Seddah, “Traduction automatique”, 30h, M2, Université Paris Sorbonne, France

Master: Laurent Romary, “Governance challenges in setting up and running an ERIC”, 1h30, Webinar RItrain – Executive Master in Management of Research Infrastructures at University of Milano-Bicocca, 24 November 2017

Master: Marie Puren, Charles Riondet, “Formation à la TEI pour des documents historiques”, 3 heures, M1-M2, Université Aix-Marseille, France, 19 October 2017

Master: Marie Puren, “Valorisation de la recherche - Humanités numériques”, 4 X 3 heures, M2 Histoire, Université Versailles-Saint-Quentin, Rennes, France, January-March 2017.

Master/PhD: Daniel Stökl Ben Ezra, “Approches numériques aux textes du judaïsme ancien”, 26 heures, M2/PhD, EPHE

Master/PhD: Marc Bui, “Introduction à la programmation Python pour les chercheurs en SHS” (2x24h) niveau M1/M2/PhD, EPHE (EPHE students and Master humanités numériques ENC-ENS-EPHE students)

Master class, Daniel Stökl Ben Ezra, “Simple and Advanced Image Treatment for Manuscript Analysis”, 2h, manuSciences summer school, 11 September 2017

Master class: Laurent Romary, “Overview of lexical models and introduction to the TEI dictionary chapter”, 1h30, Lexical Master Class, Berlin, 5 December 2017

Master class: Laurent Romary, “Querying and presenting TEI dictionary data with XSLT”, 1h30, Lexical Master Class, Berlin, 6 December 2017

Master class: Marie Puren, “Data management practices and recommendations. Managing, sharing and preserving linguistic data”, 1h30, Lexical Master Class, Berlin, 7 December 2017

Master class: Mohamed Khemakhem, “GROBID-Dictionaries”, 4 X 3 heures , Lexical Master Class, Berlin, 5-7 December 2017

EPHE, Marc Bui, “Introduction à la conception des bases de données avec SQL” (24h)

EPHE, Marc Bui, “La mise en page avec LateX pour les chercheurs en SHS” (12h)

IES Inria, Laurent Romary, “Formation à la TEI, XPath et XSLT pour HAL”, 3 journées Centre Inria Paris

IES Inria, Marie Puren, Charles Riondet, “Gestion des données de la recherche”, 7 heures, Centre Inria Paris, France, 31 March 2017

INIST, Marie Puren, “Créer un plan de gestion des données” et “Les métadonnées dans un DMP”, 2h30, INIST, Vandoeuvre-Lès-Nancy, France, 6 July 2017.

INIST: Laurent Romary, “Formation à la TEI pour les documents scientifiques”, 2 X 2 journées, INIST, Vandoeuvre Lès Nancy, France

URFIST, Marie Puren, “Données de recherche : le Plan de Gestion des Données”, 1 journée, URFIST, Rennes, France, 1st June 2017

### *10.1.9. Supervision*

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

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: 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", October 2016, EPHE, 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: Mohamed Khemakhem, "Structuration automatique de dictionnaires à partir de modèles lexicaux standardisés", September 2016, Paris Diderot, Laurent Romary

PhD in progress: Antony Perrot, "Qumran Opisthographs", started in Oct 2015, EPHE, PSL, supervised by Daniel Stökl Ben Ezra

PhD in progress: Jérémie Bosom , "Big data, internet des objets, fouille de données: élaboration de services intelligents pour le pilotage industriel", started in Oct 2015, EPHE, PSL, supervised by Marc Bui

PhD in progress: Dogu Kaan Eraslan, "Les relations entre Milet et l'Égypte à la Basse Époque (664-332 av. J.-C.)" (with a strong emphasis on computationa humanity approaches for both encoding information and extracting information, e.g. with neural image processing techniques on ancient documents), started in Oct 2015, EPHE, PSL, co-supervised by Michel Chauveau and Marc Bui

