RESEARCH CENTRE

Inria Paris Center

2022 ACTIVITY REPORT

Project-Team ALMANACH

Automatic Language Modelling and Analysis & Computational Humanities

DOMAIN

Perception, Cognition and Interaction

THEME

Language, Speech and Audio

Inria

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]
- Djamé Seddah [Inria, Researcher, on secondment (détachement) from Sorbonne Université]
- Éric Villemonte de La Clergerie [Inria, Researcher]

Post-Doctoral Fellow

- Syrielle Montariol [Inria, until May 2022]

PhD Students

- Robin Algayres [Inria, CoML project-team]
- Roman Castagné [Inria]
- Alix Chagué [Inria & Univ. Montréal]
- Floriane Chiffoleau [Inria]
- Paul-Ambroise Duquenne [META, CIFRE]
- Clémentine Fourrier [Inria, until Sep 2022]
- Matthieu Futeral-Peter [Inria, WILLOW project-team]
- Nathan Godey [Inria]
- Francis Kulumba [MINARM, from Nov 2022]
- Simon Meoni [Arkhn, CIFRE, from Dec 2022]
- Benjamin Muller [Inria, until Aug 2022]
- Tu Anh Nguyen [META, CIFRE]
- Lydia Nishimwe [Inria]
- Pedro Ortiz Suarez [Inria until Feb. 2022, Univ. Mannheim (between March and June 2022), until Jun 2022]
- Arij Riabi [Inria]
- José Rosales Nunez [CNRS]
- Lionel Tadonfouet [ORANGE, CIFRE]
- Rian Touchent [Inria, from Dec 2022]

Technical Staff

- Julien Abadji [Inria, Engineer]
- Jesujoba Alabi [Inria, Engineer, from Feb 2022]
- Wissam Antoun [Inria, Engineer, from Mar 2022]
- Niyati Sanjay Bafna [Inria, Engineer, from Oct 2022]
- Thibault Charmet [Inria, Engineer, until Jan 2022]
- Anna Chepaikina [Inria, Engineer, from Apr 2022]
- Rua Ismail [Inria, Engineer]
- Tanti Kristanti Nugraha [Inria, Engineer]
- Menel Mahamdi [Inria, Engineer, from Sep 2022]
- Virginie Mouilleron [Inria, Engineer, from Dec 2022]
- Hugo Scheithauer [Inria, Engineer]
- Yves Tadjo Takianpi [Inria, Engineer, until Jun 2022]
- Lucas Terriel [Inria, Engineer, until Oct 2022]
- You Zuo [Inria, until Nov 2022]

Interns and Apprentices

- Rishika Bhagwatkar [Inria, Intern, from Jul 2022]
- Galo Castillo Lopez [Inria, Intern, from Jun 2022 until Aug 2022]
- Kelly Christensen [Inria, Intern, from Apr 2022 until Jul 2022]
- Abderraouf Farhi [Inria, Intern, from Apr 2022 until Jul 2022]
- Jules Nuguet [Inria, Intern, from Apr 2022 until Jul 2022]
- Camille Rey [Inria, Intern, until Jun 2022]
- Nacim Talaoubrid [Inria, Intern, from Jun 2022 until Jun 2022]
- Rian Touchent [Inria, Intern, from Jun 2022 until Nov 2022]
- Pierre Vauterin [Inria, Intern, from Jun 2022 until Aug 2022]

Administrative Assistants

- Nathalie Gaudechoux [Inria]
- 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.).

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

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

(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 [97, 145] and one of the best for robust multilingual parsing as shown by our results in various shared tasks over the years [139, 136, 146], [108]. 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 [136].

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 [142]) to propose a joint model of tokenisation, normalisation, morphological analysis and syntactic analysis. We also explore adversarial learning, considering the drastic variation we face in parsing user-generated content and processing historical texts, both seen as noisy input that needs to be handled at training and decoding time.

3.2.2 Integrating context in NLP systems

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

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

A more expressive and context-aware structural representation of a linguistic production is required to accurately model UGC. Let us consider for instance the case for Syntax-based Machine Translation of social media content, as is carried out by the ALMAnaCH-led ANR project Parsiti (PI: DS). A Facebook post may be part of a discussion thread, which may include links to external content. Such information is required for a complete representation of the post's context, and in turn its accurate machine translation. Even for the presumably simpler task of POS tagging of dialogue sequences, the addition of context-based features (namely information about the speaker and dialogue moves) was beneficial [110]. In the case of UGC, working across sentence boundaries was explored for instance, with limited success, by [96] for document-wise parsing and by [128] 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.

[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

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 [94]. 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 [106] 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 [122] 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 [95] and [107]. 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 [141]). 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 [124]. 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 [105, 134], 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 [89]). Lexical simplification, another aspect of text simplification [112, 125], will also be pursued. In this regard, we have already started a collaboration with Facebook's AI Research in Paris, the UNAPEI (the largest French federation of associations defending and supporting people with intellectual disabilities and their families), and the French Secretariat of State in charge of Disabled Persons.

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

3.4 Modelling and Development of Language Resources

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

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

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

3.4.1 Construction, management and automatic annotation of Text Corpora

Corpus creation and management (including automatic annotation) is often a time-consuming and technically challenging task. In many cases, it also raises scientific issues related for instance with

[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.

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 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 [130, 132, 146] 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 [147]), 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 [129, 109, 131], while using and developing (semi)automatic

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

lexical information extraction techniques based on existing corpora [133, 135]. 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 [109, 131].

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) [127] 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 [140, 126, 138, 115]) 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 [100]. 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 (analysis of medical documents, 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

Project	GPU-hours	Real-hours	Power Consumption (kWh)	CO_2 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.

In view of recent interest about the energy consumption and carbon emission of machine learning models, and specifically of those of language models [137, 93], 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 [144]. While the ALMAnaCH team uses other 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.

[7] Jean-Zay documentation
[8] CLEPS documentations
[9] NEF documenation

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 [98] 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 [144] 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].

6 Highlights of the year

In 2022, ALMAnaCH's activity was impacted by a number of Inria-wide issues, amongst which:

- Issues related to the deployment of a new financial and human resource information system (Eksae), which has negatively affected a number of colleagues within the administrative services, with whom researchers directly or indirectly interact. For instance, not being able to easily oversee the team's budget and prepare the budget for next year has required extra effort on behalf of several people, who are already overstretched, including but not limited to our administrative assistant.

- The presentation of the institute given in our general management's addresses and publications, such as the "Rapport d'activité 2021" ("Question d'avenir"), offered a distorted and unbalanced image of our institute, hiding the primary role of research, especially in Artificial Intelligence fields such as ALMAnaCH's. This reduced emphasis on research in the presentation of Inria and the increasing lack of clarity around the future of the status of Inria researchers and the future of the role of Inria as an institution create an anxiety-inducing atmosphere, which directly affects our work and our ability to attract new members.

- The extremely degraded relationship between Inria's *Commission d'Évaluation* and Inria's current general management is difficult to bear. It has created a general lack of confidence in the intentions

[10] Jean-Zay architecture description
[11] Nvidia Tesla V100 specification
[12] Intel Xeon Gold 6248 specification
[13] Rte - éCO$_2$mix.
[14] co2.myclimate.org Estimates

of the general management and in the outcome of future hiring and promotion processes. We wish to thank the Commission d'Évaluation, and in particular its presidency, for its outstanding communication to the researchers it represents and for its continued effort in organising and participating in Inria hiring and promotion committees.

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://github.com/KaMI-tools-project/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-project/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

8 New results

8.1 Large Corpus Creation and Annotation: new advances on the OSCAR corpus

Participants: Pedro Ortiz Suarez, Julien Abadji, Rua Ismail, 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. There are some examples of freely available multilingual corpora for the training of deep learning NLP models, such as the Paracrawl corpus [91] and our own large-scale multilingual corpus OSCAR [121].[15] However, they have quality issues, especially for low-resource languages, an issue investigated in a large-scale study

[15] OSCAR web site.

we were involved in and whose initial publication in 2021 [16] will be followed by a publication in the *Transactions of the Association for Computational Linguistics*.

Recreating and updating these corpora is very complex. Since 2021, we have been developing Ungoliant [88] [19], a new pipeline that improves in a number of ways upon the goclassy pipeline used to create the original OSCAR corpus (known as OSCAR v1 or OSCAR 2019). Ungoliant is faster, modular, parameterisable, and well documented. We have since used it to create several new versions of OSCAR that are larger and based on more recent data [88] [19], and it was also used as the basis for the pipeline to create the ROOTS corpus on which the large multilingual language model BLOOM was trained during the BigScience project (See Section 8.3).

In addition to the improved pipeline, we also have also added additional information to OSCAR. The latest version of OSCAR (version 22.01) includes updated metadata, with language identification performed at the document level, resulting in a new subcorpus containing documents with sentences in multiple languages in significant proportions. Ongoing work includes a large-scale improvement of the language identification mechanism, additional annotations and more. Ungoliant is released under an open source licence and we publish the corpus under a research-only licence.

This work is still carried out in close collaboration with Pedro Ortiz, although he left ALMAnaCH mid-2022 after a successful PhD defence [69] and is now at DFKI Berlin. Other people outside of ALMAnaCH regularly interact with the "OSCAR team", especially Sebastian Nagel, from CommonCrawl.

8.2 Neural Language Modelling

Participants: Benoît Sagot, Djamé Seddah, Rachel Bawden, Pedro Ortiz Suarez, Arij Riabi, Roman Castagné, Wissam Antoun.

Pretrained language models are now ubiquitous in Natural Language Processing. Despite their success, many available models were traditionally trained only on English or on a selection of languages in a multilingual setup [99, 113]. This made practical use of such models limited, in all languages except English. One of the most visible achievements of the ALMAnaCH team was the training and release of CamemBERT in 2019, a BERT-like [99] (and more specifically a RoBERTa-like) neural language model for French trained on the French section of our large-scale web-based OSCAR corpus [121] (see Section 8.1), together with CamemBERT variants [114] and ELMo models trained on OSCAR corpora for other languages, including French [119, 120].

Since 2021, we have been investigating 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 several 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 have shown that a character-based model trained on only 99k sentences of NArabizi and fined-tuned on a small NArabizi treebank leads to performance close to that of the same architecture pre-trained on large multilingual and monolingual models [39]. We also confirmed these results a on much larger data set of noisy French user-generated content. In 2022, we have also been recently investigating character-based models in multilingual machine translation (MT) as a possible way to increase lexical sharing between related languages (described in more detail in Section 8.5) as well as in the context of modelling historical language change for cognate prediction (described in Section 8.7).

