RESEARCH CENTRE

Paris

2021 ACTIVITY REPORT

Project-Team ALMANACH

Automatic Language Modelling and Analysis & Computational Humanities

DOMAIN

Perception, Cognition and Interaction

THEME

Language, Speech and Audio

Contents

Project-Team ALMANACH

Creation of the Project-Team: 2019 July 01

Keywords

Computer sciences and digital sciences

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

A5.8. – Natural language processing

A9.1. – Knowledge

A9.2. – Machine learning

A9.4. – Natural language processing

A9.7. – AI algorithmics

Other research topics and application domains

B1.2.2. – Cognitive science

B1.2.3. – Computational neurosciences

B9.5.6. – Data science

B9.6.2. – Juridical science

B9.6.5. – Sociology

B9.6.6. – Archeology, History

B9.6.8. – Linguistics

B9.6.10. – Digital humanities

B9.7. – Knowledge dissemination

B9.7.1. – Open access

B9.7.2. – Open data

B9.8. – Reproducibility

1 Team members, visitors, external collaborators

Research Scientists

- Benoît Sagot [Team leader, Inria, Senior Researcher, HDR]
- Rachel Bawden [Inria, Researcher]
- Laurent Romary [Inria, Senior Researcher, HDR]
- Djamé Seddah [Inria, Researcher]
- Éric Villemonte de La Clergerie [Inria, Researcher]

Post-Doctoral Fellow

- Syrielle Montariol [Inria, from Apr 2021]

PhD Students

- Roman Castagné [Inria, from Oct 2021]
- Alix Chagué [Inria, from Nov 2021, Joint PhD thesis between Inria and Université de Montréal]
- Floriane Chiffoleau [Université du Mans, From Nov 2021, Joint PhD thesis between Inria and Université du Mans]
- Paul-Ambroise Duquenne [Facebook, from May 2021]
- Clementine Fourrier [Inria]
- Nathan Godey [Inria, from Dec 2021]
- Louis Martin [Facebook, until Sep 2021]
- Benjamin Muller [Inria, from Aug 2021]
- Tu Anh Nguyen [Facebook, from Apr 2021]
- Lydia Nishimwe [Inria, from Oct 2021]
- Pedro Ortiz Suarez [Inria]
- Mathilde Regnault [École Normale Supérieure de Paris, until Sep 2021]
- Arij Riabi [Inria, from Oct 2021]
- Jose Rosales Nunez [CNRS]
- Lionel Tadonfouet [Orange, CIFRE]

Technical Staff

- Julien Abadji [Inria, Engineer, from Apr 2021]
- Quentin Burthier [Inria, Engineer, from Feb 2021 until Mar 2021]
- Alix Chague [Inria, Engineer, until Oct 2021]
- Thibault Charmet [Inria, Engineer, from Feb 2021]
- Floriane Chiffoleau [Inria, Engineer, Until Oct 2021]
- Tanti Kristanti Nugraha [Inria, Engineer]

- Arij Riabi [Inria, Engineer, until Sep 2021]
- Hugo Scheithauer [Inria, Engineer, from Oct 2021]
- Yves Tadjo Takianpi [Inria, Engineer]
- Lucas Terriel [Inria, Engineer]
- Thomas Wang [Inria, Engineer, from Apr 2021 until Sep 2021]
- You Zuo [Inria, Engineer, from Oct 2021]

Interns and Apprentices

- Quentin Burthier [Inria, Jan 2021]
- Roman Castagné [Inria, from Apr 2021 until Sep 2021]
- Matthieu Futeral-Peter [Inria, from May 2021 until Oct 2021]
- Manon Ovide [Inria, from Mar 2021 until Aug 2021]
- Camille Rey [Inria, from Sep 2021]
- Sonal Sannigrahi [Inria, from Jun 2021 until Aug 2021]
- Hugo Scheithauer [Inria, from Apr 2021 until Aug 2021]

Administrative Assistant

- Meriem Guemair [Inria]

2 Overall objectives

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

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

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

The scientific positioning of ALMAnaCH extends that of its Inria predecessor, the project-team ALPAGE, a joint team with Paris-Diderot University dedicated to research in NLP and computational linguistics. ALMAnaCH remains committed to developing state-of-the-art NLP software and resources that can be used by academics and in the industry. At the same time we continue our work on language modelling in order to provide a better understanding of languages, an objective that is reinforced and addressed in the broader context of computational humanities. Finally, we remain dedicated to having an impact on the industrial world and more generally on society, via multiple types of collaboration with companies and other institutions (startup creation, industrial contracts, expertise, etc.).

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

One of the main challenges in computational linguistics is **to model and to cope with language variation**. Language varies with respect to domain and genre (news wires, scientific literature, poetry, oral transcripts…), sociolinguistic factors (age, background, education; variation attested for instance on social media), geographical factors (dialects) and other dimensions (disabilities, for instance). But language also constantly evolves at all time scales. Addressing this variability is still an open issue for NLP. Commonly used approaches, which often rely on supervised and semi-supervised machine learning methods, require very large amounts of annotated data. They still suffer from the high level of variability found for instance in **user-generated content**, **non-contemporary texts**, as well as in **domain-specific documents** (e.g. financial, legal).

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

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

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

3 Research program

3.1 Research strands

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

3.1.1 Research axis 1

Our first objective is to **stay at a state-of-the-art level in key NLP tasks** such as shallow processing, part-of-speech tagging and (syntactic) parsing, which are core expertise domains of ALMAnaCH members. This will also require us to improve the **generation of semantic representations (semantic parsing)**, and to begin to explore tasks such as machine translation, which now relies on neural architectures also used for some of the above-mentioned tasks. Given the generalisation of neural models in NLP, we will also be involved in better understanding how such models work and what they learn, something that is directly related to the investigation of language variation (Research axis 2). We will also work on the **integration of both linguistic and non-linguistic contextual information** to improve automatic

linguistic analysis. This is an emerging and promising line of research in NLP. We will have to identify, model and take advantage of each type of contextual information available. Addressing these issues will enable the development of new lines of research related to conversational content. Applications include improved information and knowledge extraction algorithms. We will especially focus on challenging datasets such as domain-specific texts (e.g. financial, legal) as well as historical documents, in the larger context of the development of digital humanities. We currently also explore the even more challenging new direction of a cognitively inspired NLP, in order to tackle the possibility to enrich the architecture of state-of-the-art algorithms, such as RNNGs, based on human neuroimaging-driven data.

3.1.2 Research axis 2

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

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

3.1.3 Research axis 3

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

3.2 Automatic Context-augmented Linguistic Analysis

This first research strand is centred around NLP technologies and some of their applications in Artificial Intelligence (AI). Core NLP tasks such as part-of-speech tagging, syntactic and semantic parsing is

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

3.2.1 Processing of natural language at all levels: morphology, syntax, semantics

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

Traditionally, a full natural language process (NLP) chain is organised as a pipeline where each stage of analysis represents a traditional linguistic field (in a *structuralism* view) from morphological analysis to purely semantic representations. The problem is that this architecture is vulnerable to error propagation and very domain sensitive: each of these stage must be compatible at the lexical and structure levels they provide. We arguably built the best performing NLP chain for French [69, 118] and one of the best for robust multilingual parsing as shown by our results in various shared tasks over the years [111, 106, 119], [78]. 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 [106].

In particular, we continue to explore the hybridisation of symbolic and statistical approaches, and extend it to neural approaches, as initiated in the context of our participation to the CoNLL 2017 multilingual parsing shared task [2] and to Extrinsic Parsing Evaluation Shared Task [3].

Fundamentally, we want to build tools that are less sensitive to variation, more easily configurable, and self-adapting. Our short-term goal is to explore techniques such as multi-task learning (cf. already [114]) to propose a joint model of tokenisation, normalisation, morphological analysis and syntactic analysis. We also explore adversarial learning, considering the drastic variation we face in parsing user-generated content and processing historical texts, both seen as noisy input that needs to be handled at training and decoding time.

3.2.2 Integrating context in NLP systems

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

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

A more expressive and context-aware structural representation of a linguistic production is required to accurately model UGC. Let us consider for instance the case for Syntax-based Machine Translation of social media content, as is carried out by the ALMAnaCH-led ANR project Parsiti (PI: DS). A Facebook post may be part of a discussion thread, which may include links to external content. Such information is required for a complete representation of the post's context, and in turn its accurate machine translation. Even for the presumably simpler task of POS tagging of dialogue sequences, the addition of context-based

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

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

features (namely information about the speaker and dialogue moves) was beneficial [83]. In the case of UGC, working across sentence boundaries was explored for instance, with limited success, by [68] for document-wise parsing and by [98] for POS tagging.

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

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

3.2.3 Information and knowledge extraction

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

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

Word clustering is of specific importance. It relies on the distributional hypothesis initially formulated by Harris, which states that words occurring in similar contexts tend to be semantically close. The latest developments of these ideas (with word2vec or GloVe) have led to the embedding of words (through vectors) in low-dimensional semantic spaces. In particular, words that are typical of several communities (see above) can be embedded in a same semantic space in order to establish mappings between them. It is also possible in such spaces to study static configurations and vector shifts with respect to variables such as time, using topological theories (such as pretopology), for instance to explore shifts in meaning over time (cf. the ANR project Profiterole concerning ancient French texts) or between communities (cf. the ANR project SoSweet). It is also worth mentioning on-going work (in computational semantics) whose goal is to combine word embeddings to embed expressions, sentences, paragraphs or even documents into semantic spaces, e.g. to explore the similarity of documents at various time periods.

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

their strong potential in industrial applications.

3.3 Computational Modelling of Linguistic Variation

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

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

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

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

3.3.1 Theoretical and empirical synchronic linguistics

Permanent members involved: all

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

Recent advances in morphology rely on quantitative and computational approaches and, sometimes, on collaboration with descriptive linguists—see for instance the special issue of the *Morphology* journal on "computational methods for descriptive and theoretical morphology", edited and introduced by [66]. In this regard, ALMAnaCH members have taken part in the design of quantitative approaches to defining and measuring morphological complexity and to assess the internal structure of morphological systems (inflection classes, predictability of inflected forms…). Such studies provide valuable insights on these prominent questions in theoretical morphology. They also improve the linguistic relevance and the development speed of NLP-oriented lexicons, as also demonstrated by ALMAnaCH members. We shall therefore pursue these investigations, and orientate them towards their use in diachronic models (cf. Section 3.3.3).

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

will be information about processing evolution (by the patients) along the novel, possibly due to the use of contextual information by humans.

3.3.2 Sociolinguistic variation

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

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

However, NLP tools are designed for standard forms of language and exhibit a drastic loss of accuracy when applied to social media varieties or non-standardised historical sources. To define appropriate tools, descriptions of these varieties are needed. However, to validate such descriptions, tools are also needed. We address this chicken-and-egg problem in an interdisciplinary fashion, by working both on linguistic descriptions and on the development of NLP tools. Recently, socio-demographic variables have been shown to bear a strong impact on NLP processing tools (see for instance [76] and references therein). This is why, in a first step, jointly with researchers involved in the ANR project SoSweet (ENS Lyon and Inria project-team Dante), we will study how these variables can be factored out by our models and, in a second step, how they can be accurately predicted from sources lacking these kinds of featured descriptions.

3.3.3 Diachronic variation

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

Very few computational models of language change have been developed. Simple deterministic finite-state-based phonetic evolution models have been used in different contexts. The PIElexicon project [93] uses such models to automatically generate forms attested in (classical) Indo-European languages but is based on an idiosyncratic and unacceptable reconstruction of the Proto-Indo-European language. Probabilistic finite-state models have also been used for automatic cognate detection and proto-form reconstruction, for example by [67] and [77]. Such models rely on a good understanding of the phonetic evolution of the languages at hand.

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

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

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

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

Diachronic evolution also applies to syntax, and in the context of the ANR project Profiterole, we are beginning to explore more or less automatic ways of detecting these evolutions and suggest modifications, relying on fine-grained syntactic descriptions (as provided by meta-grammars), unsupervised sentence clustering (generalising previous works on error mining, cf. [9]), and constraint relaxation (in meta-grammar classes). The underlying idea is that a new syntactic construction evolves from a more ancient one by small, iterative modifications, for instance by changing word order, adding or deleting functional words, etc.

3.3.4 Accessibility-related variation

Language variation does not always pertain to the textual input of NLP tools. It can also be characterised by their intended output. This is the perspective from which we investigate the issue of text simplification (for a recent survey, see for instance [113]). 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 [94]. To this end, guidelines have been developed to help writing documents that are easier to read and understand, such as the FALC ("Facile À Lire et à Comprendre") guidelines for French. [4]

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 [75, 104], our goal will be to develop tools for the computer-aided simplification of textual documents, especially administrative documents. Many of the FALC guidelines can only be linguistically expressed using complex, syntactic constraints, and the amount of available "parallel" data (aligned raw and simplified documents) is limited. We will therefore investigate hybrid techniques involving rule-based, statistical and neural approaches based on parsing results (for an example of previous parsing-based work, see [64]). Lexical simplification, another aspect of text simplification [84, 95], will also be pursued. In this regard, we have already started a collaboration with Facebook's AI Research in Paris, the UNAPEI (the largest French federation of associations defending and supporting people with intellectual disabilities and their families), and the French Secretariat of State in charge of Disabled Persons.

Accessibility can also be related to the various presentation forms of a document. This is the context in which we have initiated the OPALINE project, funded by the *Programme d'Investissement d'Avenir - Fonds pour la Société Numérique*. The objective is for us to further develop the GROBID text-extraction suite[5] in order to be able to re-publish existing books or dictionaries, available in PDF, in a format that is accessible by visually impaired persons.

3.4 Modelling and Development of Language Resources

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

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

[5] Site internet de GROBID.

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

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

3.4.1 Construction, management and automatic annotation of Text Corpora

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

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

ALMAnaCH pays a specific attention to the re-usability [6] of all resources produced and maintained within its various projects and research activities. To this end, we will ensure maximum compatibility with available international standards for representing textual sources and their annotations. More precisely we will take the TEI (*Text Encoding Initiative*) guidelines as well the standards produced by ISO committee TC 37/SC 4 as essential points of reference.

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

The process may be seen as progressively enriching the documents with new layers of annotations produced by various NLP modules and possibly validated by users, preferably in a collaborative way. It relies on the use of clearly identified representation formats for the annotations, as advocated within ISO TC 37/SC 4 standards and the TEI guidelines, but also on the existence of well-designed collaborative interfaces for browsing, querying, visualisation, and validation. ALMAnaCH has been or is working on several of the NLP bricks needed for setting such a work-flow, and has a solid expertise in the issues related to standardisation (of documents and annotations). However, putting all these elements in a

[6] From a larger point of view we intend to comply with the so-called FAIR principles.

unified work-flow that is simple to deploy and configure remains to be done. In particular, work-flow and interface should maybe not be dissociated, in the sense that the work-flow should be easily piloted and configured from the interface. An option will be to identify pertinent emerging platforms in DH (such as Transkribus) and to propose collaborations to ensure that NLP modules can be easily integrated.

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

3.4.2 Development of Lexical Resources

ALPAGE, the Inria predecessor of ALMAnaCH, has put a strong emphasis in the development of morphological, syntactic and wordnet-like semantic lexical resources for French as well as other languages (see for instance [8, 1]). Such resources play a crucial role in all NLP tools, as has been proven among other tasks for POS tagging [100, 102, 119] and parsing, and some of the lexical resource development will be targeted towards the improvement of NLP tools. They will also play a central role for studying diachrony in the lexicon, for example for Ancient to Contemporary French in the context of the Profiterole project. They will also be one of the primary sources of linguistic information for augmenting language models used in OCR systems for ancient scripts, and will allow us to develop automatic annotation tools (e.g. POS taggers) for low-resourced languages (see already [120]), especially ancient languages. Finally, semantic lexicons such as wordnets will play a crucial role in assessing lexical similarity and automating etymological research.

Therefore, an important effort towards the development of new morphological lexicons will be initiated, with a focus on ancient languages of interest. Following previous work by ALMAnaCH members, we will try and leverage all existing resources whenever possible such as electronic dictionaries, OCRised dictionaries, both modern and ancient [99, 81, 101], while using and developing (semi)automatic lexical information extraction techniques based on existing corpora [103, 105]. A new line of research will be to integrate the diachronic axis by linking lexicons that are in diachronic relation with one another thanks to phonetic and morphological change laws (e.g. XIIth century French with XVth century French and contemporary French). Another novelty will be the integration of etymological information in these lexical resources, which requires the formalisation, the standardisation, and the extraction of etymological information from OCRised dictionaries or other electronic resources, as well as the automatic generation of candidate etymologies. These directions of research are already investigated in ALMAnaCH [81, 101].