### *10.1.10. Juries*

- EVdLC: member of the PhD committee for Jakub Waszczuk at University of Blois on June 26th (Title: "Leveraging MWEs in practical TAG parsing: towards the best of the two worlds"; Supervisors: Agata Savary and Yannick Parmentier)
- MB: member of the PhD committee for Karim Sayadi at Université Paris 6 en partenariat avec EPHE on March 28th (Title: "Classification du texte numérique et numérisé. Approche fondée sur les algorithmes d'apprentissage automatique"; Supervisor: Marc Bui)
- EVdLC: examiner and member of the Master 2 jury for Jean Argouarc'h (Master de sciences cognitives, Paris Diderot; title: "Semantic models for analysis of brain activation during naturalistic text listening"; Supervisor: C. Pallier)
- BS: member of the recruiting committee for a Maître de Conférences position (NLP) at Université Paris Sorbonne

## 10.2. Standardization activities

### *10.2.1. ISO TC 37/ SC4*

- Mohamed Khemakhem and Laurent Romary: Project leaders of the ISO 24613-4 LMF "TEI Serialisation"
- Jack Bowers: Project leader of the ISO 24613-3 LMF "Etymology Extension"
- Éric de la Clergerie: Participation to AFNOR meetings, in relation with TC37/ SC4

### *10.2.2. COST ENEL & DARIAH*

- Laurent Romary, Mohamed Khemakhem, Axel Herold and Jack Bowers: Experts of a joint lexical standardisation action "TEI Lex0: Towards Best TEI P5 Encoding Practices"

## 10.3. Popularization

- BS, jointly with Emmanuel Dupoux (EHESS & ENS), gave a talk on “NLP and AI” as part of the seminar on AI organised by the Institut de l’École Normale Supérieure (21 June 2017)
- EVdLC and BS presented ALMAnaCH’s research (as well as its spin-off opensquare) at the “Rencontres Inria-industries” (forum bringing together Inria and companies) (18 November 2017)
- DSBE and BS: participation to a meeting on Digital Humanities bringing together researchers from the PSL ComUE, the University of Cambridge (UK) and the EPFL (Switzerland) (11 May 2017)
- BS: with Jean Ponce, Isabelle Ryl and Hélène Robak, represented the Inria Paris research center at the forum organised for the 30th anniversary of the DRM (the French Military Intelligence Office) (23 March 2017)
- BS and DSBE: talks during the NLP edition of the “Paris Sciences et Data” conference series.
- EVdLC: talk at the opening of the Math Olympiades 2017 about “Une palette mathématique pour appréhender le langage” (Versailles, January 25 2017)
- EVdLC: talk at the GFII DIXIT Seminar on “IA et Traitement Automatique des Langues (TAL) : Quel panorama ?” (Paris, February 24th 2017); member of the organizing committee of the forum 2017 of the GFII and panelist of the session on“de l’IA washing à la réalité industrielle, quels sont les contours du renouveau actuel de l’IA ?” (December 5th, 2017)
- EVdLC: co-animator of a new GFII Working Group “Technologies de la Connaissance”, with a focus on AI
- Loïc Grobol: participation to the 18th meeting on “Culture & Jeux Mathématiques” organized by AMIES
- MP, LR : Webinar on “Humanities and Open Science: Workflows and tools for publishing, licensing, versioning, identifiers, archiving, software...” for the International Open Access Week, 26 October 2017.
- DS gave a talk on “From Noisy Questions to Minecraft Texts: Annotation Challenges in Extreme Syntax Scenarios” at the NLP Paris Meetup, Paris, November 22th, 2017.

# 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, 35 p. [*DOI :* 10.1007/s10579-015-9295-6], https://hal.inria.fr/hal-01174492

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

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

[4] 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

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

[6] 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

[7] 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

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

[9] 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

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

[11] J. BOWERS, L. ROMARY. *Deep encoding of etymological information in TEI*, in "Journal of the Text Encoding Initiative", August 2017, n° 10, https://arxiv.org/abs/1611.10122 [*DOI :* 10.4000/JTEI.1643], https://hal.inria.fr/hal-01296498

[12] R. GARNIER, B. SAGOT. *A shared substrate between Greek and Italic*, in "Indogermanische Forschungen", September 2017, vol. 122, n° 1, pp. 29-60 [*DOI :* 10.1515/IF-2017-0002], https://hal.inria.fr/hal-01621467

[13] B. SAGOT. *Représentation de l'information sémantique lexicale : le modèle wordnet et son application au français*, in "Revue Française de Linguistique Appliquée", 2017, vol. XXII, https://hal.inria.fr/hal-01583995