Going beyond character-based models, in 2022 we focused on learning tokenisation within the language modelling architecture. We proposed MANTa, a Module for Adaptive Neural TokenizAtion [31]. MANTa is a differentiable tokeniser trained end-to-end with the language model. The resulting system offers a trade-off between the expressiveness of byte-level models and the speed of models trained using subword tokenisation. In addition, MANTA's tokenisation is highly explainable since it produces an explicit segmentation of sequences into blocks. We found that MANTa improves robustness to character perturbations and out-of-domain data, and performs comparably to other models on the general-domain GLUE benchmark [148]. Finally, we show that it is considerably faster than standard byte-level models.

We also participated in the BigScience initiative, for which see Section 8.3 and developed language models for historical French (Early Modern French), which are described in Section 8.6 of this report.

8.3 Participation in the BigScience Initiative

Participants: Benoît Sagot, Rachel Bawden, Pedro Ortiz Suarez, Roman Castagné, Clémentine Fourrier.

Finally, several ALMAnaCH members 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 250+ institutions. The goal of the project, supported by the French state in the form of large grants of computing power for the Jean Zay public supercomputer, was to create a publicly available, very large multilingual neural network language model and a publicly available, very large multilingual text dataset.

ALMAnaCH members were 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 was one of the two chairs, published a survey tokenisation in early 2022 [77] of which an updated version is in preparation. T0, the first language model of interest successfully trained and evaluated within the project also resulted in a publication, to which several ALMAnaCH members involved in the project contributed [40].

All ALMAnaCH participants co-authored the main paper describing the initiative as a whole and the resulting large-scale model, BLOOM, presented as "the world's largest open multilingual language model." [80]. Several ALMAnaCH members played a key role in developing BLOOM's training corpus [76]. As mentioned in Section 8.1, the pipeline used in BigScience to create the training corpus, sometimes referred to as the "ROOTS pipeline" (after the name given to the BLOOM training corpus), is heavily based on Ungoliant (see Section8.1), and that OSCAR represents 38% of the ROOTS corpus. Other ALMAnaCH members also participated in task-specific (partial) evaluations of BLOOM's downstream performance, notably for several shallow tasks on historical texts [41] and for MT.

8.4 Cross-lingual Transfer Learning for Low-resource Non-standard Languages

Participants: Djamé Seddah, Benoît Sagot, Benjamin Muller, Niyati Sanjay Bafna, Rachel Bawden.

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. In the context of Benjamin Muller's PhD theses, and expanding from previous work on North-African Arabic to many other languages that share similar properties, since 2020 we have been focusing on understanding the inner workings of large multilingual neural language models when fine-tuned on cross-language scenarios with various degrees of resource availability (from large to extremely scarce). This work has led to a number of publications in previous years [143, 116, 117]. One of them was re-presented this year at the French national conference TALN [43], and Benjamin Muller also successfully defended his PhD in 2022 [68].

In more recent work and as part of our collaboration with the DFKI involving the joint supervision of Niyati Bafna, we have been working on transfer learning between Hindi and related low-resource dialects of the Hindi belt. Given the lexical similarity between the dialects, transferring from standard Hindi is a promising direction, and we have been working on the training of language models that are robust to such variation through data augmentation and character-level models.

8.5 Text-based Machine Translation

[16] BigScience web site.

Participants: Rachel Bawden, Benoît Sagot, Djamé Seddah, Jose Rosales Nunez, Camille Rey, Lydia Nishimwe, Sonal Sannigrahi, Jesujoba Alabi, Benjamin Muller.

The MT research axis has been further reinforced. 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 and domain shift are among our key research topics. This section describes work in text-based MT. For speech-based MT and multimodal MT, see Sections 8.9 and 8.8 respectively.

The study of the robustness of MT models to non-standard language typically found in user-generated content (UGC) on social media sites has continued to be a major focus, with the continuation of José Rosales Nunez's PhD thesis co-supervised by Djamé Seddah with the Laboratoire de Linguistique Formelle (Université de Paris) and a new PhD thesis on the topic with Lydia Nishimwe. In the context of Lydia's thesis, we have been exploring techniques for the normalisation of non-standard to contemporary French and approaches to improve the robustness of language models by providing alternative loss functions. We are also designing and preparing a parallel test set of UGC content for the organisation of a shared task at the WMT 20223 conference.

We have also been looking into MT for low-resource and/or related languages. In collaboration with former colleagues at the University of Edinburgh, Rachel Bawden co-authored a survey on low-resource MT methods in the *Computational Linguistics Journal* in 2022. We have also been working on MT for low-resource related languages through multilingual models or in transfer settings. In the context of Sonal Sannigrahi's internship in 2021, we began exploring the effect of subword segmentation and transliteration to encourage language sharing in a multilingual setup for related (India-Aryan) languages written in different scripts. This work has continued in 2022 in view to a submission to the EAMT 2023 conference. We also explored a similar research direction but for related higher-resource (Slavic) languages as a participation to the WMT 2022 general translation task [20]. Our findings seem to agree in both experiments that transliteration does not seem to help language sharing (even in the lower resource setup) and that multilingual models tend to learn the mapping between different scripts without transliteration.

Another area of research we have been exploring is domain adaptation. In the context of the DadaNMT project led by Rachel Bawden, we have been testing the inclusion of dictionary definitions within MT models in order to compensate for vocabulary limitations when shifting between domains We envisage to submit this work at EAMT 2023, presenting an analysis of what does and does not perform well. In terms of testing MT in different domains, Rachel Bawden is a member of the organising committee of two shared tasks at WMT: the general translation task [33] and the biomedical translation task [37].

A problem that is still relevant today, despite the progress made in neural MT, is lexical ambiguity (a problem that can even be exacerbated by the translation of multiple domains). Camille Rey's internship sought to tackle this problem for English-to-French translation by coupling the standard neural MT loss with a contrastive loss trained on contrastive translations of ambiguous lexical items. The experiments show that the balance between the two losses is delicate, and it is difficlt not to introduce noise through the autoamtic creation of hte training dta. However, it coudl offer a promising perspective for future work for less-resourced languge pairs.

Finally, an important aspect of MT research is the development of effective automatic metrics. We have continued our collaboration with two researchers at the MILA research institute (Yu Lu Liu and Jackie Chi Kit Cheung), Canada and Thomas Scialom, META AI on the development of a new automatic language-model-absed metric for language generation tasks. Although we have not yet finalised the evaluation of the metric for MT itself, we have some interesting results for the similar tasks of text summarisation and simplification and transfer learning between the two tasks [75].

8.6 NLP for Early Modern French

Participants: Rachel Bawden, Pedro Ortiz Suarez, Benoît Sagot.

Early Modern French (also known as Modern French or classical 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 continued our collaboration with several researchers outside Inria (Simon Gabay from the University of Geneva and Philippe Gambette from Université Gustave-Eiffel) to develop corpora and NLP tools adapted to Early Modern French, an initiative that we call the FreEM project [42]. 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 2022, we continued our work on various aspects of the processing of Early Modern French, in collaboration with Simon Gabay (Université de Lausanne) and Philippe Gambette (Université Paris-Est): 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 in 2021 was 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 [102].

From the FreEM-*max* data, we trained a RoBERTa language model [113] for Early Modern French, which we call D'AlemBERT [30], which can be used to boost the performance across various tasks through pretraining. We have shown that it can boost the performance of PoS tagging for Early Modern French [30], that it seems to display good transfer learning potential to other periods of French with less available data, and that it can also be beneficial for named entity recognition [118]. From the FreEM-*norm* data, we developed automatic normalisation models, including rule-based and machine-translation (MT)-style approaches [22]. As well as providing normalisation, which can be a useful step before manual analysis or the application of other downstream tasks, we also exploited the models as a way to generate synthetic Early Modern French data on a large scale, which we used both to analyse the patterns of spelling change across different decades of the 17th century [101] [29] and to compare the feasibility of using synthetic data as opposed to real data [50].

Following on from work in 2021 to integrate the NLP tools developed into a processing pipeline for Early Modern French texts and encoded in TEI [92], we have continued to improve the annotation pipeline in 2022 [59]. The trained models are also freely available under open licences on the HuggingFace platform.

8.7 Cognate Prediction Using Neural Machine Translation Techniques

Participants: Clémentine Fourrier, Benoît Sagot, Rachel Bawden.

In 2022 we resumed our experiments, carried out in the context of Clémentine Fourrier's PhD thesis [67], to investigate whether, how and under which conditions neural networks can be used to learn sound correspondences between two related languages using character-based neural MT models. In particular, we focused on understanding how such models work and what they learn. So far, all linguistic interpretations about latent information captured by such models have been based on external analysis (accuracy, raw results, errors). We focused on studying what probing can tell us about both models and previous interpretations, and learnt that although our models store linguistic and diachronic information, they do not achieve it in previously assumed ways [28].

8.8 Multimodal Machine Translation

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

Image-enhanced MT In the context of Matthieu Futeral-Peter's PhD thesis and in collaboration with Ivan Laptev and Cordelia Schmid, his co-supervisors from the WILLOW project-team, we have beenn investigating how using image data can improve the processing of text. In 2022 we focused on the incorporation of the image modality in MT, i.e. Multimodal Machine Translation (MMT). One of the major challenges of MT is ambiguity, which can in some cases be resolved by accompanying context such as an image. However, recent work in MMT has shown that obtaining improvements from images is challenging, limited not only by the difficulty of building effective cross-modal representations but also by the lack of specific evaluation and training data. We developed and presented a new MMT approach based on a strong text-only MT model, which uses neural adapters and a novel guided self-attention mechanism and which is jointly trained on both visual masking and MMT [74]. We also released CoMMuTE, a Contrastive Multilingual Multimodal Translation Evaluation dataset, composed of ambiguous sentences and their possible translations, accompanied by disambiguating images corresponding to each translation. Our approach obtains competitive results over strong text-only models on standard English-to-French benchmarks and outperforms these baselines and state-of-the-art MMT systems with a large margin on our contrastive test set.

Bridging the gap between text and speech in MT using sentence embeddings In the context of Paul-Ambroise Duquenne's PhD thesis and in collaboration with Holger Schwenk, his co-superviser at META, we proposed a new modular architecture for text and speech translation, which is based on a common fixed-size multilingual and multimodal internal representation, and encoders and decoders which are independently trained [25]. We explored several variants of teacher-student training to learn text and speech encoders for multiple languages, which are compatible with the embedding space of the LASER encoder [90]. In contrast to preceding works on multilingual and multimodal representations, we also trained text decoders for multiple languages which are able to generate translations given the joint representation. Finally, we demonstrated that it is possible to train a speech decoder using raw audio only. We showed that these encoders and decoders can be freely combined to achieve very competitive performance in T2T, S2T and (zero-shot) S2S translation.

8.9 Speech modelling

Participants: Tu Anh Nguyen, Robin Algayres, Benoît Sagot.