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

In the same way as we signalled the importance of standards for the representation of interoperable corpora and their annotations, we will keep making the best use of the existing standardisation background for the representation of our various lexical resources. There again, the TEI guidelines play a central role, and we have recently participated in the "TEI Lex 0" initiative to provide a reference subset for the "Dictionary" chapter of the guidelines. We are also responsible, as project leader, of the edition of the new part 4 of the ISO standard 24613 (LMF, Lexical Markup Framework) [97] dedicated to the definition of the TEI serialisation of the LMF model (defined in ISO 24613 part 1 'Core model', 2 'Machine Readable Dictionaries' and 3 'Etymology'). We consider that contributing to standards allows us to stabilise our knowledge and transfer our competence.

3.4.3 Development of Annotated Corpora

Along with the creation of lexical resources, ALMAnaCH is also involved in the creation of corpora either fully manually annotated (gold standard) or automatically annotated with state-of-the-art pipeline processing chains (silver standard). Annotations will either be only morphosyntactic or will cover more

complex linguistic levels (constituency and/or dependency syntax, deep syntax, maybe semantics). Former members of the ALPAGE project have a renowned experience in those aspects (see for instance [112, 96, 109, 89]) and will participate to the creation of valuable resources originating from the historical domain genre.

Under the auspices of the ANR Parsiti project, led by ALMAnaCH (PI: DS), we aim to explore the interaction of extra-linguistic context and speech acts. Exploiting extra-linguistics context highlights the benefits of expanding the scope of current NLP tools beyond unit boundaries. Such information can be of spatial and temporal nature, for instance. They have been shown to improve Entity Linking over social media streams [72]. In our case, we decided to focus on a closed world scenario in order to study context and speech acts interaction. To do so, we are developing a multimodal data set made of live sessions of a first person shooter video game (Alien vs. Predator) where we transcribed all human players interactions and face expressions streamlined with a log of all in-game events linked to the video recording of the game session, as well as the recording of the human players themselves. The in-games events are ontologically organised and enable the modelling of the extra-linguistics context with different levels of granularity. Recorded over many games sessions, we already transcribed over 2 hours of speech that will serve as a basis for exploratory work, needed for the prototyping of our context-enhanced NLP tools. In the next step of this line of work, we will focus on enriching this data set with linguistic annotations, with an emphasis on co-references resolutions and predicate argument structures. The midterm goal is to use that data set to validate a various range of approaches when facing multimodal data in a close-world environment.

4 Application domains

4.1 Application domains for ALMAnaCH

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

- Information extraction, information retrieval, text mining (e.g. 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 Social and environmental responsibility

5.1 Footprint of research activities

In view of recent interest about the energy consumption and carbon emission of machine learning models, and specifically of those of language models [107, 65], we have decided to report the power consumption [7] and carbon footprint of all our experiments conducted on the Jean-Zay supercomputer during 2021. For this report, we follow the approach of [116]. While the ALMAnaCH team uses other

[7] Jean-Zay documentation

Project	GPU-hours	Real-hours	Power Consumption (kWh)	CO2 Emissions (kg)
AD011011330R1	15020	3755.00	6849.12	219.17
AD011012676	4240	1060.00	1933.44	61.87
AD011012254	23809	5952.25	10856.90	347.42
AD011011459R2	1092	273.00	497.95	15.93
Total		11040.25	20137.42	644.40

Table 1: Project ID, GPU times in hours, real node time in hours, mean power consumption including power usage effectiveness (PUE), and CO_2 emissions; for each Jean-Zay project.

computing clusters and infrastructures such as CLEPS[8] and NEF[9], these infrastructures do not allow us to use more than 4 GPUs at a time, thus we consider the power consumption and CO_2 emissions of the experiments conducted in these clusters, negligible in comparison to those of Jean-Zay. Moreover our estimates suppose peak power consumption at all times, which is the worst case scenario and which was clearly not the case at all times for all of our experiments; so we believe this more than compensates the non-reported consumption on both NEF and CLEPS.

Node infrastructure: Each of the Jean-Zay nodes[10] we use consists of 4 GPU Nvidia Tesla V100 SXM2 32GB, 192GB of RAM, and two Intel Xeon Gold 6248 processors. One Nvidia Tesla V100 card is rated at around 300W,[11] while the Xeon Gold 6248 processor is rated at 150W,[12]. For the DRAM we can use the work of [70] to estimate the total power draw of 192GB of RAM at around 20W. Thus, the total power draw of one Jean-Zay node at peak utilization adds up to around 1520W.

With this information, we use the formula proposed by [116] and compute the total power required for each setting:

$$p_t = \frac{1.20t(cp_c + p_r + gp_g)}{1000} \tag{1}$$

Where c and g are the number of CPUs and GPUs respectively, p_c is the average power draw (in W) from all CPU sockets, p_r the average power draw from all DRAM sockets, and p_g the average power draw of a single GPU. We estimate the total power consumption by adding GPU, CPU and DRAM consumption, and then multiplying by the *Power Usage Effectiveness* (PUE), which accounts for the additional energy required to support the compute infrastructure. We use a PUE coefficient of 1.20, which is the value reported by the IDRIS for the Jean-Zay supercomputer. For the real time t we have to divide the reported time for each Jean-Zay project by 4, as Jean-Zay reports the computing time of each project in GPU-hours and not in per node-hours. In table 1 we report the training times in hours, as well as the total power draw (in kWh) of each Jean-Zay project associated to the ALMAnaCH team during 2021. We use this information to compute the total power consumption (multiplying by the PUE) of each project, also reported in table 1.

We can further estimate the CO_2 emissions in kilograms of each single project by multiplying the total power consumption by the average CO_2 emissions per kWh in our region which were around 32g/kWh in average for 2021[13]. Thus the total CO_2 emissions in kg for one single model can be computed as:

$$CO_2e = 0.032p_t \tag{2}$$

All emissions are also reported in table 1. The total emission estimate for the team adds-up to 664.4kg of CO_2. The carbon footprint of a single passenger on a single trip Paris to New York, fighting economy, amounts to around 946kg of CO_2 [14].

[8] CLEPS documentations
[9] NEF documenation
[10] Jean-Zay architecture description
[11] Nvidia Tesla V100 specification
[12] Intel Xeon Gold 6248 specification
[13] Rte - éCO_2mix.
[14] co2.myclimate.org Estimates

6 Highlights of the year

Rachel Bawden was granted a junior (“tremplin”) chair in the PRAIRIE institute. She is the second ALMAnaCH member to hold a PRAIRIE chair, after Benoît Sagot, who holds a PRAIRIE chair since its creation. -

7 New software and platforms

7.1 New software

7.1.1 Enqi

Author: Benoit Sagot

Contact: Benoit Sagot

7.1.2 OSCAR

Name: Open Super-large Crawled ALMAnaCH coRpus

Keywords: Raw corpus, Multilingual corpus

Functional Description: OSCAR is a huge multilingual corpus obtained by language classification and filtering of the Common Crawl corpus using the goclassy architecture.

OSCAR is currently shuffled at line level and no metadata is provided. Thus it is mainly intended to be used in the training of unsupervised language models for natural language processing.

Data is distributed by language in both original and deduplicated form. There are currently 166 different languages available.

Release Contributions: Version 21.09 was generated using Ungoliant version v1, a new generation tool, faster and better documented/tested than the previous one, goclassy, used for OSCAR 1.0 (aka OSCAR 2019). As per OSCAR Schema v1.1, each document/record now has associated metadata. New languages with respect to version 2019: Manx, Rusyn, Scots and West Flemish. Their size and quality still has to be assessed. Removed languages with respect to version 2019: Central Bikol and Cantonese. Cantonsese was of a very low quality. Central Bikol corpus is still available on OSCAR 2019.

URL: https://oscar-corpus.com/

Publications: hal-02148693, hal-03301590, hal-03536361, hal-03177623

Contact: Pedro Ortiz Suarez

Participants: Pedro Ortiz Suarez, Benoit Sagot, Julien Abadji

7.1.3 ACCESS

Keyword: Text Simplification

Functional Description: Text simplification aims at making a text easier to read and understand by simplifying grammar and structure while keeping the underlying information identical. It is often considered an all-purpose generic task where the same simplification is suitable for all, however multiple audiences can benefit from simplified text in different ways. We adapt a discrete parametrization mechanism that provides explicit control on simplification systems based on Sequence-to-Sequence models. As a result, users can condition the simplifications returned by a model on attributes such as length, amount of paraphrasing, lexical complexity and syntactic complexity. We also show that carefully chosen values of these attributes allow out-of-the-box

Sequence-to-Sequence models to outperform their standard counterparts on simplification benchmarks. Our model, which we call ACCESS (as shorthand for AudienCe-CEntric Sentence Simplification), establishes the state of the art at 41.87 SARI on the WikiLarge test set, a +1.42 improvement over the best previously reported score.

URL: https://github.com/facebookresearch/access

Publication: hal-02445874

Contact: Louis Martin

Participants: Louis Martin, Benoit Sagot, Éric De La Clergerie, Antoine Bordes

7.1.4 ASSET

Keyword: Text Simplification

Functional Description: In order to simplify a sentence, human editors perform multiple rewriting transformations: they split it into several shorter sentences, paraphrase words (i.e. replacing complex words or phrases by simpler synonyms), reorder components, and/or delete information deemed unnecessary. Despite these varied range of possible text alterations, current models for automatic sentence simplification are evaluated using datasets that are focused on a single transformation, such as lexical paraphrasing or splitting. This makes it impossible to understand the ability of simplification models in more realistic settings. To alleviate this limitation, this paper introduces ASSET, a new dataset for assessing sentence simplification in English. ASSET is a crowdsourced multi-reference corpus where each simplification was produced by executing several rewriting transformations. Through quantitative and qualitative experiments, we show that simplifications in ASSET are better at capturing characteristics of simplicity when compared to other standard evaluation datasets for the task. Furthermore, we motivate the need for developing better methods for automatic evaluation using ASSET, since we show that current popular metrics may not be suitable when multiple simplification transformations are performed.

URL: https://github.com/facebookresearch/asset

Publication: hal-02889823

Contact: Louis Martin

Participants: Louis Martin, Benoit Sagot, Éric De La Clergerie, Antoine Bordes, Fernando Alva-Manchego, Lucia Specia, Carolina Scarton

7.1.5 EASSE

Keyword: Text Simplification

Functional Description: We introduce EASSE, a Python package aiming to facilitate and standardise automatic evaluation and comparison of Sentence Simplification (SS) systems. EASSE provides a single access point to a broad range of evaluation resources: standard automatic metrics for assessing SS outputs (e.g. SARI), word-level accuracy scores for certain simplification transformations, reference-independent quality estimation features (e.g. compression ratio), and standard test data for SS evaluation (e.g. TurkCorpus). Finally, EASSE generates easy-to-visualise reports on the various metrics and features above and on how a particular SS output fares against reference simplifications. Through experiments, we show that these functionalities allow for better comparison and understanding of the performance of SS systems.

URL: https://github.com/feralvam/easse

Contact: Louis Martin

7.1.6 tseval

Keyword: Text Simplification

Functional Description: The evaluation of text simplification (TS) systems remains an open challenge. As the task has common points with machine translation (MT), TS is often evaluated using MT metrics such as BLEU. However, such metrics require high quality reference data, which is rarely available for TS. TS has the advantage over MT of being a monolingual task, which allows for direct comparisons to be made between the simplified text and its original version. In this paper, we compare multiple approaches to reference-less quality estimation of sentence-level text simplification systems, based on the dataset used for the QATS 2016 shared task. We distinguish three different dimensions: grammaticality, meaning preservation and simplicity. We show that n-gram-based MT metrics such as BLEU and METEOR correlate the most with human judgment of grammaticality and meaning preservation, whereas simplicity is best evaluated by basic length-based metrics.

URL: https://github.com/facebookresearch/text-simplification-evaluation

Contact: Louis Martin

7.1.7 PAGnol

Keywords: Language model, French, Text generation

Functional Description: PAGnol is a collection of large French language models, geared towards free-form text generation. With 1.5 billion parameters, PAGnol-XL is the largest model available for French. PAGnol is based on the GPT-3 architecture with some GPT-2 specific components, and uses scaling laws predictions for efficient training. Using scaling laws, we efficiently rain PAGnol-XL (1.5B parameters) with the same computational budget as much smaller Bert-based models for French. PAGnol-XL is the largest model trained to date for the French language.

Contact: Djame Seddah

Partner: Lighton

7.1.8 PFSMB

Name: Parallel French Social Media Bank

Keywords: Machine translation, User-generated content, Social medias

Functional Description: The PFSMB is a collection of French-English parallel sentences manually translated from an extension of the French Social Media Bank (Seddah et al., 2012)] which contains texts collected on Facebook, Twitter, as well as from the forums of JeuxVideos.com and Doctissimo.fr. This corpus, consists of 1,554 comments in French annotated with different kind of linguistic information: Part-of-Speech tags, surface syntactic representations, as well as a normalized form whenever necessary. Comments have been translated from French to English by a native French speaker and extremely fluent, near-native, English speaker. Typographic and grammatical error were corrected in the gold translations but the language register was kept. For instance, idiomatic expressions were mapped directly to the corresponding ones in English (e.g. 'mdr' has been translated to 'lol' and letter repetitions were also kept (e.g. 'ouiii' has been translated to yesss').

Publications: hal-02270524, hal-00780895

Contact: Djame Seddah

7.1.9 EtymDB

Name: Etymological DataBase

Keyword: Lexicon

Functional Description: EtymDB is an etymological database automatically extracted from wiktionary, available in several formats (TSV, XML/TEI).

Release Contributions: Extraction from a more recent version of wiktionary, improvement of the extraction process.

URL: https://files.inria.fr/almanach/software_and_resources/default/EtymDB-en.html

Publications: hal-02678100, hal-01592061, hal-01584013

Contact: Benoit Sagot

Participants: Benoit Sagot, Clementine Fourrier

7.1.10 KaMI-Lib

Name: KaMI (Kraken Model Inspector) - Python Library

Keywords: HTR, OCR, Python, Handwritten Text Recognition, Image segmentation, Library

Functional Description: KaMI-lib (Kraken as Model Inspector) is a Python library for evaluating transcription models (handwritten text recognition and optical character recognition) trained either with the Kraken engine (http://kraken.re) or without it.

It provides a single class for comparing strings (e.g. extracted from text files) and for generating scores in order to evaluate the automatic transcription's performance. The Kraken engine is implemented in KaMI-lib in order to produce a prediction with a pre-trained transcription model and to compare it to a ground truth (in PAGE XML or XML ALTO format) associated with its image.

KaMI-lib uses different metrics to evaluate a transcription model: the Word Error Rate (WER), the Character Error Rate (CER), and the Word Accuracy (Wacc). In addition, KaMI-lib provides the edit distances and the operations performed on the different strings. It is also possible to weigh the cost of operations in order to adjust scores.

It is also possible to get different scores with text pre-processing functions applied to the ground truth and the prediction, such as deleting all diacritics, ponctuations, or numbers, ignoring upper case, etc. By doing so, KaMI-lib aims to give a better understanding of text features' impacts on transcription results. This functionality also aims to make users adapt the creation of training data according to their texts' specificites, and optimize the training process.

Documentation is available here: https://gitlab.inria.fr/dh-projects/kami/kami-lib

URL: https://gitlab.inria.fr/dh-projects/kami/kami-lib

Publications: hal-03495762, hal-03008579

Contact: Lucas Terriel

Participants: Alix Chague, Lucas Terriel, Hugo Scheithauer

7.1.11 Ungoliant

Name: Ungoliant

Keyword: Natural language processing

Functional Description: Ungoliant is a high-performance pipeline that provides tools to build corpus generation pipelines from CommonCrawl. It currently is the generation pipeline for OSCAR corpus. Ungoliant is a replacement of the goclassy pipeline.

URL: https://github.com/oscar-corpus/ungoliant

Publications: hal-03301590, hal-03536361

Contact: Julien Abadji

Participants: Julien Abadji, Pedro Ortiz Suarez, Benoit Sagot

7.1.12 HTR-United

Keywords: HTR, OCR

Functional Description: HTR-United is a Github organization without any other form of legal personality. It aims at gathering HTR/OCR transcriptions of all periods and styles of writing, mostly but not exclusively in French. It was born from the mere necessity for projects- to possess potentiel ground truth to rapidly train models on smaller corpora.