### Invited Conferences

[14] A. BAILLOT. *Zahlenwahn oder Textliebe? Digitale Philologie als Disziplin und als Weltanschauung*, in "Machines / Maschinen Les 5O ans de l'AGES", Nantes, France, Association des Germanistes de l'Enseignement Supérieur, June 2017, https://halshs.archives-ouvertes.fr/halshs-01562486

[15] L. ROMARY. *The Text Encoding Initiative as Infrastructure*, in "French-Israeli Symposium on Digital Humanities", Jerusalem, Israel, February 2017, https://hal.inria.fr/hal-01618017

### International Conferences with Proceedings

[16] M. CANDITO, B. GUILLAUME, G. PERRIER, D. SEDDAH. *Enhanced UD Dependencies with Neutralized Diathesis Alternation*, in "Depling 2017 - Fourth International Conference on Dependency Linguistics", Pisa, Italy, September 2017, https://hal.inria.fr/hal-01625466

[17] M. CONSTANT, H. MARTINEZ ALONSO. *Benchmarking Joint Lexical and Syntactic Analysis on Multiword-Rich Data*, in "MWE 2017 - 13th Workshop on Multiword Expressions", Valencia, Spain, Association for Computational Linguistics, April 2017, pp. 181 - 186, https://hal.inria.fr/hal-01677416

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

[19] 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

[20] H. MARAOUI, K. HADDAR, L. ROMARY. *Encoding prototype of Al-Hadith Al-Shareef in TEI*, in "ICALP 2017 - The 6th International Conference on Arabic Language Processing", Fes, Morocco, October 2017, 14 p. , https://hal.archives-ouvertes.fr/hal-01574543

[21] H. MARTINEZ ALONSO, Ž. AGIĆ, B. PLANK, A. SØGAARD. *Parsing Universal Dependencies without training*, in "EACL 2017 - 15th Conference of the European Chapter of the Association for Computational Linguistics", Valencia, Spain, Association for Computational Linguistics, April 2017, vol. 1, pp. 229 - 239, https://hal.inria.fr/hal-01677405

[22] H. MARTINEZ ALONSO, B. PLANK. *When is multitask learning effective? Semantic sequence prediction under varying data conditions*, in "EACL 2017 - 15th Conference of the European Chapter of the Association for Computational Linguistics", Valencia, Spain, April 2017, pp. 1-10, https://hal.inria.fr/hal-01677427

[23] H. MARTÍNEZ ALONSO, A. DELAMAIRE, B. SAGOT. *Annotating omission in statement pairs*, in "11th Linguistic Annotation Workshop", Valencia, Spain, April 2017, pp. 41-45, https://hal.inria.fr/hal-01584035

[24] S. PERNES, L. ROMARY, K. WARBURTON. *TBX in ODD: Schema-agnostic specification and documentation for TermBase eXchange*, in "LOTKS 2017- Workshop on Language, Ontology, Terminology and Knowledge Structures", Montpellier, France, September 2017, https://hal.inria.fr/hal-01581440

[25] 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

[26] B. SAGOT. *Construction automatique d'une base de données étymologiques à partir du wiktionary*, in "Traitement Automatique des Langues Naturelles 2017", Orléans, France, June 2017, https://hal.inria.fr/hal-01584013

[27] 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

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

[29] 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

[30] 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

### National Conferences with Proceedings

[31] L. GROBOL, I. TELLIER, É. DE LA CLERGERIE, M. DINARELLI, F. LANDRAGIN. *Experiences in using deep and shallow parsing to detect entity mentions in oral French*, in "TALN 2017", Orléans, France, Actes de la 24e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), Association pour le Traitement Automatique des Langues (ATALA), June 2017, https://hal.inria.fr/hal-01558711

[32] D. SEDDAH, M. CANDITO. *Building a Question Treebank for French : The French QuestionBank*, in "ACor4French - Les corpus annotés du français", Orléans, France, Acts de l'Atelier ACor4French - Les corpus annotés du français, June 2017, https://hal.inria.fr/hal-01682869

### Conferences without Proceedings

[33] A. BAILLOT, M. PUREN, C. RIONDET, D. SEILLIER, L. ROMARY. *Access to cultural heritage data. A challenge for digital humanities*, in "Digital Humanities 2017", Montréal, Canada, August 2017, https://hal.archives-ouvertes.fr/hal-01582176