Unsupervised text-based language modelling techniques and challenges are an interesting source of ideas when trying to perform unsupervised speech modelling. Speech can be made more text-like (e.g. using discrete sound units, word-like segmentation) or, conversely, language modelling techniques can be adapted to the intrinsically continuous nature of speech (e.g. using contrastive losses). We investigate these questions in collaboration with the CoML project-team at Inria Paris and with specialists of speech processing at META.

Modelling raw speech: unsupervised sequence embedding and "word" boundary detection Finding word boundaries in continuous speech is challenging as there is little or no equivalent of a "space" delimiter between words. Popular Bayesian non-parametric models for text segmentation [104, 103] use a Dirichlet process to jointly segment sentences and build a lexicon of word types. In the context of Robin Algayres's PhD thesis, co-supervised by Emmanuel Dupoux, from the CoML team at Inria Paris, EHESS and ENS, we introduced a new algorithm, named DP-Parse, which uses similar principles but only relies on an instance lexicon of word tokens, avoiding the clustering errors that arise with a lexicon of word types [14]. On the Zero Resource Speech Benchmark 2017 speech segmentation task, our model sets a new state of the art in 5 languages. The algorithm monotonically improves with better input representations,

achieving even higher scores when fed with weakly supervised inputs. Despite lacking a type lexicon, DP-Parse can be pipelined to a language model and learn semantic and syntactic representations as assessed by a new spoken word embedding benchmark.

We also introduced a simple neural encoder architecture that can be trained using an unsupervised contrastive learning objective and which gets its positive samples from data-augmented k-Nearest Neighbors search [21]. We show that when built on top of recent self-supervised audio representations (MFCCs, CPC, HuBERT and Wav2Vec 2.0, both Base and Large), this method can be applied iteratively and yield competitive performance as evaluated on two tasks: query-by-example of random sequences of speech, and spoken term discovery—a task that can be key to finding "word" boundaries, including in DPParse. On both tasks our method pushes the state of the art by a significant margin across 5 different languages. Finally, we establish a benchmark on a query-by-example task on the LibriSpeech dataset to monitor future improvements in the field.

Questioning further how speech sequences can be represented, we investigated the impact of relying first on transforming the audio into a sequence of discrete units (or pseudo-text) and then training a language model directly on such pseudo-text, as is commonly done. Note that this can be achieved either in a supervised way or unsupervisedly, as in our work cited above. In the context of Tu Anh Nguyen's PhD thesis, also co-supervised by Emmanuel Dupoux (this time as a META fellow), our scientific question was the following: is such a discrete bottleneck necessary, potentially introducing irreversible errors in the encoding of the speech signal, or could we learn a language model without discrete units at all? We studied the role of discrete versus continuous representations in spoken language modelling and showed that discretisation is indeed essential [17]. We showed that it removes linguistically irrelevant information from the continuous features, helping to improve language modeling performances. On the basis of this study, we trained a language model on the discrete units of the HuBERT features, reaching new state-of-the-art results in the lexical, syntactic and semantic metrics of the Zero Resource Speech Challenge 2021 (Track 1-Speech Only).

Dialogue-level speech language modelling In collaboration with a number of colleagues from META and also in the context of Tu Anh Nguyen's PhD thesis, we introduced dGSLM, the first "textless" model able to generate audio samples of naturalistic spoken dialogues [78]. It uses recent work on unsupervised spoken unit discovery (HuBERT + k-NN clustering) coupled with a dual-tower transformer architecture with cross-attention trained on 2000 hours of two-channel raw conversational audio (Fisher dataset) without any text or labels. We showed that our model is able to generate speech, laughter and other paralinguistic signals in the two channels simultaneously and reproduces more naturalistic and fluid turn-taking compared to a text-based cascaded model.

8.10 Hate speech detection

Participants: Djamé Seddah, Syrielle Montariol, Arij Riabi, Wissam Antoun.

In order to support the fight against radicalisation and prevent future terrorist attacks from taking place, the CounteR H2020 project brings data from diverse 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 looking at the community and its related risk factors rather than targeting and monitoring individuals.

ALMAnaCH is responsible for most of the NLP component of the project. However, issues regarding the availability of the input data prevented us from working specifically on CounteR data until very late in the project (Summer 2022). We have therefore focused on a set of adjacent domain tasks of significant social importance that are similar to CounteR's needs, namely hate speech detection towards migrants and towards women, with a focus on zero-shot cross-lingual transfer.

Zero-shot hate-speech detection in a cross-lingual scenario Given that many of our target languages are considered low-resource, we devoted a large amount of time to experiments in various scenarios

to decide which auxiliary tasks could help alleviate the lack of target domain data. Our experiments showed that in such scenarios, Named entity recognition and sentiment analysis were tasks likely to boost the classifiers as they seem to provide interesting features that boost classification performance. Interestingly, the morpho-syntactic tasks helped in the case of related languages (training data in Spanish, target language in Italian for example). Because of these results, we expanded our north-African Arabic dialect data set with a named-entity annotation layer and a sentiment analysis layer.

We conducted an extensive set of experiments in order to validate what to do in the worst case scenario (not enough data to train a classifier for a target language). Using hate-speech data sets for Italian, English and Spanish, we were able to (i) find the optimal set of auxiliary tasks in this scenario (ii) find the right architecture and (iii) optimise our models in this case. We published our findings in the prestigious AACL conference (Asian Chapter of the Association for computational linguistics) as well as in the French NLP conference, TALN, see [35, 44]

Class unbalance and multimodal classification of hateful memes One of the issues we are currently facing is the class unbalance issue that arises when labels are not evenly distributed in a dataset. As shown in our report on the radicalisation dataset, this is an important issue. Given that this was also a problem we faced when working on the related task of hate speech, we investigated and developed methods to cope with it in a multimodal setting. The architecture we proposed was based on a joint encoding of text and images using different encoders (visualBERT [111], CLIP [123] and OFA [149]), which we used to extract information from hateful memes and their associated texts. Our promising results showed that we were able to provide accurate classification of multimodal content. This work was published in the Constraints Workshop at ACL 2022 [36].

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 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 judgments between pairs of keyword sequences that we manually collected with experts from the *Cour*. The motivations, methods and results were published this year [23] and the models are freely available on the HuggingFace platform, including a model for the semi-automatic completion of keyword sequences.[17]

8.12 Patent classification

Participants: You Zuo, Benoît Sagot.

In the context of our collaboration with the INPI (the French National Institute for Intellectual Property) and in collaboration with Kim Gerdes (from the Inria- and ALMAnaCH-supported start-up company qatent and from University Paris-Saclay), we have explored patent classification into classes at multiple granularity levels, which correspond to levels in the hierarchical and very fine-grained standard patent classification scheme IPC.

[17] Link to the automatic model and the semi-automatic model.

Most previous patent classification methods have treated the task as a general text classification task, and others have tried to implement XML (eXtreme Multi-label Learning) methods designed to handle vast numbers of classes. However, they focus only on the IPC subclass level, which has fewer than 700 labels and is far from "extreme." We extracted from the INPI internal database a Corpus of French patents, which contains all parts of patent texts (title, abstract, claims and description) published from 2002 to 2021, with IPC labels at all levels. We tested different XML methods and other classification models at the subclass and group levels of the INPI-CLS dataset with about 600 and 7k labels, respectively, demonstrating the XML approach's validity to patent classification [63].

8.13 Information Extraction from Specialised Collections

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

The DataCatalogue project, jointly led with the Bibliothèque nationale de France (BnF) and the Institut national d'histoire de l'art (INHA), culminated in October 2022 with a dedicated conference showcasing the results of the project's research and development efforts, as well as other projects oriented on information extraction [54]. The development made during this year showed that the GROBID suite can be customised to parse and restructure sales catalogues into TEI-XML. The GROBID DataCatalogue module, including all associated models and datasets we created, has been made publicly available. The project was also presented during the "Atelier Culture-Inria" [61], March 2022, alongside the NER4Archives project, which ended in December 2021 [48].

Building upon our extensive experience with the XML-TEI format, we have actively participated in the scientific community's endeavours by creating a specific data model for the DataCatalogue project, and by contributing to the TEI annual conference. It took place in September 2022, and we organised a workshop oriented towards users wishing to digitise and publish historical documents [47]. Additionally, the Gallic(orpor)a project came to an end with the presentation of findings on a possible TEI encoding of the data resulting from automatic transcription pipelines [53].

Our collaboration with Assistance Publique – Hôpitaux de Paris (APHP), which began in Spring 2020 in the context of the Covid pandemic, is still prospering. The GROBID suite has been customised to effectively parse the various medical reports and documents linked to patients, paving the way for its use by medical practitioners and administrative staff. This year, we focused on incorporating named entity recognition in the GROBID workflow, enabling the generation of enriched documents (see the detailed report).

Our long-standing collaboration with the European Holocaust Research Infrastructure (ERHI) continued in 2022. We helped implement extraction information tools, such as named entity recognition and entity linking.

Even after the end of the LECTAUREP project in 2021, we have continued to disseminate the methodology and results obtained during the project, which was carried out in collaboration with the French national archives (2018-2021) [56, 57, 60, 55]. The release of the DHNord2019 conference proceedings has also led to a resurgence of the TIMEUS project in the team's activities [65, 66].

The annual DH conference provided us with an opportunity to gather our collective expertise and insights in order to establish a shared vision for a digitisation and publication pipeline dedicated to historical documents. It was a privilege to present it to the wider scientific community during this event, which took place in July 2022 [46].

We were involved in the creation of a project, supported by a DARIAH grant focusing on workflows, dedicated to harmonising HTR and OCR workflows within publication pipelines for historical documents. The project, called "Harmonizing Automatic Text Recognition", received funding in December 2022, and will officially be launched in February 2023. We are pleased to extend the findings from the DAHN project [72] and to expand our network through numerous collaborations with European institutions, including the Institut historique allemand (IHA/DHI) and the Staatbibliothek zu Berlin, among others.

8.14 Automatic Text Recognition on Historical Documents

Participants: Laurent Romary, Alix Chagué, Floriane Chiffoleau, Hugo Scheithauer, Yves Tadjo.