Datasets shared or referenced with HTR-United must, at minimum, take the form of: (i) an ensemble of ALTO XML and/or PAGE XML files containing either only information on the segmentation, either the segmentation and the corresponding transcription, (ii) an ensemble of corresponding images. They can be shared in the form of a simple permalink to ressources hosted somewhere else, or can be the contact information necessary to request access to the images. It must be possible to recompose the link between the XML files and the image without any intermediary process, (iii) a documentation on the transcription practices followed for the segmentation and the transcription. In the cases of a Github repository, this documentation must be summarized in the README.

A corpus can be sub-diveded into smaller ensembles if it seems necessary.

Release Contributions: First version.

URL: https://htr-united.github.io/

Contact: Alix Chague

-

7.2 New platforms

Over the last couple of years, we have worked towards integrating a variety of our contributions in the domain of Digital Humanities within stable platforms with the idea to make the corresponding environment reusable by other colleagues and institutions. This endeavor is in keeping with the general vision of EU infrastructures such as DARIAH to offer the humanities communities with the best up to date technologies for creating, managing, transforming and disseminating their data. We have put a specific emphasis on the stabilisation of a primary source management platform linking low level handwritten character recognition techniques with final end structured publication, with the following results:

- set up a technical environment within the collaborative project CREMMA to acquire and deploy a GPU based HTR server based upon the eScriptorium/Kraken suite. This environment, shared with the EPHE, LAMOP and Ecole Nationale des Chartes, now allows the latter for instance to provide a high quality environment for all their practical courses on digital epigraphy;

- work on a proposal for adapting the TEI standard to integrate all incoming information from HTR processes within a ready-to-publish document, which so far where only available in a scattered landscape of formats (ALTO XML with various versions, Page XML) [46]. With a change of paradigm, where TEI is given more importance earlier in a digitisation workflow compared to traditional automatic transcription formats, we hope to improve the reusability of content at various stages, and hence ease the combination of different software solutions, such as those used for automatic transcription and edition. We also aim for a better propagation of metadata elements.

- Experiment a generic instance of the open source environment TEI Publisher for the online publication of editions integrating the abovementioned standardisation proposal and combining image, HTR results (based on an implementation of the SVG W3C recommendation), annotations and final end edited results. This work has been selected as part of the TEI Publisher Community Challenge.

8 New results

8.1 Large corpus creation

Participants: Pedro Ortiz Suarez, Julien Abadji, Benoît Sagot.

Since the introduction of large language models in Natural Language Processing, large raw corpora have played a crucial role in Computational Linguistics. However, most of these large raw corpora are either available only for English or not available to the general public due to copyright issues. Nevertheless, there are some examples of freely available multilingual corpora for training Deep Learning NLP models, such as the Paracrawl corpora and our own large-scale multilingual corpus OSCAR [92].[15] However, they have quality issues, especially for low-resource languages, an issue investigated in a large-scale experiment in which we have taken part and whose initial publication in 2021 [54] will be followed by a publication in the *Transactions of the Association for Computational Linguistics*. Moreover, recreating or updating these corpora is very complex. In 2021 we have developed Ungoliant [18], a new pipeline that improves in multiple ways over the goclassy pipeline used to create the original OSCAR corpus (known as OSCAR v1 or OSCAR 2019). Ungoliant is faster, modular, parameterizable, and well documented. We used it to create a new version of OSCAR that is larger and based on recent data [18]. Moreover, this new version of OSCAR (version 21.09) includes metadata information at the document level. Ungoliant is reseased under an open source license and publish the corpus under a research-only license. Since then, we have achieved a new set of improvements and automatic annotations in order to produce the latest version of OSCAR [51].

8.2 Neural language modelling

Participants: Benoît Sagot, Djamé Seddah, Rachel Bawden, Arij Riabi, Thomas Wang, Roman Castagné.

Pretrained language models are now ubiquitous in Natural Language Processing. Despite their success, most available models have either been trained on English data or on the concatenation of data in multiple languages [71, 85]. This makes practical use of such models—in all languages except English—very limited. One of the most visible achievements of the ALMAnaCH team was the training and release of CamemBERT in 2019, a BERT-like [71] (and more specifically a RoBERTa-like) neural language model for French trained on the French section of our large-scale web-based OSCAR corpus [92] (see Section 8.1), together with CamemBERT variants [87] and ELMo models trained on OSCAR corpora for other languages, including French [90, 91].

[15] OSCAR web site.

In 2021, we investigated the impact on language modelling of shifting from token-based or subword-based models to character-based models, which were reported to improve model robustness under certain circumstances. We carried out a number of experiments on a number of such models proposed by other authors. In particular, we investigated the performance of character-based language models on North-African Arabizi (NArabizi), i.e. North-African colloquial dialectal Arabic written using an extension of the Latin script, a low-resource language variety on which ALMAnaCH has been working for several years. We show that a character-based model trained on only 99k sentences of NArabizi and fined-tuned on a small treebank of this language leads to performance close to those obtained with the same architecture pre-trained on large multilingual and monolingual models [27]. We also confirmed these results a on much larger data set of noisy French user-generated content.

We also carried out a number of experiments on previously proposed and novel ways to approximate the attention mechanism in Transformer-based language models, which is quadratic, with a linear-time mechanism.

In collaboration with the startup LightOn, we trained and published PAGnol, a collection of GPT-like generative language models for French. Using scaling laws, we efficiently train PAGnol-XL (1.5B parameters) with the same computational budget as CamemBERT, which was 13 times smaller [56]. PAGnol-XL is the largest model trained to date for the French language, and the first large-scale generative model for the language. Experiments shown that such a model, when trained on our large web-based corpus OSCAR (2019 version), can produce offensive content. It is one of the reasons that led us to develop detection tools to flag certain types of possibly offensive content in further versions of OSCAR.

Finally, a number of ALMAnaCH members have participated in the BigScience "workshop",[16] an informal 1-year-long international initiative led by the American company HuggingFace and involving 600 researchers from 50 countries and more than 250 institutions. The goal of the project, supported by the French state in the form of large grants of computing power on the Jean Zay public supercomputer, is to create a publicly available very large multilingual neural network language model and a publicly available very large multilingual text dataset. ALMAnaCH members are involved at all levels: working group chair, sub-working group chair, working group active member. First results have already been published. The working group on tokenisation, of which Benoît Sagot is one of the two chairs, has published a survey on the matter [58]. The first language model of interest successfully trained and evaluated within the project has also resulted in a publication, to which some of ALMAnaCH's member involved in the project have contributed [29].

8.3 Cross-lingual transfer learning for low-resource non-standard languages

Participants: Djamé Seddah, Benoît Sagot, Benjamin Muller.

Building NLP systems for such highly variable and low-resource languages is a difficult challenge. The recent success of large-scale multilingual pretrained neural language models (including our CamemBERT language model for French) provides us with new modelling tools to tackle it. We have studied the ability of the multilingual version of BERT to model a non-formalised North-African dialect, Arabizi, written in Latin script and often code-mixed with French, which has been a research topic for the team for several years—in particular, we have published in 2019 a treebank of this language variety, the first UGC dataset for an Arabic-related language [110]. We have shown in different scenarios that multilingual language models are able to transfer to such an unseen dialect, specifically in two extreme cases: across scripts (Arabic to Latin) and from Maltese, a related language written in the Arabic script, unseen during pretraining. The first results have led to fruitful collaborations with Yanai Elazar from Bar Illan University and Antonios Anastasopoulos from George Mason University. These works, expanding from North-African Arabic to many other languages that share similar properties, focus on the objective of understanding the inner workings of large multilingual neural language models when fine-tuned on cross-language scenarios with various degres of resource availability (from large to extremely scarce). Conducted in 2020, those works led to the publications of two papers in 2021 [59, 25], following two publications in 2020 [121, 115].

[16] BigScience web site.

8.4 Multimodal word and sentence embeddings

Participants: Matthieu Futeral-Peter, Rachel Bawden, Benoît Sagot.

In collaboration with Ivan Laptev and Cordelia Schmid from the WILLOW project-team, we have investigated how using image data can improve the representation of words and sentences. As part of Matthieu Futeral-Peter's master's internship, we study how image captioning data can be used as a pivot to align multilingual embeddings, both on the word and the sentence level, notably exploring extensions to the GlobeTrotter model [117], such as the creation of synthetic data through automatic image captioning and the addition of a nearest neighbours search to expand the number of image-caption pairs. This work is a preliminary step in the path toward multimodal machine translation, which will be one of the main subjects of Matthieu's PhD.

8.5 NLP for Early Modern French

Participants: Pedro Ortiz Suarez, Rachel Bawden, Benoît Sagot, Laurent Romary, Alix Chagué.

Early Modern French (also known as Modern French or classicial French) represents the French language from the 17th century. With the aim of helping philologists and literary experts study texts from this period, we have been collaborating with several researchers outside Inria (Simon Gabay from the University of Geneva and Philippe Gambette from Université Gustave-Eiffel) to develop various NLP tools adapted to Early Modern French. Aside the differences in topic and word choice, the texts display linguistic differences from contemporary French, including spelling differences, which encompass both typographic differences (such as the use of long s, which has become *s* in contemporary French) and those illustrative of linguistic change (*estoit→était*) and classicial influences (*sçauoir→savoir*). The NLP tools we develop must be adapted to these spelling differences and also be robust to variation, giving that no strict conventions were used during that period.

In 2021, we continued our work on various aspects of the processing of Early Modern French: historical spelling normalisation, named entity recognition, part-of-speech (PoS) tagging. These tasks rely on data from the period, and a major part of our efforts has been in the creation of adapted corpora: the FreEM corpus (short for French Early Modern), which includes a large monolingual corpus (FreEM-*max*), a smaller parallel corpus of sentences normalised into contemporary French spelling conventions (FreEm-*norm*) and a dataset annotated for locations [42].

From the FreEM-*max* data, we have also trained a RoBERTa language model [85] for Early Modern French, which we call D'AlemBERT, which will be used to boost the performance across various tasks. It has already be shown to boost the performance of PoS tagging for Early Modern French. From the FreEM-*norm* data, we have developed automatic normalisation models, including rule-based and machine-translation (MT)-style approaches. As well as providing normalisation, which can be a useful step before manual analysis or the application of other downstream tasks, we have exploited the normalisation models as a way to generate synthetic Early Modern French data on a large scale, which we then use to analyse the patterns of spelling change across different decades of the 17th century [41].

The NLP tools developed will be integrated into a processing pipeline for Early Modern French texts and encoded in TEI [32]. Articles presenting the FreEM-*max* and FreEM-*norm* corpora, the D'AlemBERT language model (and results on PoS tagging) and the normalisation models [52] are currently under peer-review at the LREC 2022 conference.

8.6 Cognate prediction vs. low-resource Machine Translation

Participants: Clémentine Fourrier, Benoît Sagot, Rachel Bawden.

In 2021 we resumed our experiments to investigate whether and under which conditions neural networks can be used to learn sound correspondences between two related languages, i.e. for the prediction of cognates of source language words in a related target language. Since the experiments published in [73], we have improved both our way to extract data from our etymological database EtymDB 2.0 [74] and our neural systems in multiple ways, leading to improved results published in [22].

More precisely, we investigated whether cognate prediction can benefit from insights from low-resource MT. We compared statistical MT (SMT) and neural MT (NMT) architectures in a bilingual setup, and studied the impact of employing data augmentation techniques commonly seen to give gains in low-resource MT: monolingual pretraining, backtranslation and multilinguality—on these techniques, see for instance [55], a survey on low-resource MT of which a member of the team is one of the main authors,[17] as well as [19].

Our experiments on several Romance languages show that cognate prediction behaves only to a certain extent like a standard lowresource MT task. In particular, MT architectures, both statistical and neural, can be successfully used for the task, but using supplementary monolingual data is not always as beneficial as using additional language data, contrarily to what is observed for MT.

8.7 Machine Translation of non-standard texts

Participants: Djamé Seddah, Rachel Bawden, Jose Rosales Nunez, Camille Rey, Lydia Nishimwe, Sonal Sannigrahi.

Machine translation (MT) in an increasingly important research topic for ALMAnaCH, which has been reinforced since the arrival of Rachel Bawden as a *Chargée de Recherches* in November 2020. Consistent with the team's main research challenges, understanding the impact of language variation on MT models and trying to improve the robustness of these models to language variation are among our key research topics.

In the context of a PhD thesis co-supervised by Djamé Seddah with the Laboratoire de Linguistique Formelle (Université de Paris), we have taken a critical look at the automatic evaluation of user-generated content (UGC). TheWe have shown that measuring the average-case performance using a standard metric on a UGC test set falls far short of giving a reliable image of the UGC translation quality. We therefore introduced PMUMT,[18] a new manually annotated dataset for the evaluation of UGC translation. We conducted several experiments on this dataset to measure the impact of different kinds of UGC specificities on translation quality, more precisely than previously possible [28].

Inspired by research on character-level MT and on robustness in character-level language models, we also explored the ability of character-based Neural MT to translate noisy UGC with a strong focus on exploring the limits of such approaches to handle productive UGC phenomena, which cannot be seen at training time. Within a strict zero-shot scenario, we studied the detrimental impact on translation performance of various user-generated content phenomena on a small annotated dataset we developed and showed that such models are indeed incapable of handling unknown characters, which leads to catastrophic translation failure once such characters are encountered. We also showed the importance of reducing the vocabulary size hyper-parameter to increase the robustness of character-based models for machine translation [26]. On this topic and in the context of an internship supervised by Rachel Bawden, we are also investigating the robustness of character-level and byte-level models for multilingual MT, whereby we will apply various perturbations to input representations (e.g. transliteration, character-shifting) to compare the impact of lexical sharing in the performance of these models.

More recently, we have developed a novel Variational Neural MT (VNMT) architecture with enhanced robustness properties, which we investigated through a detailed case-study addressing noisy French user-generated content (UGC) translation to English. The proposed model, with results comparable or superior to state-of-the-art VNMT, improves performance over UGC translation in a zero-shot evaluation scenario while keeping optimal translation scores on in-domain test sets. This work echoes the work of

[17] This paper is currently a preprint and was submitted to the *Computational Linguistics* journal, which "conditionnaly accepted" it for publication.

[18] PMUMT GitHub site.

a PhD student at LIPN on the application and extension of Variational Auto-Encoders, with whom we collaborate [21, 40].

The arrival of Rachel Bawden at the end of 2020, a specialist of MT and specifically low-resource and robust MT, who published in 2021 a survey paper on low-resource MT [55],[19] followed by the arrival two new young researchers at the end of 2021, a PhD student and a Master's student, has reinforced the team's interest in these questions and, more generally, on robustness and domain adaptation in MT. In particular, we have started the design and development of a new corpus of English UGC where tweets of varied non-standardness levels are aligned with a normalised version and a (standard) translation in French, in order to submit a new shared task for the WMT 2022 conference.

8.8 New results on text simplification

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

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

In 2017 we initiated a collaboration with the Facebook Artificial Intelligence Research (FAIR) lab in Paris and with the UNAPEI, the federation of French associations helping people with mental disabilities and their families. The objective of this collaboration is to develop tools to help the simplification of texts aimed at mentally disabled people. More precisely, the aim is to develop a computer-assisted text simplification platform (as opposed to an automatic TS system). In this context, a CIFRE PhD thesis was initiated in collaboration with the FAIR on the TS task, Louis Martin being the PhD student involved.

In 2020 and 2021, this PhD thesis extended the scope of TS to multiple languages using a totally unsupervised method and with state-of-the-art results, even compared to previous supervised models [57]. We use a large-scale mining pipeline to extract one billion sentences from the web using Common Crawl. These sentences are used to find millions of paraphrases in three languages: English, French and Spanish. This allows us to train our controllable models [88] in multiple languages, which we then adapt for the TS task. These models push the state-of-the-art further in all three languages and benchmarks. Generated simplifications are considered more fluent, and simpler than previous models.

In 2021, we also developed a French-specific simplification model based on our pre-trained language model CamemBERT [4] and compared its results with the above-mentioned unsupervised models and weakly supervised variants thereof. We also investigated a new way to address the issue of the evaluation of sentence simplification, which had been the focus of previous efforts and publications [86]. This new method is based on previous work on the evaluation of automatic summarisation models based on the ability of the models output to successfully allow for a question answering model to answer questions the input text would allow to answer [108]. Adapting this approach to the evaluation of sentence simplification proved successful [60].

This research enables us to exploit our previous works and apply them to the Cap'FALC project, whose aim is to assist the simplification of French documents by people with intellectual disabilities.

2021 was also an important year for this line of research, since Louis Martin defended his PhD in September. The Cap'FALC project, however, continues, and we hope to resume working on this socially important and scientifically challenging topic.

8.9 Threads Constitution

[19] This paper is currently a preprint and was submitted to the *Computational Linguistics* journal, which "conditionnaly accepted" it for publication.

Participants: Lionel Tadjou Tadonfouet, Éric Villemonte de La Clergerie, Laurent Romary.