[34] C. RIONDET, L. FOPPIANO. *GROBID for Humanities When engineering meets History*, in "Text as a Resource. Text Mining in Historical Science", Paris, France, Institut Historique Allemand, June 2017, https://hal.inria.fr/hal-01585693

[35] M. SEURET, D. STÖKL BEN EZRA, M. LIWICKI. *Robust Heartbeat-based Line Segmentation Methods for Regular Texts and Paratextual Elements*, in "HIP 2017 - Proceedings of the 4th International Workshop on Historical Document Imaging and Processing", Kyoto, Japan, November 2017, https://hal.archives-ouvertes.fr/hal-01677054

### Scientific Books (or Scientific Book chapters)

[36] PARTHENOS (editor). *Digital 3D Objects in Art and Humanities: challenges of creation, interoperability and preservation. White paper: A result of the PARTHENOS Workshop held in Bordeaux at Maison des Sciences de l'Homme d'Aquitaine and at Archeovision Lab. (France), November 30th - December 2nd, 2016*, PARTHENOS, Bordeaux, France, May 2017, 71 p. , https://hal.inria.fr/hal-01526713

[37] R. GARNIER, L. SAGART, B. SAGOT. *Milk and the Indo-Europeans*, in "Language Dispersal Beyond Farming", M. ROBEETS, A. SAVALYEV (editors), John Benjamins Publishing Company, December 2017, pp. 291–311 [*DOI :* 10.1075/Z.215.13GAR], https://hal.inria.fr/hal-01667476

[38] D. STÖKL BEN EZRA. *The Mishnah into French: translation issues*, in "Studies in Mishnaic Hebrew and Related Dialects : Proceedings of the Yale Symposium, May 2014", E. B. ASHER, A. KOLLER (editors), Studies in Mishnaic Hebrew and Related Fields Proceedings of the Yale Symposium on Mishnaic Hebrew, May 2014, The Program in Judaic Studies, Yale University, December 2017, pp. 349-367, https://hal.archives-ouvertes.fr/hal-01677074

### Books or Proceedings Editing

[39] A. DEGKWITZ, L. ROMARY (editors). *IFLA Satellite Meeting - Digital Humanities – Opportunities and Risks: Connecting Libraries and Research*, 2017, https://hal.inria.fr/hal-01643305

### Research Reports

[40] C. RIONDET, L. ROMARY, A. VAN NISPEN, K. J. RODRIGUEZ, M. BRYANT. *Report on Standards*, Inria Paris, March 2017, n$^{o}$ D.11.4, https://hal.inria.fr/hal-01503235

[41] L. ROMARY, P. BANSKI, J. BOWERS, E. DEGL'INNOCENTI, M. ĎURČO, R. GIACOMI, K. ILLMAYER, A. JOFFRES, F. KHAN, M. KHEMAKHEM, N. LARROUSSE, A. LITKE, M. MONACHINI, A. V. NISPEN, M. OGRODNICZUK, N. PAPADAKIS, G. PASTORE, S. PERNES, M. PUREN, C. RIONDET, M. SANZ, M. SANESI, P. SIOZOS, R. D. VALK. *Report on Standardization (draft)*, Inria, May 2017, n$^{o}$ Deliverable 4.2, https://hal.inria.fr/hal-01560563

[42] D. SEILLIER, A. BAILLOT, M. PUREN, C. RIONDET. *Survey on researchers requirements and practices towards Cultural Heritage institutions: Documentation and analysis*, Inria Paris, July 2017, https://hal.inria.fr/hal-01562860

### Scientific Popularization

[43] J. EDMOND, F. FISCHER, M. MERTENS, L. ROMARY. *The DARIAH ERIC: Redefining Research Infrastructure for the Arts and Humanities in the Digital Age*, in "ERCIM News", October 2017, n$^{o}$ 111, https://hal.inria.fr/hal-01588665