Drawing on our team's experience gained through collaborative work with patrimonial institutions, we have developed significant expertise in the field of automatic text recognition, specifically when applied to historical documents and manuscripts. This expertise has been further expanded through our engagement with the eScriptorium/Kraken solution, developed by SCRIPTA PSL. The two corresponding PhD positions which started at the end of 2021 demonstrate our team's commitment to these areas of research. We have continued to make significant contributions to the development of the eScriptorium application, including participating in weekly developer and design meetings, and direct contributions to the source code. Thanks to the developer position funded by LECTAUREP until mid-2022, we have been able to provide valuable insights and use cases to expand the reflections on the implementation of new features for the application. To enhance the accessibility and comprehensibility of HTR-related technologies, we have engaged in animating workshops with Scripta (during the DH2022 conference and the Point HTR conference organised by the BnF Datalab) and publishing in professional journals and specialised blogs [83]. After the initial tutorial built during the LECTAUREP project, we initiated the development of a better infrastructure to publish extensive and collaborative documentation on eScriptorium's functionalities. This documentation project will be continued in 2023. CREMMA's framework allowed us to fund the creation of more training data for the automatic recognition of handwritten texts in French over a period of 800 years. As a result, a total of 12 new datasets [73] were published within the HTR-United organisation, and their description added to the corresponding catalogue, thereby contributing to its improvement. In the meantime, HTR-United gained greater visibility in France and beyond [45, 82], extending its reach beyond the eScriptorium user community to include other HTR platforms such as Transkribus; the description model for training datasets was also refined based on users' feedback. Data created thanks to CREMMA and collected through HTR-United allowed us to develop a powerful generic model, Manu McFrench, for HTR on French documents from the modern and contemporary periods, which will be presented in 2023 during the ADHO annual conference. Despite the fact that the deployment of the CREMMA-funded server was not completed in 2022, we continued to grant access to our current eScriptorium infrastructure (Traces6), which was funded by the DAHN project. In 2022, we added nearly 100 new users. Towards the end of 2022, we introduced CREMMACall, a platform that simplifies the process of requesting access to the upcoming CREMMA server. By launching CREMMACall, we are able to gain a better understanding of the projects that are seeking access to our application. This will enable us to encourage them to adhere to the FAIR principles, such as using HTR-United to register any training data they may generate. This will promote transparency and accountability within the community.

9 Bilateral contracts and grants with industry

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

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 2022 began a second step for this collaboration, which involves a dedicated research engineer whose 1-year contract will be extended by another 6 months.

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 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.

Participants: Laurent Romary, Floriane Chiffoleau, Lucas Terriel.

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 A.E. (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 monitoring 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.

Participants: Djamé Seddah, Arij Riabi, Syrielle Montariol, Wissam Antoun, Galo Castillo Lopez.

10.1.2 Collaborations in European Programs, except FP7 and H2020

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.

Participants: Laurent Romary.

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.

Participants: Djamé Seddah, José Rosales, Benjamin Muller, Benoît Sagot, Éric Villemonte de La Clergerie.

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).

Participants: Laurent Romary, Benoît Sagot, Pedro Ortiz Suarez.

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.

Participants: Benoît Sagot.

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 AI
- faurecia
- Google
- Idemia
- Janssen
- Naver Labs
- Nokia
- Pfizer
- Stellantis
- Valeo
- Vertex
- DXOMARK
- Avatar Medical
- Kayrros
- Simplicity
- Sonio

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. Other participants are funded by one or the other of these two chairs.

Participants: Benoît Sagot, Rachel Bawden, Julien Abadji, Lydia Nishimwe, Nathan Godey, Roman Castagné, Rua Ismail.

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]).

Participants: Benoît Sagot, Djamé Seddah, Éric Villemonte de La Clergerie, Virginie Mouilleron, Nacim Talaoubrid, Pierre Vauterin, Rishika Bhagwatkar.

GDR LiLT

Duration: 1 Jan 2019–present.

Summary: Linguistic issues in language technology.

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

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).

Participants: Benoît Sagot, Éric Villemonte de La Clergerie.

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.

Participants: Laurent Romary, Floriane Chiffoleau, Alix Chagué, Yves Tadjo Takianpi.

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).

Participants: Benoît Sagot.

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.

Participants: Laurent Romary.

Convention (MIC) DataCatalogue

Duration: 12 Aug 2021–12 Dec 2022.

PI: Laurent Romary.

Coordinator for ALMAnaCH: Laurent Romary.

Partners:
- 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.

Participants: Laurent Romary, Hugo Scheithauer, Jules Nuguet, Abderraouf Farhi.

Sorbonne Emergence DAdaNMT

Duration: 1 Feb 2022–31 Jan 2023.

PI: Rachel Bawden.

Coordinator for ALMAnaCH: Rachel Bawden.

Summary: The aim of this project is to investigate domain adaptation for neural machine translation. We will be exploring the adaptation of models to specific, low-resource domains domains as well as training models for multiple domains.

Participants: Rachel Bawden, Jesujoba Alabi.

Contrat PIA (AMI santé numérique) OncoLab

Duration: 1 Mar 2022–1 Mar 2026.

PI: Éric de La Clergerie.

Partners:
- Arkhn
- Owkin
- Institut universitaire du cancer de Toulouse Oncopole
- Institut Curie
- Institut Bergonié
- CHU de Toulouse

Summary: The aim of the project is to make cancer data from health institutions accessible to all stakeholders involved for research and innovation purposes. The data at hand will be standardised and structured, in particular by extracting information from textual documents.

Participants: Éric de La Clergerie, Laurent Romary, Rian Touchent, Simon Meoni.

10.3 Regional initiatives

Framework agreement with Inria AP-TAL

Duration: 1 Apr 2020–present.

PIs: Laurent Romary.

Coordinator for ALMAnaCH: Laurent Romary.

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 to 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.

Participants: Laurent Romary, Tanti Kristanti.

11 Dissemination

Participants: Benoît Sagot, Laurent Romary, Éric de La Clergerie, Djamé Seddah, Rachel Bawden, Alix Chagué, Hugo Scheithauer, Floriane Chiffoleau, Tu Anh Nguyen, Lucas Terriel, Nathan Godey, Niyati Bafna, Lydia Nishimwe, Syrielle Montariol
.

11.1 Promoting scientific activities

11.1.1 Scientific events: organisation

Member of the organizing committees

- Floriane Chiffoleau: Member of the organising committee for DAHTR (Documents anciens et reconnaissance automatique des écritures manuscrites).
- Rachel Bawden: Member of the organising committee for WMT general and biomedical shared tasks.

Inria-internal events

- Djamé Seddah: Organisation of the ALMAnaCH seminar series (Nov 2017–Jul 2022)
- Rachel Bawden: Organisation of the ALMAnaCH reading group (Feb 2021–Nov 2022) and the ALMAnaCH seminar series (Jul 2022–present). 14 seminars in 2022.
- Niyati Bafna and Wissam Antoun: Organisation of the ALMAnaCH reading group (Nov 2022–present).

11.1.2 Scientific events: selection

Reviewer

- Alix Chagué: Reviewer for DAHTR (Documents anciens et reconnaissance automatique des écritures manuscrites).
- Éric de La Clergerie: Reviewer for GURT/SyntaxFest, EMNLP 2022, COLING 2022, CoG2022 (IEEE Conference on Games), Toth 2022 a nd LREC 2022.
- Nathan Godey: Reviewer for EMNLP 2022.
- Rachel Bawden : Reviewer for ACL Rolling Reviews, COLING 2022, RECITAL 2022, ACL SRW 2022 (Student research workshop) and EACL 2023.
- Benoît Sagot: Reviewer for Fifth Workshop on Computational Methods for Endangered Languages (ComputEL-5) and ACL 2022.
- Djamé Seddah: Reviewer for ACL 2022, EMNLP 2022, EACL 2023 and COLING 2022.

11.1.3 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)*.
- Éric de La Clergerie: Reviewer for *IEEE Transactions on Games* and *Pattern Recognition Letters*.
- Rachel Bawden: Reviewer for *Dialogue & Discourse*.
- Benoît Sagot: Reviewer for *ACM Computing Surveys*.
- Tu Anh Nguyen: Reviewer for *IEEE Journal of Selected Topics in Signal Processing*.

11.1.4 Invited talks

- Rachel Bawden:
  - Advanced Language Processing Winter School (ALPS) 2022 (19 Jan 2022): Social session on "Data resources in NLP: how to construct them and how to evaluate resource papers"".
  - LIFT TAL 2022: Journées Jointes des Groupements de Recherche "Linguistique Informatique, Formelle et de Terrain" et "Traitement Automatique des Langues" (15 Mar 2022): "Traduction automatique dans des scénarios à faibles ressources : application à des questions linguistiques".
  - Multispeech, Loria, Nancy (3 Mar 2022): "Low-resource MT: Few-shot learning and historical language normalisation".
  - Cognitive Science Research Group, Queen Mary University of London, UK (9 Mar 2022): "Low-resource MT: Few-shot learning and historical language normalisation".
  - Round table on "Réduction des biais et des inégalités : L'IA peut-elle avoir un rôle ?", organised by 3IA Côte d'Azur in collaboration with the 3IA (ANITI, MIAI, PR[AI]RIE) (22 Mar 2022): "Traitement automatique des langues pour tous : l'importance de s'adapter au langage varié des réseaux sociaux".
  - Pôle Modélisation du laboratoire MoDyCo, Université Paris Nanterre (25 Oct 2022): "Traduction Automatique et Linguistique : une relation en crise ?"

- Benoît Sagot:
  - Société de Linguistique de Paris (19 Feb 2022): “Le traitement automatique des langues au service de la linguistique : quelques travaux à l’interface entre linguistique et informatique”.
  - Language Technologies & Digital Humanities Conference 2022, Ljubljana, Slovenia (16 Sep 2022): “Large-scale language models: challenges and perspectives”.
  - Czech-French AI Workshop on Artificial Intelligence, Prague, Czech Republic (12 Sep 2022): “Large-scale Language Models and Their Training Corpora”.
- Hugo Scheithauer:
  - Atelier Culture-Inria, Archives Nationales, (22 Mar 2022): “DataCatalogue : présentation du projet”. Joint with Laurent Romary, Federico Nurra and Frédérique Duyrat.
  - Transkribus / eScriptorium : Transcrire, annoter et éditer numériquement des documents d’archives, Bibliothèque nationale de France (9 May 2022): “LectAuRep : Données d’archives en français des XIXe et XXe siècles”.
  - DHNord 2022, online (21 Jun 2022): “Enrichir le patrimoine écrit archivistique grâce aux technologies numériques : ingénierie du projet LectAuRep”. Joint with Aurélia Rostaing.
  - Conseil Scientifique de la Bibliothèque nationale de France (28 Jun 2022): “DataCatalogue : ingénierie de projet”. Joint with Frédérique Duyrat.
  - Segmenter et annoter les images : déconstruire pour reconstruire, Institut national d’histoire de l’art, Paris (15 Nov 2022): “LectAuRep (2018-2021) : Projet de lecture automatique de répertoires de notaires”. Joint with Aurélia Rostaing.
  - La reconnaissance des écritures manuscrites et ses usages dans les archives (29 Nov 2022): “LectAuRep : un projet de recherche et développement pour la transcription automatique des répertoires de notaires”. Joint with Aurélia Rostaing.
- Lucas Terriel:
  - Atelier Culture-Inria, Archives Nationales, (22 Mar 2022): “NER4Archives (named entity recognition for archives) : Conception et réalisation d’un outil de détection, de classification et de résolution des entités nommées dans les instruments de recherche archivistiques encodés en XML/EAD”.
- Alix Chagué:
  - Documents anciens et reconnaissance automatique des écritures manuscrites, École nationale des chartes, Paris (24 Jun 2022): “Sharing HTR datasets with standardized metadata: the HTR-United initiative”. Joint with Thibault Clérice.
  - Chairing a Round Table at DHNord2022, Lille (22 Jun 2022): “Enjeux et débats autour de l’organisation du travail dans les projets en humanités numériques”.
- Niyati Bafna:
  - Linguistic Mondays, Charles University (24 Oct 2022): “Empirical Models for an Indic Dialect Continuum”.
- Djamé Seddah:
  - ENS Lyon (Online) (21 Feb 2022): “CamemBert must die! (jk,lol)”.
  - Journée d’étude ATALA “Robustesse des Systèmes de TAL”, Maison de la Recherche, Sorbonne Univ. (25 Nov 2022): “Beyond Sesame street-based naming schemes: Camembert must Die (jk, Lol)”.