In the context of a CIFRE PhD with Orange, we have developed a corporate corpus that will be used as a reference for modelling and computing threads from conversations generated using communication and collaboration tools. The overall goal of the reconstruction of threads is to higlight the important parts of a running discussion, reviewing the upcoming commitments or deadlines, etc. Since, to our knowledge, there were no available corporate corpus for the French language, we developed a method for building such corpora, including a pseudo-anonymisation step, in compliance with the GDPR [30].

8.10 Hate speech detection

Participants: Djamé Seddah, Syrielle Montariol, Arij Riabi.

In order to support the fight against radicalization and thus prevent future terrorist attacks from taking place, the CounteR H2020 project brings data from disperse sources into an analysis and early alert platform for data mining and prediction of critical areas (e.g. communities), aiming to be a frontline community policing tool which looks at the community and its related risk factors rather than targeting and surveilling individuals. ALMAnaCH is responsible of the most part of the NLP component of the project. However, issues regarding the availability of the input data have prevented us from working specifically on CounteR data. We have therefore focused on a task of huge social importance that is similar to CounteR's needs, namely hate speech detection, with a focus on zero-shot cross-lingual transfer.

In our work in this direction, unpublished to date, we investigate a key limitation of cross-lingual transfer approaches to such tasks, namely the fact that they involve numerous linguistic and cultural differences and discrepencies from one language or one country to the other. We investigated how training on multilingual auxiliary tasks (sentiment analysis, named entity recognition, and tasks relying on syntactic information) impacts the zero-shot transfer of hate speech detection models across languages. We showed the positive impact of these tasks, particularly named entity recognition, for bridging the gap between languages. But we also shown that, in some cases, the language model training data prevents hate speech detection models from benefiting from the *knowledge proxy* brought by auxiliary tasks fine-tuning.

8.11 Similar case detection for the *Cour de Cassation*

Participants: Thibault Charmet, Rachel Bawden, Benoît Sagot.

As part of our LabIA project, funded by the DINUM, in collaboration with the *Cour de Cassation*, we have developed tools for (i) the automatic generation of keyword sequences (*titrages*), which are currently used by experts at the *Cour* to detect rulings that are similar in their application of the law, and (ii) the development of various similarity measures, which can be used to find similar documents. The similarity measures we use also integrate the keyword sequences predicted by our models, with improved correlations to expert judgments of similarity. This is important because it enables us to increase the coverage of these keyword sequences, which are currently only available for about 20% of all rulings, and to increase the number of sequences per ruling, since the granularity of the sequences and the word choice is highly variable, which otherwise limits the recall of a similarity search process. To validate results, similarity measures are compared against similarity judgements between pairs of keyword sequences that we manually collected with experts from the *Cour*. The motivations, methods and results have been submitted for peer review at the LREC 2022 conference.

8.12 Models for the representation of lexical content

Participants: Laurent Romary.

For several years, the ALMAnaCH team has taken a leadership role in defining standards for representing lexical content, either as a result of digitising legacy dictionaries or through the creation of new lexical resources serving as a basis for computational linguistics processes.

The bridge between the general ISO standardisation work and the community-based TEI guidelines has been further developed in the context of the DARIAH working group on lexical resources, whose objective consists in identifying univocal TEI-based constructs for a variety of lexical features encountered in machine readable dictionaries. The work carried out by the group (see web site), whose quality was recognised by the TEI community as a whole with the Rahtz Prize for TEI Ingenuity, has been and implemented in several major dictionary endeavours worldwide as exemplified in [79] and, in 2021, [20]. In the context of the ELEXIS project, we also took part in the publication of a reference document on best practices in electronic lexicographic resource development, the Lexicographic Data Seal of Compliance [61].

Regarding our long-lasting efforts on the encoding of etymological information in the TEI has resulted in the publication of a description of TEI Lex-0 Etym, the etymological component of the TEI Lex-0 initiative, which aims at defining a terser subset of the TEI guidelines for the representation of etymological features in dictionary entries [53]. Going beyond the basic provision of etymological mechanisms in the TEI guidelines, TEI Lex-0 Etym proposes a systematic representation of etymological and cognate descriptions by means of embedded constructs based on the (for etymologies) and (for etymons and cognates) elements. In particular, given that all the potential contents of etymons are highly analogous to those of dictionary entries in general, the contents presented herein heavily re-use many of the corresponding features and constraints introduced in other components of the TEI Lex-0 to the encoding of etymologies and etymons. The TEI Lex-0 Etym model is also closely aligned to ISO 24613-3 on modelling etymological data and the corresponding TEI serialisation available in ISO 24613-4.

8.13 Information extraction from specialised collections

Participants: Laurent Romary, Alix Chagué, Floriane Chiffoleau, Lucas Terriel, Hugo Scheithauer, Tanti Kristanti, Yves Tadjo.

Building up on our long-standing contribution to the development of the GROBID suite, the ALMAnaCH team has pursued its effort in extending the scope and performance of the GROBID modules in relation to various ongoing collaboration schemes.

In the continuity of Mohamed Khemakhem's PhD thesis [80], which has demonstrated the possibility to parse lexical entries from legacy dictionary content in a very fine-grained way, we investigated the applicability of the same architecture on catalogue-like textual objects, as previsouly demonstrated in [82] and in 2021 in [23] on the example of exhibition catalogues. The DataCatalog project, jointly led with the Bibliothèque nationale de France (BnF) and the Institut national d'histoire de l'art (INHA), was launched in late 2021. It aims to develop a GROBID module for automatically structuring catalogues in TEI-XML, and to publish files produced with it on an interface and allowing visualizations and queries on segmented zones. We hope to give researchers a new way to access these documents where accessing the information is made easier and more accurate. The BnF and the INHA provide a vast collection of sale catalogues on which the custom TEI model created for the project is based.

The occurrence of the COVID pandemic in Spring 2020 has also been an opportunity to start a project with the Parisian hospital network (APHP). One of the goals of the project (and the action that ended up being the main focus of our attention) was to see how the GROBID suite could be further expanded to parse the variety of medical reports and documents associated with a patient so that doctors can trace the precise relevant information (anamnesis, symptoms, treatments etc.). The excellent results obtained

in 2020 have led us to build up a stable collaboration with the APHP to make the GROBID workflow an essential building block of their document processing chain.

Following the provision of a direct grant from the Ministry of Higher Education and Research (MESRI - DAHN project) and benefiting from a collaborative framework with the French National Archives (Lectaurep project [34, 35]), we have explored how our experience in extracting information from legacy documents could be made part of wider understanding of the components of a generic digitisation workflow of documents initially available as images [39, 36]. In particular, in collaboration with colleagues from the ÉPHÉ (École Pratique des Hautes Études), we contributed to the further development of the eScriptorium platform, which aims to provide an annotation, training and transcription environment for handwritten text recognition (HTR) tasks, for any language and any writing system, down to the understanding of how to encode the data at hand in the TEI format [46] and publish it in the TEI Publisher environment [47].

We have made contributions in a number of complementary directions, including the following:

- We have defined methods for the coherent management and sharing of ground truth, i.e reference annotated data with potential interest across a variety of time periods, languages, writing systems and domains. To this end, HTR-United (web site) aims at gathering various annotated corpora along with their metadata for the creation of HTR models [37];
- We have worked on the development of several refined HTR models and carefully evaluated its performance levels thanks to KaMI (Kraken as Model Inspector) [45, 48];
- We have worked in the context of the NER4archives project on semi-automatic methods and tools to recognise named entities in archival research instruments encoded in XML/EAD [38];
- We have worked on the automatic transcription and edition of digitised sources on work in the textile industry [43]; and of the digitised WWI correspondance of Paul d'Estournelles de Constant [39]. We have also updated and complemented the encoding and edition of several corporas including the *Berliner Intelektuelle and testimonies from the Holocaust* (web site).

9 Bilateral contracts and grants with industry

Participants: Benoît Sagot, Rachel Bawden, Djamé Seddah, Éric Villemonte de La Clergerie, Louis Martin, Tu Anh Nguyen, Paul-Ambroise Duquenne, You Zuo, Thibault Charmet.

9.1 Bilateral contracts with industry

Ongoing contracts:

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

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

Facebook A collaboration on text simplification ("français Facile À Lire et à Comprendre", FALC) is ongoing with Facebook's Parisian FAIR laboratory. It involved 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 PhD thesis was defended in 2021. This collaboration is part of a larger initiative called Cap'FALC involving (at least) these three partners as well as the relevant ministries. Funding received as a consequence of the CIFRE PhD thesis: 60,000 euros Moreover, two new CIFRE PhD theses on other topics (1. language modelling applied to speech conversational data; 2. sentence-level vector representations) have started in 2021.

Orange

Winespace The collaboration with this start-up company, dedicated to information extraction from wine descriptions to develop a wine recommendation system, was carried out in 2020 following previous discussions, in collaboration with Inria Bordeaux's "InriaTech" structure. In 2021, we designed and prepared a second step for this collaboration, which will involve a dedicated research engineer who will begin their work in 2022. We also presented the results of the collaboration's first step at a regional data science event, Dataquitaine [31].

INPI A collaboration with the Institut National de la Propriété Industrielle (France's patent office) started in 2021. A research engineer was hired for a one-year contract to work on patent classification. This project informally collaborates with the qatent startup, on which see below.

Cour de cassation A LabIA project started in early 2021. A research engineer was hired for a one-year contract to work on the automatic analysis of Cour de Cassation rulings and on the automatic assessment of the similarity between two rulings.

9.2 Active collaborations without a contract

Science Miner ALMAnaCH (following ALPAGE) has collaborated since 2014 with this company founded by Patrice Lopez, a specialist in machine learning techniques and initiator of the Grobid and NERD (now entity-fishing) suites. Patrice Lopez provides scientific support for the corresponding software components in the context of the Parthenos, EHRI and Iperion projects, as well as in the context of the Inria anHALytics initiative, aiming to provide a scholarly dashboard on scientific papers available from the HAL national publication repository.

qatent The startup company qatent, dedicated to computer-aided patent creation, was created in 2021. It is supported by the Inria Startup Studio and by the ALMAnaCH team. Regular interactions take place between qatent founders and members on the one hand and ALMAnaCH members on the other hand, in particular those involved in the collaboration with INPI. This creates a stimulating informal collaboration between all three entities around NLP for patents, which might foster further activity in this domain (e.g. a future PhD thesis).

LightON LightON builds Optical Processor Units, a specialized line of processor able to outperform GPU on certains tasks. We're working with them to see if we can use their technology to speed up the training of large language models. We recently submitted a grant proposal to access 250k gpu hours on Jean Zay in order to scale their algorithms. This informal collaboration with this company has already resulted in the design, training and publication of the PAGnol generative language model for French (cf. PAGnol's web site).

AXA ReV AXA ReV is the R&D Lab of the Axa Insurance group, located in Paris. This collaboration focuses on neural models interpretability and establishing "explainable" benchmarks as a end-goal for research on question answering.

10 Partnerships and cooperations

10.1 European initiatives

10.1.1 FP7 & H2020 Projects

H2020 EHRI "European Holocaust Research Infrastructure"

Duration: 1 May 2015–31 Aug 2024.

PI: Conny Kristel (NIOD-KNAW, NL).

Coordinator for ALMAnaCH: Laurent Romary.

Partners:
- Archives Générales du Royaume et Archives de l'État dans les provinces (Belgium)
- Aristotelio Panepistimio Thessalonikis (Greece)
- Dokumentačné Stredisko Holokaustu Občianske Združenie (Slovakia)
- Fondazione Centro Di Documentazione Ebraica Contemporanea -CDEC - ONLUS (Italy)
- International Tracing Service (Germany)
- Kazerne Dossin Memoriaal, Museum Endocumentatiecentrum Over Holocausten Mensenrechten (Belgium)
- Koninklijke Nederlandse Akademie Van Wetenschappen - KNAW (Netherlands)
- Magyarorszagi Zsido Hitkozsegek Szovetsege Tarsadalmi Szervezet (Hungary)
- Masarykův ústav a Archiv AV ČR, v. v. i. (Czech Republic)
- Memorial de La Shoah (France)
- Stiftung Zur Wissenschaftlichen Erforschung Der Zeitgeschichte - Institut Fur Zeitgeschichte IFZ (Germany)
- Stowarzyszenie Centrum Badan Nad Zaglada Zydow (Poland)
- The United States Holocaust Memorial Museum (United States)
- The Wiener Holocaust Library (UK)
- Vilniaus Gaono žydų istorijos muziejus (Lithuania)
- Wiener Wiesenthal Institut Fur Holocaust-Studien - VWI (Austria)
- Yad Vashem The Holocaust Martyrs And Heroes Remembrance Authority (Israel)
- Židovské muzeum v Praze (Czech Republic)
- Żydowski Instytut Historyczny im. Emanuela Ringelbluma (Poland)

Summary: Transforming archival research on the Holocaust, by providing methods and tools to integrate and provide access to a wide variety of archival content.

H2020 CounteR

Duration: 1 May 2021–30 Apr 2024.

PI: Catalin Truffin.

Coordinator for ALMAnaCH: Djamé Seddah.

Partners:
- Assist Software SRL (Romania)
- Insikt Intelligence S.L. (Spain)
- IMAGGA Technologies LTD (Bulgaria)
- Icon Studios LTD (Malta)
- Consorzio Interuniversitario Nazionale per l'Informatica (Italy)
- Eötvös Loránd Tudományegyetem (Hungary)
- Universita Cattolica del Sacro Cuore (Italy)
- Malta Information Technology Law Association (Malta)
- European Institute Foundation (Bulgaria)
- Association Militants des Savoirs (France)

- Eticas Research and Consulting S.L. (Spain)
- Elliniki Etairia Tilepikoinonion kai Tilematikon Efarmogon AE (Greece)
- Ministério da Justiça (Portugal)
- Hochschule für den Öffentlichen Dienst in Bayern (Germany)
- Iekslietu Ministrijas Valsts Policija [State Police Of The Ministry Of Interior] (Latvia)
- Serviciul de Protectie si Paza (Romania)
- Glavna Direktsia Natsionalna Politsia (Bulgaria)
- Ministère de l'Intérieur (France)

Summary: In order to support the fight against radicalization and thus prevent future terrorist attacks from taking place, the CounteR project brings data from disperse sources into an analysis and early alert platform for data mining and prediction of critical areas (e.g. communities), aiming to be a frontline community policing tool which looks at the community and its related risk factors rather than targeting and surveilling individuals. The system will incorporate state of the art NLP technologies combined with expert knowledge into the psychology of radicalization processes to provide a complete solution for law enforcement authorities to understand the when, where and why of radicalization in the community.

10.1.2 Other European programs/initiatives

ERIC DARIAH

Duration: 1 Sep 2014–31 Aug 2034.

Coordinator for ALMAnaCH: Laurent Romary.

Summary: Coordinating Digital Humanities infrastructure activities in Europe (17 partners, 5 associated partners). L. Romary is a former president of DARIAH's board of director.

10.2 National initiatives

10.2.1 ANR

ANR ParSiTi

Duration: 1 Nov 2016–31 Mar 2022.

PI: Djamé Seddah.

Coordinator for ALMAnaCH: Djamé Seddah.

Partners:
- LISN
- LIPN

Summary: Context-aware parsing and machine translation of user-generated content.

ANR BASNUM

Duration: 1 Oct 2018–30 Jun 2023.

PI: Geoffrey Williams (Université de Grenoble).

Coordinator for ALMAnaCH: Laurent Romary.

Partners:
- Université de Bretagne Sud
- Université Grenoble Alpes
- LaTTICe

Summary: Digitalisation and computational annotation and exploitation of Henri Basnage de Beauval's encyclopedic dictionary (1701).

ANR PARSE-ME

Duration: 1 Oct 2015–30 Sep 2021.

PI: Matthieu Constant (ATILF).

Coordinator for ALMAnaCH: Djamé Seddah.

Summary: Multi-word expressions in parsing.

ANR Profiterole

Duration: 1 Oct 2017–31 Jan 2021.

PI: Sophie Prévost (LaTTICe).

Coordinator for ALMAnaCH: Éric de La Clergerie.

Partners:
- LaTTICe
- LLF
- IRHIM

Summary: Modelling and analysis of Medieval French. ALMAnaCH members are associated to LLF (U. de Paris) for this project.

ANR TIME-US

Duration: 1 Oct 2016–31 Dec 2021.

PI: Manuela Martini (LARHRA).

Coordinator for ALMAnaCH: Éric de La Clergerie.