### Other Publications

[44] J. NIVRE, Ž. AGIĆ, L. AHRENBERG, L. ANTONSEN, M. J. ARANZABE, M. ASAHARA, L. ATEYAH, M. ATTIA, A. ATUTXA, L. AUGUSTINUS, E. BADMAEVA, M. BALLESTEROS, E. BANERJEE, S. BANK, V. BARBU MITITELU, J. BAUER, K. BENGOETXEA, R. A. BHAT, E. BICK, V. BOBICEV, C. BÖRSTELL, C. BOSCO, G. BOUMA, S. BOWMAN, A. BURCHARDT, M. CANDITO, G. CARON, G. CEBIROĞLU ERYIĞIT, G. G. A. CELANO, S. CETIN, F. CHALUB, J. CHOI, S. CINKOVÁ, Ç. ÇÖLTEKIN, M. CONNOR, E. DAVIDSON, M. D. MARNEFFE, V. D. PAIVA, A. D. D. ILARRAZA, P. DIRIX, K. DOBROVOLJC, T. DOZAT, K. DROGANOVA, P. DWIVEDI, M. ELI, A. ELKAHKY, T. ERJAVEC, R. FARKAS, H. FERNANDEZ ALCALDE, J. FOSTER, C. FREITAS, K. GAJDOŠOVÁ, D. GALBRAITH, M. GARCIA, M. GÄRDENFORS, K. GERDES, F. GINTER, I. GOENAGA, K. GOJENOLA, M. GÖKIRMAK, Y. GOLDBERG, X. GÓMEZ GUINOVART, B. GONZÁLES SAAVEDRA, M. GRIONI, N. GRŪZĪTIS, B. GUILLAUME, N. HABASH, J. HAJIČ, J. HAJIČ JR., L. HÀ MỸ, K. HARRIS, D. HAUG, B. HLADKÁ, J. HLAVÁČOVÁ, F. HOCIUNG, P. HOHLE, R. ION, E. IRIMIA, T. JELÍNEK, A. JOHANNSEN, F. JØRGENSEN, H. KAŞIKARA, H. KANAYAMA, J. KANERVA, T. KAYADELEN, V. KETTNEROVÁ, J. KIRCHNER, N. KOTSYBA, S. KREK, V. LAIPPALA,

L. LAMBERTINO, T. LANDO, J. LEE, P. LÊ HÔNG, A. LENCI, S. LERTPRADIT, H. LEUNG, C. Y. LI, J. LI, K. LI, N. LJUBEŠIĆ, O. LOGINOVA, O. LYASHEVSKAYA, T. LYNN, V. MACKETANZ, A. MAKAZHANOV, M. MANDL, C. MANNING, C. MĂRĂNDUC, D. MAREČEK, K. MARHEINECKE, H. MARTÍNEZ ALONSO, A. MARTINS, J. MAŠEK, Y. MATSUMOTO, R. MCDONALD, G. MENDONÇA, N. MIEKKA, A. MISSILÄ, C. MITITELU, Y. MIYAO, S. MONTEMAGNI, A. MORE, L. MORENO ROMERO, S. MORI, B. MOSKALEVSKYI, K. MUISCHNEK, K. MÜÜRISEP, P. NAINWANI, A. NEDOLUZHKO, G. NEŠPORE-BĒRZKALNE, L. NGUYỄN THỊ, H. NGUYỄN THỊ MINH, V. NIKOLAEV, H. NURMI, S. OJALA, P. OSENOVA, R. ÖSTLING, L. ØVRELID, E. PASCUAL, M. PASSAROTTI, C. PEREZ, G. PERRIER, S. PETROV, J. PIITULAINEN, E. PITLER, B. PLANK, M. POPEL, L. PRETKALNIŅA, P. PROKOPIDIS, T. PUOLAKAINEN, S. PYYSALO, A. RADEMAKER, L. RAMASAMY, T. RAMA, V. RAVISHANKAR, L. REAL, S. REDDY, G. REHM, L. RINALDI, L. RITUMA, M. ROMANENKO, R. ROSA, D. ROVATI, B. SAGOT, S. SALEH, T. SAMARDŽIĆ, M. SANGUINETTI, B. SAULĪTE, S. SCHUSTER, D. SEDDAH, W. SEEKER, M. SERAJI, M. SHEN, A. SHIMADA, D. SICHINAVA, N. SILVEIRA, M. SIMI, R. SIMIONESCU, K. SIMKÓ, M. ŠIMKOVÁ, K. SIMOV, A. SMITH, A. STELLA, M. STRAKA, J. STRNADOVÁ, A. SUHR, U. SULUBACAK, Z. SZÁNTÓ, D. TAJI, T. TANAKA, T. TROSTERUD, A. TRUKHINA, R. TSARFATY, F. TYERS, S. UEMATSU, Z. UREŠOVÁ, L. URIA, H. USZKOREIT, S. VAJJALA, D. V. NIEKERK, G. V. NOORD, V. VARGA, E. V. D. L. CLERGERIE, V. VINCZE, L. WALLIN, J. N. WASHINGTON, M. WIRÉN, T. WONG, Z. YU, Z. ŽABOKRTSKÝ, A. ZELDES, D. ZEMAN, H. ZHU. *Universal Dependencies 2.1*, November 2017, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University - Corpus - Project code: 15-10472S; Project name: Morphologically and Syntactically Annotated Corpora of Many Languages, https://hal.inria.fr/hal-01682188