11.1.5 Scientific expertise

- Lydia Nishimwe:
  - Expert for Consultation by the Cour des Comptes (May 2022).
- Rachel Bawden:
  - Expert for Consultation by the Cour des Comptes (May 2022).
- Benoît Sagot:
  - Expert for Inria focus group on the evolution of the administrative research assistant position.

11.1.6 Research administration

- Éric de La Clergerie:
  - Member of the scientific board for the RECOLNAT infrastructure (Website).
- Rachel Bawden:
  - Member of the scientific board for Société de Linguistique de Paris (Administratrice).
- 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.
- Benoît Sagot:
  - 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)).
  - Member of the scientific board for Société de Linguistique de Paris (Administrateur).

11.2 Teaching - Supervision - Juries

11.2.1 Teaching

- 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.
  - 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.
  - Tutorial for the workshop "HTR@BnF", Bibliothèque nationale de France, Paris. eScriptorium (4 hours), coorganised with Daniel Stoekl Ben Ezra;Peter Stokes. Bibliothèque nationale de France, Paris.10 May 2022.
  - Tutorial for the eScriptorium workshop at the DH2022 conference, University of Tokyo. Hands-on Introduction to eScriptorium, an Open-Source Platform for HTR (3.5 hours), coorganised with Daniel Stoekl Ben Ezra;Peter Stokes. University of Tokyo.26 Jul 2022.
- 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.
  - Master's course (M2) as part of the Master "Documentation et Humanités Numériques". Introduction à Python et à l'algorithmie (9 hours). École du Louvre, France.
  - Master's course (M2) as part of the Master "Documentation et Humanités Numériques". Collecte et publication de jeux de données avec Python (4.3 hours). école du Louvre, France.
  - 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.

- Syrielle Montariol:
  - Master's course (M2) as part of the Master in statistics. Text Mining (24 hours), coorganised with Guillaume Gravier. ENSAI.

- Benoît Sagot:
  - Master's course (M2) as part of the Master "Mathématiques, Vision Apprentissage". Speech and Language Processing (20 hours), coorganised with Emmanuel Dupoux and Paul Michel. ENS Paris-Saclay, France.
  - Certificate Data AI Product Owner Société Générale: on site session, PariSanté Campus. Introduction to NLP (7 hours). PariSanté Campus.5 Jul 2022.
  - Certificate Data AI Product Owner Société des Ingénieurs de l'Automobile, PariSanté Campus. Introduction to NLP (14 hours). PariSanté Campus.11 Jul 2022–13 Jul 2022.
  - Certificate Data AI Product Owner Société Générale: online session, Online. Introduction to NLP, part 1 (3.5 hours). Online.13 Oct 2022.
  - Certificate Data AI Product Owner Société Générale: online session, Online. Introduction to NLP, part 2 (3.5 hours). Online.15 Nov 2022.

- Floriane Chiffoleau:
  - 2-day training, URFIST de Bretagne et des Pays de la Loire, France. Introduction à la TEI (12 hours). URFIST de Bretagne et des Pays de la Loire, France.1 Nov 2022–1 Dec 2022.

- Djamé Seddah:
  - Tutorial for Advanced Language Processing winter school , Online. "CamemBert must die! (jk,lol. Beyond Sesame street-based naming schemes: Camembert vs CharacterBert, a study on the performance robustness of large monolingual language models and their character-based counterparts" (1 hour). Online.20 Jan 2022.

11.2.2 Supervision

PhD

- Axel Herold: "Extraction of etymological information from digital dictionaries" (1 Oct 2016–present). Primary affiliation: Berlin-Brandenburg Academy of Sciences. Supervised by Laurent Romary.
- José Carlos Rosales Núñez: "Machine translation for user-generated content" (1 Jun 2018–present). Primary affiliation: LISN, CNRS. Supervised by Guillaume Wisniewski and Djamé Seddah.
- Pedro Ortiz Suarez: "NLP and IE from 17th century encyclopedia" (1 Oct 2018–30 Apr 2022). Supervised by Laurent Romary and Benoît Sagot. PhD defended on 27 Jun 2022.
- Benjamin Muller: "NLP for social media texts" (1 Oct 2018–22 Jul 2022). Supervised by Benoît Sagot and Djamé Seddah. PhD defended on 17 Nov 2022.

- Clémentine Fourrier: “NLP for historical low-resource situations” (1 Oct 2019–30 Sep 2022). Supervised by Laurent Romary, Benoît Sagot and Rachel Bawden. PhD defended on 26 Sep 2022.
- Robin Algayres: “Unsupervised Automatic Speech Recognition in low resource conditions” (1 Oct 2019–present). Primary affiliation: CoML Inria project-team. Supervised by Emmanuel Dupoux and Benoît Sagot.
- Lionel Tadjou Tadonfouet: “Conversations Disentanglement” (1 Mar 2020–present). CIFRE PhD with Orange. Supervised by Laurent Romary, Éric de La Clergerie and Fabrice Bourge (CIFRE advisor).
- Tu Anh Nguyen: “Unsupervised acquisition of linguistic representations from speech (audio) data” (19 Apr 2021–present). CIFRE PhD with META AI Paris. Supervised by Benoît Sagot and Emmanuel Dupoux (CIFRE advisor).
- Paul-Ambroise Duquenne: “Study of vector spaces for sentence representation” (15 May 2021–present). CIFRE PhD with META AI Paris. Supervised by Benoît Sagot and Holger Schwenk (CIFRE advisor).
- Lydia Nishimwe: “Robust Neural Machine Translation” (1 Oct 2021–present). Supervised by Benoît Sagot and Rachel Bawden.
- Roman Castagné: “Neural language modelling” (1 Oct 2021–present). Supervised by Benoît Sagot and Éric de La Clergerie.
- 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.
- Floriane Chiffoleau: “Training data and creation of models for the text recognition of typewritten or handwritten corpus of archival collection” (15 Oct 2021–present). Primary affiliation: Université du Mans. Supervised by Anne Baillot and Laurent Romary.
- Matthieu Futeral-Peter: “Text-image multimodal models” (1 Nov 2021–present). Primary affiliation: WILLOW, Inria. Supervised by Ivan Laptev and Rachel Bawden.
- Alix Chagué: “Methodology for the creation of training data and the application of handwritten text recognition to the Humanities.” (1 Nov 2021–present). Secondary affiliation: Université de Montréal and CRIHN. Supervised by Laurent Romary, Emmanuel Château-Dutier and Michael Sinatra.
- Nathan Godey: “Neural language modelling” (1 Dec 2021–present). Supervised by Benoît Sagot and Éric de La Clergerie.
- Francis Kulumba: “Disambiguation of authors, institutions and bibliographic references in scientific publications” (1 Nov 2022–present). Supervised by Laurent Romary and Guillaume Vimont.
- Rian Touchent: “Information Extraction on French Electronic Health Records” (1 Dec 2022–present). Supervised by Laurent Romary and Éric de La Clergerie.
- Simon Meoni: (1 Dec 2022–present). CIFRE PhD with Arkhn. Supervised by Laurent Romary and Éric de La Clergerie.

Interns

- Camille Rey: “Neural Machine Translation” (20 Sep 2021–19 Mar 2022). Supervised by Rachel Bawden and Benoît Sagot.
- Kelly Christensen: “Linguistic annotation of digital facsimiles” (1 Apr 2022–31 Jul 2022). Supervised by Benoît Sagot and Simon Gabay.

- Jules Nuguet: “Data visualization experiment for Tei Publisher on the DataCatalogue project” (4 Apr 2022–31 Jul 2022). Primary affiliation: CESR. Supervised by Hugo Scheithauer and Laurent Romary.
- Abderraouf Farhi: “Automatic data extraction (DataCatalogue project)” (4 Apr 2022–31 Jul 2022).
- Camille Rey: “Neural Machine Translation” (1 May 2022–30 Jun 2022). Supervised by Rachel Bawden and Benoît Sagot.
- Rian Touchent: “Adaptation of camemBERT to the medical domain” (1 Jun 2022–8 Nov 2022). Primary affiliation: CentraleSupélec. Supervised by Éric de La Clergerie.
- Nacim Talaoubrid: “Annotation of named entities in the Narabizi Treebank” (1 Jun 2022–30 Jun 2022). Supervised by Djamé Seddah.
- Galo Castillo Lopez: “Improving hate speech detection in few shots scenarios” (6 Jun 2022–31 Aug 2022). Supervised by Djamé Seddah.
- Pierre Vauterin: “Integration of a speech recognition module in a First Person Shooter” (13 Jun 2022–5 Aug 2022). Supervised by Djamé Seddah.
- Rishika Bhagwatkar : “Exploring multimodal interactions ” (1 Jul 2022–31 Dec 2022). Supervised by Djamé Seddah.

Engineers

- Tanti Kristanti Nugraha: “Entity fishing for scholarly literature in the humanities” (1 Nov 2017–present).
- Yves Tadjo Takianpi: “Digital humanities” (1 Sep 2020–27 Jun 2022).
- Lucas Terriel: “Digital humanities, standardisation, named-entities processing in finding aids” (1 Nov 2020–31 Oct 2022).
- Thibault Charmet: “Automatic analysis of decisions by the Cour de Cassation (the French Supreme Court)” (1 Feb 2021–31 Jan 2022). Supervised by Rachel Bawden and Benoît Sagot.
- Julien Abadji: “Large-scale multilingual corpus development and extension (OSCAR corpus)” (1 Apr 2021–present). Supervised by Benoît Sagot.
- You Zuo: “Automatic patent classification” (1 Oct 2021–30 Nov 2022). Supervised by Benoît Sagot.
- Hugo Scheithauer: “Training segmentation models for sales catalogues with GROBID” (1 Oct 2021–present). Supervised by Laurent Romary.
- Rua Ismail: “Language identification for large-scale multilingual raw corpus development” (17 Jan 2022–present). Supervised by Benoît Sagot.
- Jesujoba Alabi: “Domain adaptation for neural machine translation” (1 Feb 2022–14 Jun 2022). Supervised by Rachel Bawden.
- Wissam Antoun: “Language models for languages displaying high variabilty, in particular Arabic dialects used on social media” (1 Mar 2022–present). Supervised by Djamé Seddah and Benoît Sagot.
- Anna Chepaikina: “Automatic generation of oenological descriptions” (31 Mar 2022–present). Supervised by Benoît Sagot.
- Menel Mahamdi: “Automatic extraction and annotation of information regarding the ecological impact of projects handled by the French Ministry for the Ecological Transition” (1 Sep 2022–present). Supervised by Éric de La Clergerie.