Partners:
- LARHRA
- TELEMMe
- Labo ICT
- IRHIS
- Centre Maurice Halbwachs-EHESS

Summary: Digital study of remuneration and time budget textile trades in XVIIIth and XIXth century France. ALMAnaCH members are associated to CEDREF (U. de Paris) for this project.

ANR CulturIA

Duration: 1 Nov 2021–31 Oct 2024.

PI: Alexandre Gefen.

Coordinator for ALMAnaCH: Benoît Sagot.

Partners:
- UMR THALIM
- UPR Centre Internet et Société

Summary: This project aims at building a cultural history of Artificial Intelligence, based on a mixed method, combining the methods of the history of ideas and the history of collective imaginations with a search of scientific literature and ethnographic field work among AI creators.

3IA PRAIRIE

Duration: 1 Oct 2019–31 Dec 2023.

PI: Isabelle Ryl.

Coordinators for ALMAnaCH: Benoît Sagot and Rachel Bawden.

Partners:
- Inria
- CNRS
- Institut Pasteur
- PSL
- Université de Paris
- Amazon
- Google DeepMind
- Facebook
- faurecia
- GE Healthcare
- Google
- Idemia
- Janssen
- Naver Labs
- Nokia
- Pfizer
- Stellantis
- Valeo
- Vertex

Summary: The PRAIRIE Institute (PaRis AI Research InstitutE) is one of the four French Institutes of Artificial Intelligence, which were created as part of the national French initiative on AI announced by President Emmanuel Macron on May 29, 2018. PRAIRIE's objective is to become within five years a world leader in AI research and higher education, with an undeniable impact on economy and technology at the French, European and global levels. It brings together academic members ("PRAIRIE chairs") who excel at research and education in both the core methodological areas and the interdisciplinary aspects of AI, and industrial members that are major actors in AI at the global level and a very strong group of international partners. Benoît Sagot holds a PRAIRIE chair. Rachel Bawden holds a junior PRAIRIE chair.

LabEx EFL

Duration: 1 Oct 2010–30 Sep 2024.

PI: Barbara Hemforth (LLF).

Coordinators for ALMAnaCH: Benoît Sagot, Djamé Seddah and Éric de La Clergerie.

Summary: Empirical foundations of linguistics, including computational linguistics and natural language processing. ALMAnaCH's predecessor team 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. Several ALMAnaCH members are now "individual members" of the LabEx EFL. B. Sagot serves as deputy head (and former head) of one of the scientific strands of the LabEx, namely strand 6 dedicated to language resources. Benoît Sagot and D; Seddah are

(co-)heads of a number of scientific “operations” within strands 6, 5 (“computational semantic analysis”) and 2 (“experimental grammar”). Main collaborations are related to 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] and LLF [CNRS and Paris-Diderot]).

GDR LiLT

Duration: 1 Jan 2019–present.

Summary: Linguistic issues in language technology.

10.2.2 Other National Initiatives

Informal initiative Cap’FALC

Duration: 1 Jan 2018–present.

Coordinator for ALMAnaCH: Benoît Sagot.

Partners:
- UNAPEI
- FAIR

Summary: The text simplification algorithm developed within Cap’FALC is based on neural models for natural language processing. It will work similarly to a spell checker, which marks passages in a text, offers solutions but does not correct without a human validation step. The tool is intended to represent a valuable aid for disabled people responsible for transcribing texts in FALC, not to replace their intervention at all stages of the drafting; only their expertise can validate a text as being accessible and easy to read and understand. Cap’FALC is endorsed by the French Secretary of State for Disabled People and supported by Malakoff Humanis via the CCAH (National Disability Action Coordination Committee).

Convention (MIC, Archives Nationales) LECTAUREP

Duration: 1 Jan 2018–4 Nov 2021.

PI: Laurent Romary.

Coordinator for ALMAnaCH: Laurent Romary.

Partners:
- ÉPHÉ
- Archives Nationales
- Ministère de la culture

Summary: Development of a platform for the transcription, reading and automatic analysis of notarial deeds present in the National Archives.

Convention (MIC, Archives Nationales) DAHN

Duration: 1 Jun 2019–30 Apr 2022.

PI: Laurent Romary.

Coordinator for ALMAnaCH: Laurent Romary.

Partners:
- ÉPHÉ
- Université du Mans
- Ministère de la culture

Summary: Digitalisation and computational exploitation of archives of historical interest.

Convention (MIC, Archives Nationales) NER4archives

Duration: 1 Jan 2020–23 Sep 2021.

PI: Laurent Romary.

Coordinator for ALMAnaCH: Laurent Romary.

Partners:
- Ministère de la culture
- Archives Nationales

Summary: Named entity recognition for finding aids in XML-EAD, a standard for encoding descriptive information regarding archival records.

TGIR Huma-Num

Duration: 1 Jan 2013–present.

Summary: ALMAnaCH is a member of the CORLI consortium on "corpora, languages and interactions" (B. Sagot is a member of the consortium's board).

DIM Matériaux Anciens et Patrimoniaux

Duration: 1 Jan 2017–present.

PI: Étienne Anheim, Loïc Bertrand, Isabelle Rouget.

Coordinator for ALMAnaCH: Laurent Romary.

Summary: The DIM "Matériaux anciens et patrimoniaux" (MAP) is a region-wide research network. Its singularity relies on a close collaboration between human sciences, experimental sciences such as physics and chemistry, scientific ecology and information sciences, while integrating socio-economical partners from the cultural heritage environment. Based on its research, development and valorization potential, we expect such an interdisciplinary network to raise the Ile-de-France region up to a world-top position as far as heritage sciences and research on ancient materials are concerned.

BNF Datalab Gallic(orpor)a

Duration: 1 Oct 2021–31 Dec 2021.

PI: Benoît Sagot.

Coordinator for ALMAnaCH: Benoît Sagot.

Partners:
- École Nationale des Chartes
- Université de Genève

Summary: Consolidate and apply a processing chain for ancient Gallica documents in long diachrony, from the first French manuscripts to revolutionary prints. (end date to be confirmed).

Convention (MIC) DataCatalogue

Duration: 12 Aug 2021–12 Dec 2022.

PI: Laurent Romary.

Coordinator for ALMAnaCH: Laurent Romary.

Partner:
- Ministère de la culture, INHA, Bibliothèque Nationale de France

Summary: The project aims at contributing to the proper transition between a basic digitalisation of cultural heritage content and the actual usage of the corresponding content within a "collection as data" perspective. To acheive this, we experiment news methods for extracting the logical structure of scanned (and OCRed) catalogues and standardise their content for publication towards curators, researchers, or wider users.

10.3 Regional initiatives

Framework agreement with Inria AP-TAL

Duration: 1 Apr 2020–present.

PIs: Laurent Romary, Éric de La Clergerie and Benoît Sagot.

Coordinators for ALMAnaCH: Laurent Romary, Éric de La Clergerie and Benoît Sagot.

Partner:
- APHP

Summary: Within the AP-TAL and HopiTAL projects, ALMAnaCH is involved in collaborative work with APHP and other Inria teams whose goal is ti help dealing with the COVID-19 pandemics. ALMAnaCH's contributions are related to the deployment of NLP techniques on COVID-19-related non-structured text data.

11 Dissemination

11.1 Promoting scientific activities

11.1.1 Scientific events: selection

- Djamé Seddah: Programme chair for EACL 2021 Demo track and IWPT 2021 Shared Task.

Reviewer and member the conference program committees

- Alix Chagué: Reviewer for FF21 (Fantastic Futures 2021).
- Clémentine Fourrier: Subreviewer for EMNLP 2021 and ACL 2021.
- Éric de La Clergerie: Reviewer for EACL 2021.
- Hugo Scheithauer: Reviewer for FF21 (Fantastic Futures 2021).
- Louis Martin: Reviewer for CHI 2021, EACL 2021 and AAAI 2021.
- Pedro Ortiz Suarez: Reviewer for EACL 2021 (Demo track), ACL 2021, EMNLP 2021, CHR2021, ARR October 2021 and ARR November 2021.
- Rachel Bawden: Reviewer for EACL 2021 (Demo track), NAACL SRW 2021 (Student research workshop), ACL-IJCNLP 2021, WMT 2021, RECITAL 2021 (TALN workshop), ACL-IJCNLP SRW 2021 (Student research workshop), EMNLP 2021 and ACL Rolling Reviews (November).
- Laurent Romary: Reviewer for CHR 2021 (2d Computational Humanities Research conference), ISA-17 (Seventeenth Joint ACL-ISO Workshop on Interoperable Semantic Annotation) and 9th Workshop on the Challenges in the Management of Large Corpora.
- Benoît Sagot: Reviewer for EACL 2021 (Demo track), LREC 2020, AdaptNLP 2021, IJCAI 2021, ACL 2021 and Programme committee member for CTTS 2021 (Current Trends in Text Simplification Workshop).
- Djamé Seddah: Reviewer for ACL 2021, EMNLP 2021, LAW workshop 2021, WNUT workshop 2021, Insight on Negative Results workshop 2021 and BlackBox NLP workshop 2021.
- Syrielle Montariol: Reviewer for ACL 2021.

11.1.2 Journal

Member of the editorial boards

- Rachel Bawden: Member of the editorial board for *Northern European Journal of Language Technology*.

Reviewer - Reviewing Activities

- Alix Chagué: Reviewer for *JDMDH (Journal of Data Mining & Digital Humanities)*.
- Pedro Ortiz Suarez: Reviewer for *JDMDH (Journal of Data Mining & Digital Humanities)*.
- Rachel Bawden: Reviewer for *Bulletin de la Société de Linguistique de Paris* and *Computational Linguistics* (Book Review).
- Benoît Sagot: Reviewer for *Bulletin de la Société de Linguistique de Paris 2021*, *Transactions on Affective Computing* and *ACM Computing Surveys*.
- Djamé Seddah: Reviewer for *Transactions on Asian and Low-Resource Language Information Processing*.

11.1.3 Invited talks

- Louis Martin:
  - Inria Paris (12 Feb 2021): “Cap’FALC, le projet d’outil numérique Facile à Lire et à Comprendre !”.
- Alix Chagué:
  - Séminaire du MATE-SHS (tuto@mate) ; online (11 Mar 2021): “Comment faire lire des gribouillis à mon ordinateur ?”.
  - Séminaire “Sciences du patrimoine - sciences du texte. Confrontation des méthodes” organisé par l’École nationale des chartes, Paris (20 May 2021): “CREMMA : une infrastructure mutualisée pour la reconnaissance d’écritures manuscrites et la patrimonialisation numérique” [33].
- Pedro Ortiz Suarez:
  - Séminaires du Master Sciences du Langage, Université Paris Nanterre (23 Nov 2021): “Les Modèles de Langue pour le Français Contemporain et Historique”.
- Lucas Terriel:
  - Les Futurs Fantastiques - 3e Conférence Internationale sur l’Intelligence Artificielle appliquée aux Bibliothèques, Archives et Musées, AI4LAM; Bibliothèque nationale de France, Dec 2021, Paris, France (9 Dec 2021): “NER4Archives (named entity recognition for archives) : méthodes et outils semi-automatiques pour reconnaître les entités nommées dans les instruments de recherche archivistiques encodés en XML/EAD”.
- Éric de La Clergerie:
  - Séminaire des fondamentaux - Université de Genève (26 Oct 2021): “Vers un corpus outillé en Humanité Numérique”.
  - Session Inria Alumni, Paris (CNAM) (19 Nov 2021): “TAL & IA Embeddings? Deep Learning? Language Models?”.
  - Colloque final TimeUS, Lyon 9-10 Septembre 2021 (9 Sep 2021): “Empowering a corpus in Digital Humanities”.

  - École 43, Bluenove (17 Sep 2021): “TAL? Classification ? Clustering? Topic Modeling? Embeddings? Deep Learning?”.

- Syrielle Montariol:

  - Josef Stefan Institute, Ljubjana (online) (6 Oct 2021): “Detecting Omissions of Risk Factors in Company Annual Reports”.
  - Hong Kong Polytechnic University (online) (23 Jun 2021): “Models of diachronic semantic change: Static, Dynamic, Contextualised.”.
  - Inria Lille (9 Dec 2021): “Models of diachronic semantic change”.

11.1.4 Scientific expertise

- Djamé Seddah

  - Reviewer for the ANR (evaluation AAP ASTRID).

- Benoît Sagot

  - Expert for Consultation by the Conseil d’État on the impact of AI research on public policies (21/07/2021).
  - Expert for Consultation by government AI policy decision makers (16/09/2021).

11.1.5 Research administration

- Djamé Seddah

  - Member of the scientific board for the ANR (evaluation AAP ASTRID).

- Laurent Romary

  - President of the scientific board for ABES (Website).
  - Member of the scientific board for the ELEXIS Interoperability and Sustainability Committee (ISC) (ELEXIS is the European Lexicographic Infrastructure.).
  - Member of the scientific board for the Schloss Dagstuhl Scientific Advisory Board (Website).
  - Member of the international advisory board for the Research Infrastructure project LINDAT/CLARIAH-CZ.

- Pedro Ortiz Suarez

  - Member of the user committee for the Jean Zay supercomputer (Invited member).

- Benoît Sagot

  - Member of the scientific board for Société de Linguistique de Paris (Responsable de la communication numérique 1).
  - Member of the scientific board for Inria Paris’s Comité des Projets (member of the Scientific Board of the Inria Paris research centre (Bureau du Comité des Projets)).

- Rachel Bawden

  - Member of the scientific board for Société de Linguistique de Paris (Responsable de la communication numérique 2).

11.2 Teaching - Supervision - Juries

11.2.1 Teaching

- Clémentine Fourrier: Eq. Master's course (M1,M2) Module TDLog - Techniques de Developpement Logiciel (computer science course in Python) (12.25 hours). École des Ponts ParisTech, France.

- Alix Chagué: Master's course (M2) as part of the Master "Documentation et Humanités Numériques". Méthodologie de la recherche et préprofessionalisation (18 hours), coorganised with Françoise Dalex. École du Louvre, France.

- Alix Chagué: Master's course (M2) as part of the Master "Documentation et Humanités Numériques". Typologie, formats et outils d'exploitation des documents numériques : introduction à XML TEI (6 hours). École du Louvre, France.

- Alix Chagué: Master's course (M2) as part of the Master "Documentation et Humanités Numériques". Introduction à Python et à l'algorithmie (6 hours). École du Louvre, France.

- Syrielle Montariol: Master's course (M2) as part of the Machine Learning and AI for Economics and Finance Summer school. Advanced NLP for Finance (3 hours). Université Paris Dauphine.

- Hugo Scheithauer: Master's course (M2) as part of the Master "Documentation et Humanités Numériques". Collecte et publication de jeux de données avec Python (3.3 hours). École du Louvre, France.

- Alix Chagué: Master's course (M2) as part of the Master "Documentation et Humanités Numériques". Méthodologie de la recherche et préprofessionalisation (18 hours), coorganised with Françoise Dalex. École du Louvre, France.

- Arij Riabi: Bachelor's course (L1) as part of the Licence sciences pour l'ingenieur. Programmation pour les sciences (39 hours). Université Paris-Est Créteil Val de Marne, France.

- Louis Martin: Master's course (M2) as part of the African Master's in Machine Intelligence. Deep Natural Language Processing (25 hours), coorganised with Antoine Bordes and Angela Fan. Virtual.

- Benoît Sagot: Master's course (M2) as part of the Master "Mathématiques, Vision Apprentissage". Speech and Language Processing (20 hours), coorganised with Emmanuel Dupoux. ENS Paris-Saclay, France.