[45] M. PUREN, C. RIONDET, D. SEILLIER, L. ROMARY. *The Standardization Survival Kit (SSK): For a wider use of standards within Arts and Humanities*, July 2017, Digital Humanities Benelux Conference 2017, Poster, https://hal.archives-ouvertes.fr/hal-01587687

[46] L. ROMARY, J. EDMOND. *Sustainability in DARIAH*, April 2017, 10 p. , Sustainability of Digital Research Infrastructures for the Arts and Humanities, https://hal.inria.fr/hal-01516487

[47] L. ROMARY, C. RIONDET. *Ongoing maintenance and customization of archival standards using ODD (EAC-CPF revision proposal)*, December 2017, EAC-CPF revision proposal, https://hal.inria.fr/hal-01677185

[48] L. ROMARY, C. RIONDET. *Towards multiscale archival digital data*, September 2017, working paper or preprint, https://hal.inria.fr/hal-01586389

[49] L. ROMARY. *How to Open up? (Digital) Libraries at the Service of (Digital) Scholars*, April 2017, Fiesole Collection Development Retreat, https://hal.inria.fr/hal-01513674

[50] V. VANDEN DAELEN, J. EDMOND, P. LINKS, M. PRIDDY, L. REIJNHOUDT, V. TOLLAR, A. VAN NISPEN, C. HAUWAERT, C. RIONDET. *La publication durable digitale des guides d'archives de l'histoire du 20ème siècle*, November 2017, working paper or preprint, https://hal.inria.fr/hal-01632366

## References in notes

[51] 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

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

[53] 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

[54] 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

[55] 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

[56] P. DENIS, B. SAGOT. *Coupling an annotated corpus and a lexicon for state-of-the-art POS tagging*, in "Language Resources and Evaluation", 2012, vol. 46, n$^{o}$ 4, pp. 721-736 [*DOI :* 10.1007/s10579-012-9193-0], https://hal.inria.fr/inria-00614819

[57] 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

[58] 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

[59] 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

[60] 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.

[61] H. A. MARTÍNEZ, 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

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

[63] 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.

[64] 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

[65] 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

[66] L. ROMARY, E. DEGL'INNOCENTI, K. ILLMAYER, A. JOFFRES, E. KRAIKAMP, N. LARROUSSE, M. OGRODNICZUK, M. PUREN, C. RIONDET, D. SEILLIER. *Standardization survival kit (Draft)*, Inria, October 2016, n° Deliverable 4.1, https://hal.inria.fr/hal-01513531

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

[68] 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

[69] 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

[70] 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

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

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

[73] 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

[74] 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

[75] 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

[76] 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

[77] 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

[78] 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

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

[80] G. WALTHER, G. JACQUES, B. SAGOT. *Uncovering the inner architecture of Khaling verbal morphology*, in "3rd Workshop on Sino-Tibetan Languages of Sichuan", Paris, France, September 2013, https://hal.inria.fr/hal-00927278

[81] G. WALTHER, G. JACQUES, B. SAGOT. *The Opacity-Compactness Tradeoff: Morphomic Features for an Economical Account of Khaling Verbal Inflection*, in "16th International Morphology Meeting (IMM 16)", Budapest, Hungary, May 2014, https://hal.inria.fr/hal-01114854