- Niyati Bafna: “Linguistically inspired language models for closely related languages” (1 Oct 2022–present). Secondary affiliation: DFKI. Supervised by Benoît Sagot, Rachel Bawden, Josef van Genabith and Cristina España-Bonet.
- Mouilleron Virginie: “Correction and Annotations of the Alien vs Predator data set” (1 Dec 2022–present). Supervised by Djamé Seddah.

Postdocs

- Syrielle Montariol: “Word usage change for emerging communities and radicalisation detection in social media” (1 Apr 2021–31 May 2022). Supervised by Djamé Seddah.

11.2.3 Juries

PhD

- Rachel Bawden
  - Member of the PhD committee as examiner for Antoine Simoulin at Université Paris Cité on 7 Jul 2022. Title: *Sentence embeddings and their relation with sentence structures.*
  - Member of the PhD committee as examiner for Baptiste Rozière at Université Paris Dauphine-PSL on 12 Jul 2022. Title: *Traduction Non Supervisée de Langages de Programmation.*
  - Member of the PhD committee as examiner for Jitao Xu at Université Paris-Saclay on 2 Dec 2022. Title: *Writing in two languages: neural machine translation as an assistive bilingual writing tool.*
  - Member of the PhD committee as co-supervisor for Clémentine Fourrier at Inria on 26 Sep 2022. Title: *Neural Approaches to Historical Word Reconstruction.*
- Benoît Sagot
  - Member of the PhD committee as co-director for Clémentine Fourrier at Inria on 26 Sep 2022. Title: *Neural Approaches to Historical Word Reconstruction.*
  - Member of the PhD committee as president for Léo Laugier at Télécom Paris (IPP) on 8 Nov 2022. Title: *Analysis and Control of Online Interactions through Neural Natural Language Processing.*
  - Member of the PhD committee as director for Benjamin Muller at Inria on 17 Nov 2022. Title: *How can we make language models better at handling the diversity and variability of natural languages?.*
  - Member of the PhD committee as co-director for Pedro Ortiz Suarez at Inria on 27 Jun 2022. Title: *A Data-driven Approach to Natural Language Processing for Contemporary and Historical French.*
- Djamé Seddah
  - Member of the PhD committee as co-supervisor for Benjamin Muller at Inria on 17 Nov 2022. Title: *How can we make language models better at handling the diversity and variability of natural languages?.*
  - Member of the PhD committee as examiner for Graziella De Martino at Université de Bari (Italie) on 22 Nov 2022. Title: *Machine Learning and Process Discovery to Support Legal Transcript Writing* .
- Éric de La Clergerie
  - Member of the PhD committee as co-supervisor for Mathilde Regnault at Université Sorbonne Nouvelle, Paris, France on 16 Jun 2022. Title: *Annotation et analyse syntaxiques de corpus hétérogènes.*

  - Member of the PhD committee as examiner for Laura Noreskal at Université Paris Nanterre on 14 Dec 2022. Title: *Détection automatique de constructions erronées: structures coordonnées dans les rédactions des étudiants*.

- Laurent Romary
  - Member of the PhD committee as director for Pedro Ortiz Suarez at Inria on 27 Jun 2022. Title: *A Data-driven Approach to Natural Language Processing for Contemporary and Historical French*.
  - Member of the PhD committee as director for Clémentine Fourrier at Inria on 26 Sep 2022. Title: *Neural Approaches to Historical Word Reconstruction*.

Master

- Rachel Bawden
  - Member of the Master's committee as co-supervisor for Camille Rey at INALCO on 12 Jul 2022. Title: *Améliorer la désambiguïsation lexicale en traduction automatique neuronale*.
- Benoît Sagot
  - Member of the Master's committee as co-director for Camille Rey at INALCO on 12 Jul 2022. Title: *Améliorer la désambiguïsation lexicale en traduction automatique neuronale*.
  - Member of the Master's committee as examiner for Quang Anh Nguyen at MVA & École polytechnique on 1 Sep 2022. Title: *Few shots learning for text classification: Topic categorization of news data for credit surveillance*.
  - Member of the Master's committee as examiner for Angelo Ortiz at MVA & Télécom ParisTech and onepoint on 9 Sep 2022. Title: *Word-Sense Disambiguation by Graph Analysis*.
  - Member of the Master's committee for Clarine Vongpaseut at MVA & Sinequa on 26 Sep 2022. Title: *Contextualized topic modelling*.
  - Member of the Master's committee for Félix Lefebvre at MVA & Inria Saclay on 1 Nov 2022. Title: *Large-scale embedding of heterogeneous information*.
- Alix Chagué
  - Member of the Master's committee as co-supervisor for Jessica Benammar at École du Louvre, Paris, France on 16 Sep 2022. Title: *Les spécificités de la mise en place d'une documentation pour une collection privée. Bilan et perspectives de la documentation de la collection de Christian Giacomotto*.
  - Member of the Master's committee as co-supervisor for Margaux Granier at École du Louvre, Paris, France on 15 Sep 2022. Title: *Le film institutionnel, une typologie particulière d'archive audiovisuelle: enjeux de documentation, pratiques et valorisation d'un outil interne au service de la mémoire et de la communication. Le cas particulier de la base de données "Audiovisuel" de la maison Hermès*.
  - Member of the Master's committee as co-supervisor for Agathe Souleau at École du Louvre, Paris, France on 20 Sep 2022. Title: *Intégration d'une collection scientifique au musée des Arts et Métiers: enjeux de la gestion physique d'un fonds de l'IGN, de son inventaire à sa valorisation*.
  - Member of the Master's committee as co-supervisor for Anne-Claire Durand at École du Louvre, Paris, France on 20 Sep 2022. Title: *Initier à la mise en place d'un système de gestion informatisé des biens culturels pour le Ministère de l'Eduction nationale, de la Jeunesse et des Sports, et le Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation*.
  - Member of the Master's committee as co-supervisor for Julien Christol at École du Louvre, Paris, France on 21 Sep 2022. Title: *Vers une définition et une (re)valorisation de la documentation sur des collections photographiques relatives aux missions chrétiennes. Le cas du musée du quai Branly-Jacques Chirac*.

  - Member of the Master's committee as co-supervisor for Lisa Gianoni at École du Louvre, Paris, France on 5 Oct 2022. Title: *Définir la stratégie documentaire d'un fonds photographique en quête de patrimonialisation. La mise en place de bonnes pratiques pour le traitement documentaire des archives de la Brigade des Sapeurs-Pompiers de Paris.*
  - Member of the Master's committee as co-supervisor for Erine Walendowski at École du Louvre, Paris, France on 4 Oct 2022. Title: *L'évolution des pratiques de mise en ligne des collections: le cas de la Collection du ministre du Service historique de la Défense.*

- Hugo Scheithauer

  - Member of the Master's committee as examiner for Quentin Bernet at École du Louvre, Paris, France on 29 Sep 2022. Title: *Iconographie numérique. Quelles nouvelles perspectives la science de la donnée offre-t-elle à l'Histoire de l'Art ?.*
  - Member of the Master's committee as examiner for Margaux Coïc at École du Louvre, Paris, France on 5 Oct 2022. Title: *La documentation des expositions, du projet d'exposition à sa mise à disposition auprès des publics. Cas d'étude des pratiques documentaires de musées français et internationaux..*

CSD

- Rachel Bawden

  - Member of the CSD committee for Antoine Yang at Inria on 15 Jun 2022. Title: *Multimodal video representation with cross-modal learning.*

- Benoît Sagot

  - Member of the CSD committee for Nathanaël Beau at Université Paris Cité & onespace on 1 Sep 2022. Title: *Génération de code métier Python à partir d'une description en langage naturel.*
  - Member of the CSD committee for Marine Courtin at Université Sorbonne Nouvelle, Paris, France on 24 Oct 2022. Title: *Emergence empirique de structures syntaxiques: interactions entre la détection de frontières d'énoncés et la structuration syntaxique.*
  - Member of the CSD committee for Yixuan Li at Université Sorbonne Nouvelle, Paris, France on 25 Oct 2022. Title: *Tech-mining on Chinese Patents.*
  - Member of the CSD committee for Marc Benzahra at Université Paris Saclay on 6 Jan 2022. Title: *Vers une évaluation universelle du niveau de complexité des textes.*

- Éric de La Clergerie

  - Member of the CSD committee for Mishra Shrey at Université Paris sciences et lettres on 7 Oct 2022. Title: *Vers une base de connaissances de résultats mathématiques.*
  - Member of the CSD committee for Maya Sahraoui at Sorbonne Université, Paris, France on 31 Mar 2022. Title: *Enrichissement joint, bases de connaissances - textes - images, par machine learning dans le contexte de l'identification en biodiversité.*

Hiring committees

- Rachel Bawden:

  - Member of the Commission des emplois scientifiques (CES) hiring committee at Inria (Paris Centre). Delegations, postdocs and PhDs.
  - External Member of the MC hiring committee at Télécom Paris (Data, Intelligence and Graphs (DIG)).

- Benoît Sagot:

- Member of the AER hiring committee at Inria (Saclay Centre).
- Member of the AER hiring committee at Inria (Lyon Centre).
- Member of the AER hiring committee at Inria (Grenoble Centre).
- Member of the CR-ISFP hiring committee at Inria (Nancy Centre).
- Member of the CR-ISFP hiring committee at Inria (Paris Centre).

11.3 Popularization

11.3.1 Articles and contents

Authored articles

- Floriane Chiffoleau authored an article for the Digital Intellectuals Blog (Outreach article), "Recognizing and encoding the corpus' named entities". Online, 1 Mar 2022.
- Alix Chagué authored an article for LECTAUREP Blog (Outreach article), "Publication d'un modèle de transcription et de la vérité de terrain !" Online, 12 May 2022.

Articles with citation

- Rachel Bawden was cited in an article by larecherche.fr (Media article), "Bloom : un nouveau modèle de génération automatique de textes". Online, 19 Aug 2022.
- Benoît Sagot was cited in an article by Le Monde (supplément Sciences), "Au-delà de l'intelligence artificielle, le chatbot ChatGPT doit ses talents d'orateur à l'humain". Print + Online, 21 Dec 2022.