- Alix Chagué: Animated Training and Workshop to us eScriptorium and build HTR projects, Ecole nationale des chartes, Paris (18 Feb 2021): eScriptorium et l'HTR (3 hours)

- Floriane Chiffoleau: Animated 2-day training, URFIST de Bretagne et des Pays de la Loire, France (1 Sep 2021): Introduction à la TEI (12 hours)

- Pedro Ortiz Suarez: Animated Tutorial for the workshop "Philologie computationnelle: au delà de l'encodage du texte", Université de Lausanne (3 Dec 2021): Reconnaître les entités nommées (3 hours)

- Rachel Bawden: Animated Tutorial for the workshop "Philologie computationnelle: au delà de l'encodage du texte", Université de Lausanne (3 Dec 2021): Normalisation de la langue (3 hours)

- Benoît Sagot: Animated Tutorial for the workshop "Philologie computationnelle: au delà de l'encodage du texte", Université de Lausanne (3 Dec 2021): Modèles de langue, histoire et objectifs (keynote talk) (1 hours)

11.2.2 Supervision

PhD

- PhD in progress: Axel Herold, “Extraction of etymological information from digital dictionaries” (1 Oct 2016–present). Supervised by Laurent Romary.
- PhD defended: Louis Martin, “Automatic text simplification” (1 Jun 2018–31 Aug 2021). Supervised by Benoît Sagot and Éric de La Clergerie. PhD defended on 27 Oct 2021.
- PhD in progress: José Carlos Rosales Núñez, “Machine translation for user-generated content” (1 Jun 2018–present). Supervised by Guillaume Wisniewksi and Djamé Seddah.
- PhD in progress: Pedro Ortiz Suarez, “NLP and IE from 17th century encyclopedia” (1 Oct 2018–present). Supervised by Laurent Romary and Benoît Sagot.
- PhD in progress: Benjamin Muller, “NLP for social media texts” (1 Oct 2018–present). Supervised by Benoît Sagot and Djamé Seddah.
- PhD in progress: Clémentine Fourrier, “Neural architecture development, computational historical linguistics” (1 Oct 2019–present). Supervised by Laurent Romary, Benoît Sagot and Rachel Bawden.
- PhD in progress: Robin Algayres, “Unsupervised Automatic Speech Recognition in low resource conditions” (1 Oct 2019–present). Supervised by Emmanuel Dupoux and Benoît Sagot.
- PhD in progress: Lionel Tadjou Tadonfouet, “Conversations Disentanglement” (1 Mar 2020–present). Supervised by Laurent Romary and Éric de La Clergerie.
- PhD in progress: Tú Anh Nguyễn, “Unsupervised acquisition of linguistic representations from speech (audio) data” (19 Apr 2021–present). Supervised by Benoît Sagot.
- PhD in progress: Paul-Ambroise Duquenne, “Study of vector spaces for sentence representation” (15 May 2021–present). Supervised by Benoît Sagot.
- PhD in progress: Lydia Nishimwe, “Robust Neural Machine Translation” (1 Oct 2021–present). Supervised by Benoît Sagot and Rachel Bawden.
- PhD in progress: Roman Castagné, “Neural language modelling” (1 Oct 2021–present). Supervised by Benoît Sagot and Éric de La Clergerie.
- PhD in progress: Arij Riabi, “NLP for low-resource, non-standardised language varieties, especially North-African dialectal Arabic written in Latin script” (1 Oct 2021–present). Supervised by Djamé Seddah and Laurent Romary.
- PhD in progress: Floriane Chiffoleau, “Training data and creation of models for the text recognition of typewritten or handwritten corpus of archival collection” (15 Oct 2021–present). Supervised by Anne Baillot and Laurent Romary.
- PhD in progress: Matthieu Futeral-Peter, “Text-image multimodal models” (1 Nov 2021–present). Supervised by Ivan Laptev and Rachel Bawden.
- PhD in progress: Alix Chagué, “Methodology for the creation of training data and the application of handwritten text recognition to the Humanities.” (1 Nov 2021–present). Supervised by Laurent Romary, Emmanuel Château-Dutier and Michael Sinatra.
- PhD in progress: Nathan Godey, “Neural language modelling” (1 Dec 2021–present). Supervised by Benoît Sagot.

11.2.3 Juries

PhD

- Benoît Sagot
  - Reviewer of the PhD committee for Silvia García Mendéz at the Universidad de Vigo on 1 Feb 2021. Title: *Contribution to Natural Language Generation for Spanish*
  - Reviewer of the PhD committee for Adelle Abdallah at the Université Paris 8 Vincennes à Saint-Denis & Université Libanaise on 2 Jul 2021. Title: *Catégorisation sémantique et information grammaticale en arabe*
  - Director of the PhD committee for Louis Martin at the Sorbonne Université, Paris, France on 27 Oct 2021. Title: *Simplification automatique de phrases à l'aide de méthodes contrôlables et non supervisées*
  - Reviewer of the PhD committee for Hicham El Boukkouri at the Université Paris-Saclay on 18 Nov 2021. Title: *Domain Adaptation of Word Embeddings Through the Exploitation of In-domain Corpora and Knowledge Bases*
- Djamé Seddah
  - Examiner of the PhD committee for Syrielle Montariol at the Université Paris-Saclay on 8 Feb 2021. Title: *Models of diachronic semantic change using word embeddings*
- Éric de La Clergerie
  - Co-supervisor of the PhD committee for Louis Martin at the Sorbonne Université, Paris, France on 27 Oct 2021. Title: *Simplification automatique de phrases à l'aide de méthodes contrôlables et non supervisées*

Master

- Alix Chagué
  - Examiner of the Master committee for Paul Jean at the Ecole du Louvre, Paris, France on 28 Jun 2021. Title: *Enjeux et outils d'un référencement en données géographiques : le cas du projet DatArt*
  - Examiner of the Master committee for Morgane Mosnier at the Ecole du Louvre, Paris, France on 28 Jun 2021. Title: *Les outils numériques du service des Ressources documentaires du musée d'Archéologie nationale Enjeux et évolutions*
  - Examiner of the Master committee for Hugo Scheithauer at the Ecole nationale des chartes, Paris, France on 27 Sep 2021. Title: *La reconnaissance d'entités nommées appliquées à des données issues de la transcription automatique de documents manuscrits patrimoniaux : expérimentations et préconisations à partir du projet LECTAUREP*
  - Examiner of the Master committee for Evi Ronne at the Ecole du Louvre, Paris, France on 4 Oct 2021. Title: *Enjeux d'une création de thésaurus genres iconographiques pour la description des collections du Château de Versailles*
  - Examiner of the Master committee for Paola Fava at the Ecole du Louvre, Paris, France on 5 Oct 2021. Title: *Enrichir la documentation de la muséographie d'un château-musée : le cas de la photothèque numérique du Château de Versailles*
  - Examiner of the Master committee for Faustine Le Garrec at the Ecole du Louvre, Paris, France on 7 Oct 2021. Title: *Enrichissement documentaire des collections en ligne du Centre national des arts plastiques : analyse comparée de stratégies de valorisation numérique*
  - Examiner of the Master committee for Paul Kervegan at the Ecole du Louvre, Paris, France on 7 Oct 2021. Title: *Du premier traitement à la donnée ouverte. Enjeux et méthodologie du référencement des travaux scientifiques de l'INA grm.*

  - Examiner of the Master committee for Camille Graindorge at the Ecole du Louvre, Paris, France on 11 Oct 2021. Title: *Enjeu des référentiels de la collection de photographies et propositions de solutions pour la valorisation du contexte colonial : le cas du musée du quai Branly-Jacques Chirac*
  - Examiner of the Master committee for Clara Lelièvre at the Ecole du Louvre, Paris, France on 11 Oct 2021. Title: *Enjeux et méthodes d'un liage de référentiels géographiques : l'exemple du projet de recherche ALEGORIA*
  - Examiner of the Master committee for Juliette Aujard at the Ecole du Louvre, Paris, France on 13 Oct 2021. Title: *Proposition d'une méthodologie de traitement documentaire d'un fonds de planches contact Numérisation, reconditionnement et valorisation du fonds Marc Riboud conservé au Musée National des Arts Asiatiques Guimet*
  - Examiner of the Master committee for Salomé Sieurac at the Ecole du Louvre, Paris, France on 13 Oct 2021. Title: *Description documentaire des enregistrements sonores d'événements musicaux. Etude comparative, normes, standards de description et valorisation de contenus numériques à partir de l'exemple de la Philharmonie de Paris.*
  - Examiner of the Master committee for Justin Kalinowski at the Ecole du Louvre, Paris, France on 14 Oct 2021. Title: *Objectifs et propositions d'amélioration des données documentaires descriptives des biens culturels de l'académie de France à Rome*
  - Examiner of the Master committee for Eva Mesko at the Ecole du Louvre, Paris, France on 30 Nov 2021. Title: *Les archives audiovisuelles. Processus de patrimonialisation. Enjeux au sein d'entreprises privées.*

- Benoît Sagot

  - Director of the Master committee for Roman Castagné at the MVA & ENPC on 20 Sep 2021. Title: *Quelle tokenisation pour les modèles de langue ?*
  - Co-director of the Master committee for Matthieu Futeral-Peter at the MVA & ENSAE on 11 Oct 2021. Title: *Exploration of multilingual and multimodal word embeddings*
  - Examiner of the Master committee for Hugo Laurençon at the MVA & ENPC on 20 Sep 2021. Title: *Modélisation non supervisée du langage à l'aide de frontières de mots explicites ou implicites*
  - Examiner of the Master committee for Ines Florez de la Colina at the MVA & CentraleSupelec on 19 Oct 2021. Title: *Designing high-performance, compact and efficient language models for a specialised search engine*
  - Examiner of the Master committee for Julie Dessaint at the MVA & Télécom ParisTech on 14 Sep 2021. Title: *Extracting information from text*
  - Examiner of the Master committee for Lucie Galland at the MVA on 10 Sep 2021. Title: *Adaptive conversational agent using reinforcement learning*
  - Examiner of the Master committee for Lucile Saulnier at the MVA on 13 Sep 2021. Title: *Exploring Language Models*
  - Examiner of the Master committee for Mahdi Kallel at the MVA & Télécom ParisTech on 21 Oct 2021. Title: *Few shot learning with language models*
  - Examiner of the Master committee for Mohamed Amine Hachicha at the MVA & CentraleSupelec on 19 Oct 2021. Title: *Model Architecture and Automatic Training Optimization for Text Classification*
  - Examiner of the Master committee for Omar Souaidi at the MVA & CentraleSupelec on 15 Oct 2021. Title: *Conversational Speech Recognition for Low-Ressource Languages*
  - Examiner of the Master committee for Valentin Taillandier at the MVA on 17 Sep 2021. Title: *Improving the Factual Knowledge Encoding of Large Language Models*

- Rachel Bawden

  - Co-supervisor of the Master committee for Matthieu Futeral-Peter at the MVA & ENSAE on 11 Oct 2021. Title: *Exploration of multilingual and multimodal word embeddings*

11.3 Popularization

11.3.1 Articles and contents

- Clémentine Fourrier authored an article for “Comment les applications savent-elles ce que vous allez écrire?”. Je Science donc Je Suis online Journal, 1 Mar 2021.
- Alix Chagué, with LECTAUREP Blog, authored an article for “Résolution de bug : Something went wrong during the segmentation”. Online, 25 Mar 2021.
- Floriane Chiffoleau, with the Digital Intellectuals Blog, authored an article for “Storing and sharing the project”. Online, 19 Apr 2021.
- Alix Chagué, with LECTAUREP Blog, authored an article for “Traces6 : notre serveur principal”. Online, 21 Apr 2021.
- Hugo Scheithauer, with LECTAUREP Blog, authored an article for “Un exemple d’exploitation des données produites grâce à la reconnaissance d’écriture manuscrite : la reconnaissance d’entités nommées”. Online, 3 Jun 2021.
- Alix Chagué, with LECTAUREP Blog, authored an article for “Création de modèles de transcription pour le projet LECTAUREP #1”. Online, 9 Jun 2021.
- Floriane Chiffoleau, with the Digital Intellectuals Blog, authored an article for “Publication of my digital edition – Developing my TEI Publisher application”. Online, 16 Jun 2021.
- Floriane Chiffoleau, with the Digital Intellectuals Blog, authored an article for “Availability and high quality: distributing the facsimile with NAKALA”. Online, 2 Sep 2021.
- Alix Chagué, with LECTAUREP Blog, authored an article for “Création de modèles de transcription pour le projet LECTAUREP #2”. Online, 6 Oct 2021.
- Floriane Chiffoleau, with the Digital Intellectuals Blog, authored an article for “Publication of my digital edition – Online launch of the TEI Publisher application”. Online, 10 Dec 2021.

11.3.2 Education

- Rachel Bawden, with The Latymer School, Edmonton, UK, participated in “Shorts posts by alumni of the Latymer School for the International Day of Women and Girls in Science.”. Online, 11 Feb 2021.
- Clémentine Fourrier participated in “Rendez-Vous des Jeunes Mathématiciennes et Informaticiennes Inria 2021”. Online, 1 Oct 2021 (2 hours).

11.3.3 Interventions

- Alix Chagué, with ADEMEC (Association des Diplômés et Etudiants des Masters de l’Ecole des chartes), was the main organiser of “ADEMEC Monthly Workshops”. Ecole nationale des chartes, 1 Dec 2020 (2 hours).
- Clémentine Fourrier, with Académie de Paris , gave a talk on “IA et language: sur quoi, comment, pourquoi?”. Paris, 1 Mar 2021 (2 hours).
- Rachel Bawden gave a talk on “Natural Language Processing: Applications and Neural Modelling”. Online, 6 Jul 2021 (1.5 hours).
- Benoît Sagot, with Kili Technologies, gave a talk at “Data Centric AI day” on “Resources for multilingual NLP in the neural era: the examples of OSCAR and CamemBERT”. Online (Paris), 9 Nov 2021 (1 hour).
- Alix Chagué & Hugo Scheithauer, with Brace your digital scholarly edition!, gave a talk on “From eScriptorium to TEI Publisher”. Berlin, 19 Nov 2021 (0.5 hours).

- Benoît Sagot, with Groupe X-IA, gave a talk at "Soirée X-IA" on "Développement, utilisation et limites des modèles de langue, avec un focus sur OSCAR et CamemBERT". Paris, 23 Nov 2021 (3 hours).
- Lucas Terriel, Alix Chagué & Hugo Scheithauer, with Les Futurs Fantastiques - 3e Conférence Internationale sur l'Intelligence Artificielle appliquée aux Bibliothèques, Archives et Musées, AI4LAM, gave a talk on "Atelier : Production d'un modèle affiné de reconnaissance d'écriture manuscrite avec eScriptorium et évaluation de ses performances. Évaluer son modèle HTR/OCR avec KaMI (Kraken as Model Inspector)". Online, 1 Dec 2021 (2 hours).
- Clémentine Fourrier gave a talk on "La recherche en TAL". IHP, 16 Dec 2021 (2 hours).
- Pedro Ortiz Suarez, with Fundación Vía Libre, gave a talk on "Inteligencia Artificial Abierta". Online, 28 Dec 2021 (1 hour).

12 Scientific production

12.1 Major publications

[1] D. Fišer and B. Sagot. 'Constructing a poor man's wordnet in a resource-rich world'. In: *Language Resources and Evaluation* 49.3 (2015), pp. 601–635. DOI: 10.1007/s10579-015-9295-6. URL: https://hal.inria.fr/hal-01174492.

[2] G. Jawahar, B. Sagot and D. Seddah. 'What does BERT learn about the structure of language?' In: *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy, July 2019. URL: https://hal.inria.fr/hal-02131630.

[3] P. Lopez and L. Romary. 'HUMB: Automatic Key Term Extraction from Scientific Articles in GROBID'. In: *SemEval 2010 Workshop*. ACL SigLex event. Uppsala, Sweden, July 2010, pp. 248–251. URL: https://hal.inria.fr/inria-00493437.

[4] L. Martin, B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, É. Villemonte de La Clergerie, D. Seddah and B. Sagot. 'CamemBERT: a Tasty French Language Model'. In: *ACL 2020 - 58th Annual Meeting of the Association for Computational Linguistics*. Seattle / Virtual, United States, July 2020. DOI: 10.18653/v1/2020.acl-main.645. URL: https://hal.inria.fr/hal-02889805.

[5] P. J. Ortiz Suárez, B. Sagot and L. Romary. 'Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures'. In: *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Ed. by P. Bański, A. Barbaresi, H. Biber, E. Breiteneder, S. Clematide, M. Kupietz, H. Lüngen and C. Iliadi. Cardiff, United Kingdom: Leibniz-Institut für Deutsche Sprache, July 2019. DOI: 10.14618/IDS-PUB-9021. URL: https://hal.inria.fr/hal-02148693.

[6] C. Ribeyre, É. Villemonte de La Clergerie and 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*. Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Denver, USA, United States, June 2015. URL: https://hal.archives-ouvertes.fr/hal-01174533.

[7] L. Romary. 'TEI and LMF crosswalks'. In: *JLCL - Journal for Language Technology and Computational Linguistics* 30.1 (2015). URL: https://hal.inria.fr/hal-00762664.

[8] 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. URL: https://hal.inria.fr/inria-00521242.

[9] B. Sagot and É. Villemonte de La Clergerie. 'Error mining in parsing results'. In: *The 21st International Conference of the Association for Computational Linguistics (ACL 2006)*. Sydney, Australia, July 2006, pp. 329–336. URL: https://hal.inria.fr/hal-02270412.

[10] D. Seddah, B. Sagot, M. Candito, V. Mouilleron and V. Combet. 'The French Social Media Bank: a Treebank of Noisy User Generated Content'. Anglais. In: *COLING 2012 - 24th International Conference on Computational Linguistics*. Kay, Martin and Boitet, Christian. Mumbai, Inde, Dec. 2012. URL: http://hal.inria.fr/hal-00780895.