Media interviews

- Djamé Seddah was interviewed as part of Radio France (7 mn interview at France Culture's La Methode Scienfique), "le reportage du jour". Radio broadcast + Online, 5 Jan 2022.

11.3.2 Education

- Tu Anh Nguyen was a speaker at MASSP - Math and Science Summer Program (Summer School for High school students on Data Science), "Introductory courses on "Unsupervised Learning" & "Deep Learning"". Online, 7 Aug 2022 (6 hours).
- Rachel Bawden was a speaker at the Rendez-Vous des Jeunes Mathématiciennes et Informaticiennes Inria 2022 (Presentation on NLP to high school girls), "Traitement Automatique des Langues (TAL)". Inria Paris, 1 Oct 2022 (1 hour).

11.3.3 Interventions

- Éric de La Clergerie, with IRCAM, Sorbonne Université, gave a talk at Deep Voice Paris (Professional workshop), "Le biais est dans le texte?". Paris, 17 Jun 2022 (1.5 hours).
- Djamé Seddah, with IRCAM, Sorbonne Université, gave a talk at Deep Voice Paris (Professional workshop), "Le biais est dans le texte?". Paris, 17 Jun 2022 (1.5 hours).
- Éric de La Clergerie gave a talk at Healthcare NLP meetup (Professional workshop), "Addressing the processing of medical textual documents". Paris, 26 Oct 2022 (1 hour).
- Alix Chagué, with AI4LAM, gave a talk at AI4LAM Community Call (Professional workshop), "HTR-United: Commons for Automatic Text Recognition?". Online, 15 Nov 2022 (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. Algayres, T. Ricoul, J. Karadayi, H. Laurençon, S. Zaiem, A. Mohamed, B. Sagot and E. Dupoux. 'DP-Parse: Finding Word Boundaries from Raw Speech with an Instance Lexicon'. In: *Transactions of the Association for Computational Linguistics* 10 (19th Sept. 2022), pp. 1051–1065. DOI: 10.1162/tacl_a_00505. URL: https://hal.inria.fr/hal-03831873.

[15] B. Haddow, R. Bawden, A. V. Miceli Barone, J. Helcl and A. Birch. 'Survey of Low-Resource Machine Translation'. In: *Computational Linguistics* 48.3 (2022), pp. 673–732. URL: https://hal.inria.fr/hal-03479757.

[16] J. Kreutzer, I. Caswell, 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'. In: *Transactions of the Association for Computational Linguistics* 10 (31st Jan. 2022), pp. 50–72. DOI: 10.1162/tacl_a_00447. URL: https://hal.inria.fr/hal-03177623.

[17] T. A. Nguyen, B. Sagot and E. Dupoux. 'Are discrete units necessary for Spoken Language Modeling?' In: *IEEE Journal of Selected Topics in Signal Processing* (23rd Aug. 2022). URL: https://hal.inria.fr/hal-03831707.

National journals

[18] A. Chagué and L. Romary. 'Artificial Intelligence, openiing a field of possibilities'. In: *Arabesques* 107 (2nd Sept. 2022), pp. 4–5. DOI: 10.35562/arabesques.3043. URL: https://hal.inria.fr/hal-04030241.

International peer-reviewed conferences

[19] J. Abadji, P. Ortiz Suarez, L. Romary and B. Sagot. 'Towards a Cleaner Document-Oriented Multilingual Crawled Corpus'. In: Thirteenth Language Resources and Evaluation Conference - LREC 2022. Proceedings of the Thirteenth Language Resources and Evaluation Conference. Marseille, France, 20th June 2022. URL: https://hal.inria.fr/hal-03536361.

[20] J. O. Alabi, L. Nishimwe, B. Muller, C. Rey, B. Sagot and R. Bawden. 'Inria-ALMAnaCH at the WMT 2022 shared task: Does Transcription Help Cross-Script Machine Translation?' In: *Proceedings of the Seventh Conference on Machine Translation*. EMNLP 2022 - Seventh Conference on Machine Translation (WMT22 - Workshop on Statistical Machine Translation). Abu Dhabi, United Arab Emirates, 2022. URL: https://hal.inria.fr/hal-03836180.

[21] R. Algayres, A. Nabli, B. Sagot and E. Dupoux. 'Speech Sequence Embeddings using Nearest Neighbors Contrastive Learning'. In: Interspeech 2022 - 23rd INTERSPEECH Conference. Incheon, South Korea, 18th Sept. 2022. URL: https://hal.inria.fr/hal-03831888.

[22] R. Bawden, J. Poinhos, E. Kogkitsidou, P. Gambette, B. Sagot and S. Gabay. 'Automatic Normalisation of Early Modern French'. In: LREC 2022 - 13th Language Resources and Evaluation Conference. Marseille, France, 20th June 2022, pp. 3354–3366. DOI: 10.5281/zenodo.5865428. URL: https://hal.inria.fr/hal-03540226.

[23] T. Charmet, I. Cherichi, M. Allain, U. Czerwinska, A. Fouret, B. Sagot and R. Bawden. 'Complex Labelling and Similarity Prediction in Legal Texts: Automatic Analysis of France's Court of Cassation Rulings'. In: LREC 2022 - 13th Language Resources and Evaluation Conference. Marseille, France, 20th June 2022. URL: https://hal.inria.fr/hal-03663110.

[24] A. Chepaikina, R. Bossy, C. Roussey and S. Bernard. 'Thesaurus Enrichment via Coordination Extraction'. In: 16th International Conference on Metadata and Semantics Research (MTSR 2022). London, United Kingdom, 7th Nov. 2022. URL: https://hal.inria.fr/hal-03933526.

[25] P.-A. Duquenne, H. Gong, B. Sagot and H. Schwenk. 'T-Modules: Translation Modules for Zero-Shot Cross-Modal Machine Translation'. In: EMNLP 2022 - 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, United Arab Emirates, 7th Dec. 2022. URL: https://hal.inria.fr/hal-03834732.

[26] G. Felhi, J. Le Roux and D. Seddah. 'Exploiting Inductive Bias in Transformers for Unsupervised Disentanglement of Syntax and Semantics with VAEs'. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. NAACL 2022 - Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle, United States: Association for Computational Linguistics, 10th July 2022, pp. 5763–5776. DOI: 10.18653/v1/2022.naacl-main.423. URL: https://hal.science/hal-03812847.

[27] C. Fourrier and S. Montariol. 'Caveats of Measuring Semantic Change of Cognates and Borrowings using Multilingual Word Embeddings'. In: LChange'22 - 3rd International Workshop on Computational Approaches to Historical Language Change 2022. Dublin, Ireland, 26th May 2022. URL: https://hal.inria.fr/hal-03635005.

[28] C. Fourrier and B. Sagot. 'Probing Multilingual Cognate Prediction Models'. In: ACL 2022 - Findings of the Association for Computational Linguistics. Dublin, Ireland, 22nd May 2022. URL: https://hal.inria.fr/hal-03614691.

[29] S. Gabay, R. Bawden, P. Gambette, J. Poinhos, E. Kogkitsidou and B. Sagot. 'Le changement linguistique au XVIIe s. : nouvelles approches scriptométriques'. In: *Actes du 8e Congrès Mondial de Linguistique Française*. CMLF 2022 - 8e Congrès Mondial de Linguistique Française. Vol. 138. SHS Web of conferences. Orléans, France: EDP Sciences, 2022, pp. 02006.1–14. DOI: 10.1051/shsconf/202213802006. URL: https://hal.science/hal-03681556.

[30] S. Gabay, P. Ortiz Suarez, A. Bartz, A. Chagué, R. Bawden, P. Gambette and B. Sagot. 'From FreEM to D'AlemBERT: a Large Corpus and a Language Model for Early Modern French'. In: 13th Language Resources and Evaluation Conference - LREC 2022. Proceedings of the 13th Language Resources and Evaluation Conference. Marseille, France, 20th June 2022, pp. 3367–3374. URL: https://hal.inria.fr/hal-03596653.

[31] N. Godey, R. Castagné, E. Villemonte de La Clergerie and B. Sagot. 'MANTa: Efficient Gradient-Based Tokenization for Robust End-to-End Language Modeling'. In: EMNLP 2022 - The 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, United Arab Emirates, Dec. 2022. URL: https://hal.science/hal-03844262.

[32] O. Goldman, F. Tinner, H. Gonen, B. Muller, V. Basmov, S. Kirimi, L. Nishimwe, B. Sagot, D. Seddah, R. Tsarfaty and D. Ataman. 'The MRL 2022 Shared Task on Multilingual Clause-level Morphology'. In: 1st Shared Task on Multilingual Clause-level Morphology. Abu Dhabi, United Arab Emirates, 8th Dec. 2022. URL: https://hal.inria.fr/hal-03878174.

[33] T. Kocmi, R. Bawden, O. Bojar, A. Dvorkovich, C. Federmann, M. Fishel, T. Gowda, Y. Graham, R. Grundkiewicz, B. Haddow, R. Knowles, P. Koehn, C. Monz, M. Morishita, M. Nagata, T. Nakazawa, M. Novák, M. Popel, M. Popović and M. Shmatova. 'Findings of the 2022 Conference on Machine Translation (WMT22)'. In: *Proceedings of the Seventh Conference on Machine Translation (WMT)*. WMT 2022 - Seventh Conference on Machine Translation. Abu Dhabi, United Arab Emirates, 2022. URL: https://hal.inria.fr/hal-03932367.

[34] L. Martin, A. Fan, É. Villemonte de la Clergerie, A. Bordes and B. Sagot. 'MUSS: Multilingual Unsupervised Sentence Simplification by Mining Paraphrases'. In: LREC 2022 - 13th Language Resources and Evaluation Conference. Marseille, France, 20th June 2022. URL: https://hal.inria.fr/hal-03834719.

[35] S. Montariol, A. Riabi and D. Seddah. 'Multilingual Auxiliary Tasks Training: Bridging the Gap between Languages for Zero-Shot Transfer of Hate Speech Detection Models'. In: *Findings of AACL 2022*. AACL-IJCNLP 2022 - 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing. Online, France, 20th Nov. 2022. URL: https://hal.inria.fr/hal-03840070.

[36] S. Montariol, É. Simon, A. Riabi and D. Seddah. 'Fine-tuning and Sampling Strategies for Multimodal Role Labeling of Entities under Class Imbalance'. In: *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*. CONSTRAINT 2022 - Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations. Dublin, Ireland: Association for Computational Linguistics, 27th May 2022, pp. 55–65. DOI: 10.18653/v1/2022.constraint-1.7. URL: https://hal.inria.fr/hal-03840060.