[11] R. Tsarfaty, D. Seddah, Y. Goldberg, S. Kübler, Y. Versley, M. Candito, J. Foster, I. Rehbein and 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.

[12] R. Tsarfaty, D. Seddah, S. Kübler and J. Nivre. 'Parsing Morphologically Rich Languages: Introduction to the Special Issue'. In: *Computational Linguistics*. Special Issue on Parsing Morphologically-Rich Languages 39.1 (Mar. 2013), p. 8. DOI: 10.1162/COLI_a_00133. URL: https://hal.inria.fr/hal-00780897.

[13] É. Villemonte de La Clergerie. 'Improving a symbolic parser through partially supervised learning'. In: *The 13th International Conference on Parsing Technologies (IWPT)*. Naria, Japan, Nov. 2013. URL: https://hal.inria.fr/hal-00879358.

12.2 Publications of the year

International journals

[14] R. Bawden. '[Book Review] Understanding Dialogue: Language Use and Social Interaction'. In: *Computational Linguistics* (2021). URL: https://hal.inria.fr/hal-03324500.

[15] L. Foppiano, S. Dieb, A. Suzuki, P. Baptista de Castro, S. Iwasaki, A. Uzuki, M. G. Esparza Echevarria, Y. Meng, K. Terashima, L. Romary, Y. Takano and M. Ishii. 'SuperMat: Construction of a linked annotated dataset from superconductors-related publications'. In: *Science and Technology of Advanced Materials: Methods* 1.1 (13th July 2021). DOI: 10.1080/27660400.2021.1918396. URL: https://hal.inria.fr/hal-03101177.

[16] N. Truan and L. Romary. 'Building, Encoding, and Annotating a Corpus of Parliamentary Debates in XML-TEI: A Cross-Linguistic Account'. In: *Journal of the Text Encoding Initiative* (2021). URL: https://halshs.archives-ouvertes.fr/halshs-03097333.

[17] F. Uiterwaal, F. Niccolucci, S. Bassett, S. Krauwer, H. Hollander, F. Admiraal, L. Romary, G. BRUSEKER, C. Meghini, J. Edmond and M. Hedges. 'From disparate disciplines to unity in diversity How the PARTHENOS project has brought European humanities Research Infrastructures together'. In: *International Journal of Humanities and Arts Computing* 15.1-2 (Oct. 2021), pp. 101–116. DOI: 10.3366/ijhac.2021.0264. URL: https://hal.inria.fr/hal-03402145.

International peer-reviewed conferences

[18] J. Abadji, P. J. Ortiz Suárez, L. Romary and B. Sagot. 'Ungoliant: An Optimized Pipeline for the Generation of a Very Large-Scale Multilingual Web Corpus'. In: CMLC 2021 - 9th Workshop on Challenges in the Management of Large Corpora. Limerick / Virtual, Ireland, 13th July 2021. DOI: 10.14618/ids-pub-10468. URL: https://hal.inria.fr/hal-03301590.

[19] F. Arthaud, R. Bawden and A. Birch. 'Few-shot learning through contextual data augmentation'. In: EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics. Proceedings of the 16th conference of the European Chapter of the Association for Computational Linguistics. Kiev / Virtual, Ukraine, 19th Apr. 2021. URL: https://hal.inria.fr/hal-03121971.

[20] R. Costa, A. Salgado, A. F. Khan, S. Carvalho, L. Romary, B. Almeida, M. Ramos, M. Khemakhem, R. Silva and T. Tasovac. 'MORDigital: The Advent of a New Lexicographical Portuguese Project'. In: eLex 2021 - Seventh biennial conference on electronic lexicography. Brno, Czech Republic, 5th July 2021. URL: https://hal.inria.fr/hal-03195362.

[21] G. Felhi, J. L. Roux and D. Seddah. 'Challenging the Semi-Supervised VAE Framework for Text Classification'. In: Second Workshop on Insights from Negative Results in NLP (colocated with EMNLP). Proceedings of the Second Workshop on Insights from Negative Results in NLP. Punta Cana, Dominican Republic: Association for Computational Linguistics, 10th Nov. 2021. URL: https://hal.inria.fr/hal-03540081.

[22] C. Fourrier, R. Bawden and B. Sagot. 'Can Cognate Prediction Be Modelled as a Low-Resource Machine Translation Task?' In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Bangkok, Thailand, 1st Aug. 2021. URL: https://hal.inria.fr/hal-03243380.

[23] S. Gabay, B. Topalov, C. Corbières, L. Rondeau du Noyer, B. Joyeux-Prunel and L. Romary. 'Automating Artl@s – extracting data from exhibition catalogues'. In: EADH 2021 - Second International Conference of the European Association for Digital Humanities. Krasnoyarsk, Russia, 21st Sept. 2021. URL: https://hal.archives-ouvertes.fr/hal-03331838.

[24] S. Montariol and A. Allauzen. 'Transport Optimal pour le Changement Sémantique à partir de Plongements Contextualisés'. In: *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*. TALN 2021 - Traitement Automatique des Langues Naturelles. Lille / Virtuel, France: ATALA, 2021, pp. 235–244. URL: https://hal.archives-ouvertes.fr/hal-03265889.

[25] B. Muller, Y. Elazar, B. Sagot and D. Seddah. 'First Align, then Predict: Understanding the Cross-Lingual Ability of Multilingual BERT'. In: EACL 2021 - The 16th Conference of the European Chapter of the Association for Computational Linguistics. Kyiv / Virtual, Ukraine, 19th Apr. 2021. URL: https://hal.inria.fr/hal-03239087.

[26] J. C. R. Núñez, G. Wisniewski and D. Seddah. 'Noisy UGC Translation at the Character Level: Revisiting Open-Vocabulary Capabilities and Robustness of Char-Based Models'. In: W-NUT 2021 - 7th Workshop on Noisy User-generated Text (colocated with EMNLP 2021). Proceedings of the Seventh W-NUT workshop (colocated with EMNLP 2021). Punta Cana, Dominican Republic, 24th Oct. 2021. URL: https://hal.inria.fr/hal-03540174.

[27] A. Riabi, B. Sagot and D. Seddah. 'Can Character-based Language Models Improve Downstream Task Performance in Low-Resource and Noisy Language Scenarios?' In: Seventh Workshop on Noisy User-generated Text (W-NUT 2021, colocated with EMNLP 2021). Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021). punta cana, Dominican Republic, 10th Jan. 2022. URL: https://hal.inria.fr/hal-03527328.

[28] J. C. Rosales Nunez, D. Seddah and G. Wisniewski. 'Understanding the Impact of UGC Specificities on Translation Quality'. In: W-NUT 2021 - Seventh Workshop on Noisy User-generated Text (colocated with EMNLP 2021). Proceedings of the Seventh W-NUT workshop (colocated with EMNLP 2021). Punta Cana, Dominican Republic, 24th Jan. 2022. URL: https://hal.inria.fr/hal-03540175.

[29] V. Sanh, A. Webson, C. Raffel, S. H. Bach, L. Sutawika, Z. Alyafeai, A. Chaffin, A. Stiegler, T. Le Scao, A. Raja, M. Dey, M. S. Bari, C. Xu, U. Thakker, S. Sharma, E. Szczechla, T. Kim, G. Chhablani, N. V. Nayak, D. Datta, J. Chang, M. T.-J. Jiang, H. Wang, M. Manica, S. Shen, Z.-X. Yong, H. Pandey, M. McKenna, R. Bawden, T. Wang, T. Neeraj, J. Rozen, A. Sharma, A. Santilli, T. Fevry, J. A. Fries, R. Teehan, T. Bers, S. Biderman, L. Gao, T. Wolf and A. M. Rush. 'Multitask Prompted Training Enables Zero-Shot Task Generalization'. In: ICLR 2022 - Tenth International Conference on Learning Representations. Online, Unknown Region, 2021. URL: https://hal.inria.fr/hal-03540072.

[30] L. T. Tadjou, F. Bourge, T. Marie, L. Romary and E. Villemonte de la Clergerie. 'Building A Corporate Corpus For Threads Constitution'. In: Student Research Workshop associated with the International Conference on Recent Advances in Natural Language Processing (RANLP'2021). Online, Bulgaria, 1st Sept. 2021. URL: https://hal.inria.fr/hal-03351533.

National peer-reviewed Conferences

[31] A. Gérard, B. Sagot and E. Pons. 'Le Traitement Automatique des Langues au service du vin'. In: Dataquitaine 2021 - IA, Recherche Opérationnelle & Data Science. Bordeaux / Virtual, France, 25th Feb. 2021. URL: https://hal.inria.fr/hal-03146219.

Conferences without proceedings

[32] A. Bartz, J. Janes, L. Romary, P. Gambette, R. Bawden, P. Ortiz Suarez, B. Sagot and S. Gabay. 'Expanding the content model of annotationBlock'. In: Next Gen TEI, 2021 - TEI Conference and Members' Meeting. Virtual, United States, 25th Oct. 2021. URL: https://hal.archives-ouvertes.fr/hal-03380805.

[33] A. Chagué. 'CREMMA : Une infrastructure mutualisée pour la reconnaissance d'écritures manuscrites et la patrimonialisation numérique'. In: Sciences du patrimoine - sciences du texte. Confrontation des méthodes. Paris, France, 20th May 2021. URL: https://hal.inria.fr/hal-03541887.

[34] A. Chagué and R. Aurélia. 'LECTAUREP : Lecture Automatique des Répertoires de Notaires Parisiens'. In: Fantastic Futures 2021 / Futures Fantastiques 2021. Paris, France, 8th Dec. 2021. URL: https://hal.inria.fr/hal-03479303.

[35] A. Chagué and R. Aurélia. 'LECTAUREP: Paris Notary Record Books Automated Reading'. In: Fantastic Futures 2021 / Futures Fantastiques 2021. Paris, France, 8th Dec. 2021. URL: https://hal.inria.fr/hal-03479258.

[36] A. Chagué and F. Chiffoleau. 'An accessible and transparent pipeline for publishing historical egodocuments'. In: WPIP21 - What's Past is Prologue: The NewsEye International Conference. Virtual, Austria, 16th Mar. 2021. URL: https://hal.archives-ouvertes.fr/hal-03173038.

[37] A. Chagué, T. Clérice and L. Romary. 'HTR-United : Mutualisons la vérité de terrain !' In: DHNord2021 - Publier, partager, réutiliser les données de la recherche : les data papers et leurs enjeux. Lille, France, 15th Nov. 2021. URL: https://hal.archives-ouvertes.fr/hal-03398740.

[38] P. Charbonnier, L. Terriel, F. Clavaud, L. Romary, G. Piraino and V. Verdese. 'NER4Archives (named entity recognition for archives) : méthodes et outils semi-automatiques pour reconnaître les entités nommées dans les instruments de recherche archivistiques encodés en XML/EAD'. In: Les Futurs Fantastiques - 3e Conférence Internationale sur l'Intelligence Artificielle appliquée aux Bibliothèques, Archives et Musées. Paris, France, 9th Dec. 2021. URL: https://hal.archives-ouvertes.fr/hal-03495486.

[39] F. Chiffoleau, A. Baillot and M. Ovide. 'A TEI-based publication pipeline for historical egodocuments -the DAHN project'. In: Next Gen TEI, 2021 - TEI Conference and Members' Meeting. Virtual, United States, 25th Oct. 2021. URL: https://hal.archives-ouvertes.fr/hal-03451421.

[40] G. Felhi, J. Le Roux and D. Seddah. 'Towards Unsupervised Content Disentanglement in Sentence Representations via Syntactic Roles'. In: CtrlGen: Controllable Generative Modeling in Language and Vision. virtual, France, 13th Jan. 2022. URL: https://hal.inria.fr/hal-03540084.

[41] S. Gabay, P. Gambette, R. Bawden, J. Poinhos, E. Kogkitsidou and B. Sagot. 'Variation graphique dans les documents d'Ancien Régime : Nouvelles approches scriptométriques'. In: Journée d'étude : « Pour une histoire de la langue 'par en bas': textes privés et variation des langues dans le passé ». Paris, France, 16th Sept. 2021. URL: https://hal.inria.fr/hal-03357080.

[42] S. Gabay and P. J. Ortiz Suárez. 'A dataset for automatic detection of places in (early) modern French texts'. In: NASSCFL 2021 - 50th Annual North American Society for Seventeenth-Century French Literature Conference. Iowa City / Virtual, United States, 28th May 2021, p. 5. URL: https://hal.archives-ouvertes.fr/hal-03187097.

[43] J.-D. Généro, A. Chagué, V. Le Fourner and M. Puren. 'Transcribing and editing digitized sources on work in the textile industry'. In: Rémunérations et usages du temps des hommes et des femmes dans le textile en France de la fin du XVIIe au début du XXe siècle. Lyon, France, 9th Sept. 2021. URL: https://hal.archives-ouvertes.fr/hal-03340622.

[44] A. Riabi, T. Scialom, R. Keraron, B. Sagot, D. Seddah and J. Staiano. 'Synthetic Data Augmentation for Zero-Shot Cross-Lingual Question Answering'. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Punta cana, Dominican Republic, 6th Nov. 2021. DOI: 10.18653/v1/2021.emnlp-main.562. URL: https://hal.inria.fr/hal-03109187.

[45] H. Scheithauer, A. Chagué, R. Aurélia, L. Terriel, L. Romary, M.-F. Limon-Bonnet, B. Davy, G. Piraino, F. Beltrami, D. Habib, N. Denis and m. durand marc. 'Production d'un modèle affiné de reconnaissance d'écriture manuscrite avec eScriptorium et évaluation de ses performances'. In: Les Futurs Fantastiques - 3e Conférence Internationale sur l'Intelligence Artificielle appliquée aux Bibliothèques, Archives et Musées, AI4LAM. Paris, France, 8th Dec. 2021. URL: https://hal.inria.fr/hal-03538195.

[46] H. Scheithauer, A. Chagué, S. Gabay, L. Romary, J. Janes and C. Jahan. 'From page to content – which TEI representation for HTR output?' In: Next Gen TEI, 2021 - TEI Conference and Members' Meeting. Weaton (virtual), United States, Sept. 2019. URL: https://hal.archives-ouvertes.fr/hal-03380807.

[47] H. Scheithauer, A. Chagué and L. Romary. 'From eScriptorium to TEI Publisher'. In: Brace your digital scholarly edition! Berlin, Germany, 19th Nov. 2021. URL: https://hal.inria.fr/hal-03538115.

[48] L. Terriel. 'Atelier : Production d'un modèle affiné de reconnaissance d'écriture manuscrite avec eScriptorium et évaluation de ses performances. Évaluer son modèle HTR/OCR avec KaMI (Kraken as Model Inspector)'. In: Les Futurs Fantastiques - 3e Conférence Internationale sur l'Intelligence Artificielle appliquée aux Bibliothèques, Archives et Musées. Paris, France, 1st Dec. 2021. URL: https://hal.archives-ouvertes.fr/hal-03495762.

[49] L. Yeganova, D. Wiemann, M. Neves, F. Vezzani, A. Siu, I. J. Unanue, M. Oronoz, N. Mah, A. Névéol, D. Martinez, R. Bawden, G. Di Maria Di Nunzo, R. Roller, P. Thomas, C. Grozea, O. Perez De Viñaspre, M. V. Navarro and A. J. Yepes. 'Findings of the WMT 2021 Biomedical Translation Shared Task: Summaries of Animal Experiments as New Test Set'. In: EMNLP 2021 - Sixth Conference on Machine Translation. Punta Cana, Dominican Republic, 10th Nov. 2021. URL: https://hal.archives-ouvertes.fr/hal-03435096.

Doctoral dissertations and habilitation theses

[50] L. Martin. 'Automatic sentence simplification using controllable and unsupervised methods'. Sorbonne Université, 27th Oct. 2021. URL: https://tel.archives-ouvertes.fr/tel-03543971.

Reports & preprints

[51] J. Abadji, P. Ortiz Suarez, L. Romary and B. Sagot. *Towards a Cleaner Document-Oriented Multilingual Crawled Corpus*. 19th Jan. 2022. URL: https://hal.inria.fr/hal-03536361.

[52] R. Bawden, J. Poinhos, E. Kogkitsidou, P. Gambette, B. Sagot and S. Gabay. *Automatic Normalisation of Early Modern French*. 23rd Jan. 2022. DOI: 10.5281/zenodo.5865428. URL: https://hal.inria.fr/hal-03540226.

[53] J. Bowers, A. Herold, L. Romary and T. Tasovac. *TEI Lex-0 Etym – towards terse recommendations for the encoding of etymological information*. 13th Jan. 2021. URL: https://hal.inria.fr/hal-03108781.

[54] I. Caswell, J. Kreutzer, L. Wang, A. Wahab, D. van Esch, N. Ulzii-Orshikh, A. Tapo, N. Subramani, A. Sokolov, C. Sikasote et al. *Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets*. 23rd Mar. 2021. URL: https://hal.inria.fr/hal-03177623.

[55] B. Haddow, R. Bawden, A. V. Miceli Barone, J. Helcl and A. Birch. *Survey of Low-Resource Machine Translation*. 14th Dec. 2021. URL: https://hal.inria.fr/hal-03479757.

[56] J. Launay, G. L. Tommasone, B. Pannier, F. Boniface, A. Chatelain, A. Cappelli, I. Poli and D. Seddah. *PAGnol: An Extra-Large French Generative Model*. LightON, 16th Oct. 2021. URL: https://hal.inria.fr/hal-03540159.

[57] L. Martin, A. Fan, É. de la Clergerie, A. Bordes and B. Sagot. *Multilingual Unsupervised Sentence Simplification*. 13th Jan. 2021. URL: https://hal.inria.fr/hal-03109299.

[58] S. J. Mielke, Z. Alyafeai, E. Salesky, C. Raffel, M. Dey, M. Gallé, A. Raja, C. Si, W. Y. Lee, B. Sagot and S. Tan. *Between words and characters: A Brief History of Open-Vocabulary Modeling and Tokenization in NLP*. 22nd Jan. 2022. URL: https://hal.inria.fr/hal-03540069.

[59] B. Muller, B. Sagot and D. Seddah. *Can Multilingual Language Models Transfer to an Unseen Dialect? A Case Study on North African Arabizi*. 7th Mar. 2021. URL: https://hal.inria.fr/hal-03161677.

[60] T. Scialom, L. Martin, J. Staiano, É. V. de la Clergerie and B. Sagot. *Rethinking Automatic Evaluation in Sentence Simplification*. 16th Apr. 2021. URL: https://hal.inria.fr/hal-03199901.

[61] T. Tasovac, L. Romary, E. Tóth-Czifra and I. Marinski. *Lexicographic Data Seal of Compliance*. ELEXIS; DARIAH, 30th July 2021. URL: https://hal.archives-ouvertes.fr/hal-03344267.

12.3 Other

Scientific popularization

[62] A. Chagué. *Comment faire lire des gribouillis à mon ordinateur ?* 11th Mar. 2021. URL: https://hal.archives-ouvertes.fr/hal-03170345.

[63] A. Chagué and R. Aurélia. 'Présentation du projet Lectaurep (Lecture automatique de répertoires)'. In: Atelier sur la transcription des écritures manuscrites - BnF DataLab. Paris, France, 26th Jan. 2021. URL: https://hal.archives-ouvertes.fr/hal-03122019.

12.4 Cited publications

[64] M. J. Aranzabe, A. D. De Ilarraza and I. Gonzalez-Dios. 'Transforming complex sentences using dependency trees for automatic text simplification in Basque'. In: *Procesamiento del lenguaje natural* 50 (2013), pp. 61–68.

[65] E. M. Bender, T. Gebru, A. McMillan-Major and S. Shmitchell. 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?' In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. Virtual Event, Canada: Association for Computing Machinery, 2021, pp. 610–623. DOI: 10.1145/3442188.3445922. URL: https://doi.org/10.1145/3442188.3445922.

[66] O. Bonami and B. Sagot. 'Computational methods for descriptive and theoretical morphology: a brief introduction'. In: *Morphology*. Computational methods for descriptive and theoretical morphology 27.4 (2017), pp. 1–7. DOI: 10.1017/CBO9781139248860. URL: https://hal.inria.fr/hal-01628253.

[67] A. Bouchard-Côté, D. Hall, T. Griffiths and D. Klein. 'Automated Reconstruction of Ancient Languages using Probabilistic Models of Sound Change'. In: *Proceedings of the National Academy of Sciences* 110 (2013), pp. 4224–4229.

[68] J. C. K. Cheung and G. Penn. 'Utilizing Extra-sentential Context for Parsing'. In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. EMNLP '10. Cambridge, Massachusetts, 2010, pp. 23–33.

[69] M. Constant, M. Candito and 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, Oct. 2013, pp. 46–52. URL: https://hal.archives-ouvertes.fr/hal-00932372.

[70] S. Desrochers, C. Paradis and V. M. Weaver. 'A Validation of DRAM RAPL Power Measurements'. In: *Proceedings of the Second International Symposium on Memory Systems*. MEMSYS '16. Alexandria, VA, USA: Association for Computing Machinery, 2016, pp. 455–470. DOI: 10.1145/2989081.2989088. URL: https://doi.org/10.1145/2989081.2989088.

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

[72] Y. Fang and M.-W. Chang. 'Entity Linking on Microblogs with Spatial and Temporal Signals'. In: *TACL* 2 (2014), pp. 259–272. URL: https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/323.

[73] C. Fourrier and B. Sagot. 'Comparing Statistical and Neural Models for Learning Sound Correspondences'. In: *LT4HALA 2020 : First Workshop on Language Technologies for Historical and Ancient Languages*. Due to the COVID-19 pandemic, the workshop will not take place. However, the proceedings are published online. Marseille, France, May 2020. URL: https://hal.inria.fr/hal-02529929.

[74] C. Fourrier and B. Sagot. 'Methodological Aspects of Developing and Managing an Etymological Lexical Resource: Introducing EtymDB 2.0'. In: *LREC 2020 - 12th Language Resources and Evaluation Conference*. Due to COVID19 pandemic, the 12th edition is cancelled. The LREC 2020 Proceedings are available at http://www.lrec-conf.org/proceedings/lrec2020/index.html. Marseille, France, May 2020. URL: https://hal.inria.fr/hal-02678100.

[75] J. E. Hoard, R. Wojcik and 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.

[76] D. Hovy and T. Fornaciari. 'Increasing In-Class Similarity by Retrofitting Embeddings with Demographic Information'. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 671–677. URL: http://aclweb.org/anthology/D18-1070.

[77] D. Hruschka, S. Branford, E. Smith, J. Wilkins, A. Meade, M. Pagel and T. Bhattacharya. 'Detecting Regular Sound Changes in Linguistics as Events of Concerted Evolution'. In: *Current Biology* 1.25 (2015), pp. 1–9.

[78] G. Jawahar, B. Muller, A. Fethi, L. Martin, É. Villemonte de La Clergerie, B. Sagot and D. Seddah. 'ELMoLex: Connecting ELMo and Lexicon features for Dependency Parsing'. In: *CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Brussels, Belgium, Oct. 2018. DOI: 10.18653/v1/K18-2023. URL: https://hal.inria.fr/hal-01959045.

[79] F. Khan, L. Romary, A. Salgado, J. Bowers, M. Khemakhem and T. Tasovac. 'Modelling Etymology in LMF/TEI: The Grande Dicionário Houaiss da Língua Portuguesa Dictionary as a Use Case'. In: *LREC 2020 - 12th Language Resources and Evaluation Conference*. Due to COVID19 pandemic, the 12th edition is cancelled. The LREC 2020 Proceedings are available at http://www.lrec-conf.org/proceedings/lrec2020/index.html. Marseille, France, May 2020. URL: https://hal.inria.fr/hal-02618067.

[80] M. Khemakhem. 'Standard-based Lexical Models for Automatically Structured Dictionaries'. Theses. Université de Paris, Oct. 2020. URL: https://tel.archives-ouvertes.fr/tel-03153438.

[81] M. Khemakhem, L. Foppiano and L. Romary. 'Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields'. In: *electronic lexicography, eLex 2017*. Leiden, Netherlands, Sept. 2017. URL: https://hal.archives-ouvertes.fr/hal-01508868.

[82] M. Khemakhem, S. Gabay, B. Joyeux-Prunel, L. Romary, L. Saint-Raymond and L. Rondeau Du Noyer. 'Information Extraction Workflow for Digitised Entry-based Documents'. In: *DARIAH Annual event 2020*. Zagreb / Virtual, Croatia, May 2020. URL: https://hal.archives-ouvertes.fr/hal-02508549.

[83] S. Kübler, M. Scheutz, E. Baucom and R. Israel. 'Adding Context Information to Part Of Speech Tagging for Dialogues'. In: *NEALT Proceedings Series*. Ed. by M. Dickinson, K. Muurisep and M. Passarotti. Vol. 9. 2010, pp. 115–126.

[84] A.-L. Ligozat, C. Grouin, A. Garcia-Fernandez and D. Bernhard. 'Approches à base de fréquences pour la simplification lexicale'. In: *TALN-RÉCITAL 2013* (2013), p. 493.

[85] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov. 'RoBERTa: A Robustly Optimized BERT Pretraining Approach'. In: *arXiv preprint arXiv:1907.11692* (2019).

[86] L. Martin, S. Humeau, P.-E. Mazaré, A. Bordes, É. Villemonte de La Clergerie and B. Sagot. 'Reference-less Quality Estimation of Text Simplification Systems'. In: *1st Workshop on Automatic Text Adaptation (ATA)*. Tilburg, Netherlands, Nov. 2018. URL: https://hal.inria.fr/hal-01959054.

[87] L. Martin, B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, É. Villemonte de La Clergerie, D. Seddah and B. Sagot. 'CamemBERT: a Tasty French Language Model'. Web site: https://camembert-model.fr. Oct. 2019. URL: https://hal.inria.fr/hal-02445946.

[88] L. Martin, B. Sagot, É. Villemonte de La Clergerie and A. Bordes. 'Controllable Sentence Simplification'. Code and models: https://github.com/facebookresearch/access. Oct. 2019. URL: https://hal.inria.fr/hal-02445874.

[89] H. Martínez Alonso, D. Seddah and 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, Dec. 2016. URL: https://hal.inria.fr/hal-01584054.

[90] P. J. Ortiz Suárez, Y. Dupont, B. Muller, L. Romary and B. Sagot. 'Establishing a New State-of-the-Art for French Named Entity Recognition'. In: *LREC 2020 - 12th Language Resources and Evaluation Conference*. Due to COVID19 pandemic, the 12th edition is cancelled. The LREC 2020 Proceedings are available at http://www.lrec-conf.org/proceedings/lrec2020/index.html. Marseille, France, May 2020. URL: https://hal.inria.fr/hal-02617950.

[91] P. J. Ortiz Suárez, L. Romary and B. Sagot. 'A Monolingual Approach to Contextualized Word Embeddings for Mid-Resource Languages'. In: *ACL 2020 - 58th Annual Meeting of the Association for Computational Linguistics*. Seattle / Virtual, United States, July 2020. DOI: 10.18653/v1/2020.acl-main.156. URL: https://hal.inria.fr/hal-02863875.

[92] P. J. Ortiz Suárez, B. Sagot and L. Romary. 'Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures'. In: *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Ed. by P. Bański, A. Barbaresi, H. Biber, E. Breiteneder, S. Clematide, M. Kupietz, H. Lüngen and C. Iliadi. Cardiff, United Kingdom: Leibniz-Institut für Deutsche Sprache, July 2019. DOI: 10.14618/IDS-PUB-9021. URL: https://hal.inria.fr/hal-02148693.

[93] J. Pyssalo. 'System PIE: the Primary Phoneme Inventory and Sound Law System for Proto-Indo-European'. PhD thesis. University of Helsinki, 2013.

[94] L. Rello, R. Baeza-Yates, S. Bott and 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, p. 15.

[95] L. Rello, R. Baeza-Yates, L. Dempere-Marco and 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.

[96] C. Ribeyre, M. Candito and D. Seddah. 'Semi-Automatic Deep Syntactic Annotations of the French Treebank'. In: *The 13th International Workshop on Treebanks and Linguistic Theories (TLT13)*. Proceedings of TLT 13. Tübingen Universität. Tübingen, Germany, Dec. 2014. URL: https://hal.inria.fr/hal-01089198.

[97] L. Romary, M. Khemakhem, F. Khan, J. Bowers, N. Calzolari, M. George, M. Pet and P. Bański. 'LMF Reloaded'. In: *AsiaLex 2019: Past, Present and Future*. Istanbul, Turkey, June 2019. URL: https://hal.inria.fr/hal-02118319.

[98] A. M. Rush, R. Reichart, M. Collins and 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*. EMNLP-CoNLL '12. Jeju Island, Korea, 2012, pp. 1434–1444.

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

[100] B. Sagot. *External Lexical Information for Multilingual Part-of-Speech Tagging*. Research Report RR-8924. Inria Paris, June 2016. URL: https://hal.inria.fr/hal-01330301.

[101] B. Sagot. 'Extracting an Etymological Database from Wiktionary'. In: *Electronic Lexicography in the 21st century (eLex 2017)*. Leiden, Netherlands, Sept. 2017, pp. 716–728. URL: https://hal.inria.fr/hal-01592061.

[102] B. Sagot and H. Martínez Alonso. 'Improving neural tagging with lexical information'. In: *15th International Conference on Parsing Technologies*. Pisa, Italy, Sept. 2017, pp. 25–31. URL: https://hal.inria.fr/hal-01592055.

[103] B. Sagot, D. Nouvel, V. Mouilleron and 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. URL: https://hal.inria.fr/hal-00832078.

[104] C. Scarton, M. De Oliveira, A. Candido Jr, C. Gasperin and 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.

[105] Y. Scherrer and 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*. European Language Resources Association. Reykjavik, Iceland, May 2014. URL: https://hal.inria.fr/hal-01022298.

[106] S. Schuster, É. Villemonte de La Clergerie, M. Candito, B. Sagot, C. D. Manning and 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*. Proceedings of the 2017 Shared Task on Extrinsic Parser Evaluation. Pisa, Italy, Sept. 2017, pp. 47–59. URL: https://hal.inria.fr/hal-01592051.

[107] R. Schwartz, J. Dodge, N. A. Smith and O. Etzioni. 'Green AI'. In: *Commun. ACM* 63.12 (Nov. 2020), pp. 54–63. DOI: 10.1145/3381831. URL: https://doi.org/10.1145/3381831.

[108] T. Scialom, P.-A. Dray, S. Lamprier, B. Piwowarski, J. Staiano, A. Wang and P. Gallinari. 'QuestEval: Summarization Asks for Fact-based Evaluation'. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 6594–6604. DOI: 10.18653/v1/2021.emnlp-main.529. URL: https://aclanthology.org/2021.emnlp-main.529.

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

[110] D. Seddah, F. Essaidi, A. Fethi, M. Futeral, B. Muller, P. J. Ortiz Suárez, B. Sagot and A. Srivastava. 'Building a User-Generated Content North-African Arabizi Treebank: Tackling Hell'. In: *ACL 2020 - 58th Annual Meeting of the Association for Computational Linguistics*. Seattle / Virtual, Canada, July 2020. DOI: 10.18653/v1/2020.acl-main.107. URL: https://hal.inria.fr/hal-02889804.

[111] D. Seddah, B. Sagot and 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. URL: https://hal.inria.fr/hal-00703124.

[112] D. Seddah, B. Sagot, M. Candito, V. Mouilleron and V. Combet. 'The French Social Media Bank: a Treebank of Noisy User Generated Content'. In: *COLING 2012 - 24th International Conference on Computational Linguistics*. Kay, Martin and Boitet, Christian. Mumbai, India, Dec. 2012. URL: https://hal.inria.fr/hal-00780895.

[113] M. Shardlow. 'A survey of automated text simplification'. In: *International Journal of Advanced Computer Science and Applications* 4.1 (2014), pp. 58–70.

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

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

[116] E. Strubell, A. Ganesh and A. McCallum. 'Energy and Policy Considerations for Deep Learning in NLP'. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 3645–3650. DOI: 10.18653/v1/P19-1355. URL: https://aclanthology.org/P19-1355.

[117] D. Surís, D. Epstein and C. Vondrick. *Globetrotter: Unsupervised Multilingual Translation from Visual Alignment*. 2020. arXiv: 2012.04631 [cs.CL].

[118] É. Villemonte de La Clergerie. 'Jouer avec des analyseurs syntaxiques'. In: *TALN 2014*. ATALA. Marseilles, France, July 2014. URL: https://hal.inria.fr/hal-01005477.

[119] É. Villemonte de La Clergerie, B. Sagot and D. Seddah. 'The ParisNLP entry at the ConLL UD Shared Task 2017: A Tale of a #ParsingTragedy'. In: *Conference on Computational Natural Language Learning*. Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Vancouver, Canada, Aug. 2017, pp. 243–252. DOI: 10.18653/v1/K17-3026. URL: https://hal.inria.fr/hal-01584168.

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

[121] 'When Being Unseen from mBERT is just the Beginning: Handling New Languages With Multilingual Language Models'.