[37] M. Neves, A. Jimeno Yepes, A. Siu, R. Roller, P. Thomas, M. Vicente Navarro, L. Yeganova, D. Wiemann, G. M. Di Nunzio, F. Vezzani, C. Gérardin, R. Bawden, D. Johan Estrada, S. Lima-López, E. Farré-Maduell, M. Krallinger, C. Grozea and A. Névéol. 'Findings of the WMT 2022 Biomedical Translation Shared Task: Monolingual Clinical Case Reports'. In: *Proceedings of the Seventh Conference on Machine Translation*. WMT22 - Seventh Conference on Machine Translation. Abu Dhabi, United Arab Emirates, 2022, pp. 694–723. URL: https://hal.inria.fr/hal-03932275.

[38] T. A. Nguyen, E. Kharitonov, J. Copet, Y. Adi, W.-N. Hsu, A. Elkahky, P. Tomasello, R. Algayres, B. Sagot, A. Mohamed and E. Dupoux. 'Generative Spoken Dialogue Language Modeling'. In: SLT-2022 - IEEE Spoken Language Technology Workshop. Doha-Qatar, Qatar, 13th Feb. 2023. URL: https://hal.inria.fr/hal-03985368.

[39] 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.

[40] 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, 2022. URL: https://hal.inria.fr/hal-03540072.

[41] F. de Toni, C. Akiki, J. de La Rosa, C. Fourrier, E. Manjavacas, S. Schweter and D. van Strien. 'Entities, Dates, and Languages: Zero-Shot on Historical Texts with T0'. In: BigScience 2022 - International Workshop on Challenges & Perspectives in Creating Large Language Models 2022. Dublin, Ireland, 27th May 2022. URL: https://hal.inria.fr/hal-03639144.

National peer-reviewed Conferences

[42] S. Gabay, P. Ortiz Suarez, R. Bawden, A. Bartz, P. Gambette and B. Sagot. 'Le projet FREEM : ressources, outils et enjeux pour l'étude du français d'Ancien Régime'. In: *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*. TALN 2022 - Traitement Automatique des Langues Naturelles. Avignon, France: ATALA, 2022, pp. 154–165. URL: https://hal.science/hal-03701524.

[43] B. Muller, A. Anastasopoulos, B. Sagot and D. Seddah. 'When Being Unseen from mBERT is just the Beginning : Handling New Languages With Multilingual Language Models'. In: *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*. TALN 2022 - 29° conférence sur le Traitement Automatique des Langues Naturelles. Avignon, France: ATALA, 2022, pp. 450–451. URL: https://hal.science/hal-03701503.

[44] A. Riabi, S. Montariol and D. Seddah. 'Tâches Auxiliaires Multilingues pour le Transfert de Modèles de Détection de Discours Haineux'. In: *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*. Traitement Automatique des Langues Naturelles. Avignon, France: ATALA, 2022, pp. 413–423. URL: https://hal.archives-ouvertes.fr/hal-03701522.

Conferences without proceedings

[45] A. Chagué and T. Clérice. 'Sharing HTR datasets with standardized metadata: the HTR-United initiative'. In: Documents anciens et reconnaissance automatique des écritures manuscrites. Paris, France, 23rd June 2022. URL: https://hal.inria.fr/hal-03703989.

[46] A. Chagué, H. Scheithauer, L. Terriel, F. Chiffoleau and Y. Tadjo-Takianpi. 'Take a sip of TEI and relax: a proposition for an end-to-end workflow to enrich and publish data created with automatic text recognition'. In: Digital Humanities 2022 : Responding to Asian Diversity. Tokyo, Japan, 25th July 2022. URL: https://hal.inria.fr/hal-03739767.

[47] F. Chiffoleau and H. Scheithauer. 'From a collection of documents to a published edition : how to use an end-to-end publication pipeline'. In: TEI 2022 - Text Encoding Initiative 2022 Conference. Newcastle, United Kingdom, 12th Sept. 2022. URL: https://hal.science/hal-03780316.

[48] F. Clavaud, L. Romary, P. Charbonnier, L. Terriel, G. Piraino and V. Verdese. 'NER4Archives (named entity recognition for archives) : Conception et réalisation d'un outil de détection, de classification et de résolution des entités nommées dans les instruments de recherche archivistiques encodés en XML/EAD.' In: Atelier Culture-INRIA. Pierrefitte sur Seine, France, 22nd Mar. 2022. URL: https://hal.science/hal-03625734.

[49] 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.

[50] S. Gabay, R. Bawden, B. Sagot and P. Gambette. 'Vers l'étude linguistique sur données artificielles: Le cas des systèmes graphiques en diachronie longue'. In: Variation(s) en français. Nancy, France, 17th Nov. 2022. URL: https://hal-univ-eiffel.archives-ouvertes.fr/hal-03856660.

[51] L. Grobol, M. Regnault, P. Ortiz Suarez, B. Sagot, L. Romary and B. Crabbé. 'BERTrade: Using Contextual Embeddings to Parse Old French'. In: *Proceedings of the 13th Language Resources and Evaluation Conference*. 13th Language Resources and Evaluation Conference. Marseille, France, 21st June 2022. URL: https://hal.science/hal-03736840.

[52] E. H. Karim, W. Antoun, F. Le Ber and V. Pitchon. 'Reconnaissance des entités nommées pour l'analyse des pharmacopées médiévales'. In: EGC 2023 - Extraction et Gestion des Connaissances. Lyon, France, 16th Jan. 2023. URL: https://hal.science/hal-03934557.

[53] A. Pinche, K. Christensen and S. Gabay. 'Between automatic and manual encoding: Towards a generic TEI model for historical prints and manuscripts'. In: TEI 2022 conference : Text as data. Newcastle, United Kingdom, 13th Sept. 2022. DOI: 10.5281/zenodo.7092214. URL: https://hal.science/hal-03780302.

[54] L. Romary and H. Scheithauer. 'DataCatalogue : enjeux et réalisations'. In: Un outil numérique pour interroger les catalogues de vente : le projet DataCatalogue. Paris, France, 21st Oct. 2022. URL: https://hal.archives-ouvertes.fr/hal-03829309.

[55] A. Rostaing and H. Scheithauer. 'Enrichir le patrimoine écrit archivistique grâce aux technologies numériques : Ingénierie du projet LectAuRep (Lecture automatique de répertoires)'. In: DHNord 2022 - Travailler en Humanités Numériques : collaborations, complémentarités et tensions. Online, France, 20th June 2022. URL: https://hal.archives-ouvertes.fr/hal-03792952.

[56] A. Rostaing and H. Scheithauer. 'LectAuRep : Un projet de recherche et développement pour la transcription automatique de répertoires de notaires'. In: La reconnaissance des écritures manuscrites et ses usages dans les archives. Pierrefitte-sur-Seine, France, 29th Nov. 2022. URL: https://hal.archives-ouvertes.fr/hal-03894910.

[57] A. Rostaing and H. Scheithauer. 'LectAuRep (2018-2021) :Projet de lecture automatique de répertoires de notaires'. In: Segmenter et annoter les images : déconstruire pour reconstruire. Paris, France, 15th Nov. 2022. URL: https://hal.archives-ouvertes.fr/hal-03855439.

[58] Y. Rychener, X. Renard, D. Seddah, P. Frossard and M. Detyniecki. 'On the Granularity of Explanations in Model Agnostic NLP Interpretability'. In: XKDD 2022 - ECML PKDD 2022 International Workshop on eXplainable Knowledge Discovery in Data Mining. Grenoble, France, 19th Sept. 2022. URL: https://hal.science/hal-03936558.

[59] B. Sagot, L. Romary, R. Bawden, P. J. Ortiz Suárez, K. Christensen, S. Gabay, A. Pinche and J.-B. Camps. 'Gallic(orpor)a : Extraction, annotation et diffusion de l'information textuelle et visuelle en diachronie longue: Restitution des travaux'. In: DataLab de la BnF : Restitution des travaux 2022. Paris, France, 9th Dec. 2022. URL: https://hal.science/hal-03930542.

[60] H. Scheithauer. 'LectAuRep : Données d'archives en français des XIXe et XXe siècles'. In: Transkribus / eScriptorium : Transcrire, annoter et éditer numériquement des documents d'archives. Paris, France, 9th May 2022. URL: https://hal.inria.fr/hal-03666884.

[61] H. Scheithauer, L. Romary, F. Duyrat and F. Nurra. 'Datacatalogue: project presentation'. In: Atelier Culture-Inria. Pierrefitte-sur-Seine, France, 22nd Mar. 2022. URL: https://hal.inria.fr/hal-03618381.

[62] Y. Zuo, Y. Li, A. Parias García and K. Gerdes. 'Technological taxonomies for hypernym and hyponym retrieval in patent texts'. In: ToTh 2022 - Terminology & Ontology: Theories and applications. Chambéry, France, 2022. URL: https://hal.science/hal-03850399.

[63] Y. Zuo, H. Mouzoun, S. Ghamri Doudane, K. Gerdes and B. Sagot. 'Patent Classification using Extreme Multi-label Learning: A Case Study of French Patents'. In: SIGIR 2022 - PatentSemTech workshop - 3rd Workshop on Patent Text Mining and Semantic Technologies. Madrid, Spain, 2022. URL: https://hal.science/hal-03850405.

Scientific book chapters

[64] J. Bowers. 'Pathways and patterns of metaphor and metonymy in Mixtepec-Mixtec body-part terms'. In: *The Grammar of Body-Part Expressions: A view from the Americas*. Roberto Zariquiey; Pilar M. Valenzuela, 2022, pp. 91–135. DOI: 10.1093/oso/9780198852476.003.0004. URL: https://hal.inria.fr/hal-02075731.

[65] A. Chagué, V. Le Fourner, M. Martini and E. Villemonte de La Clergerie. 'Deux siècles de sources disparates sur l'industrie textile en France : comment automatiser les traitements d'un corpus non-uniforme ?' In: *La fabrique numérique des corpus en sciences humaines et sociales*. Presses Universitaires du Septentrion; https://www.septentrion.com/livre/?GCOI=27574100990460, 16th Dec. 2022. URL: https://shs.hal.science/halshs-03973309.

[66] V. Le Fourner, A. Chagué, M. Martini and A. Albert. 'Structurer automatiquement un corpus homogène issu de la reconnaissance d'écriture manuscrite : les jugements du Conseil des prud'hommes des tissus parisiens'. In: *La fabrique numérique des corpus en sciences humaines et sociales*. Presses Universitaires du Septentrion, 16th Dec. 2022, https://www.septentrion.com/livre/?GCOI=27574100990460. URL: https://shs.hal.science/halshs-03969037.

Doctoral dissertations and habilitation theses

[67] C. Fourrier. 'Neural Approaches to Historical Word Reconstruction'. Université PSL (Paris Sciences & Lettres), 26th Sept. 2022. URL: https://hal.inria.fr/tel-03793299.

[68] B. Muller. 'How Can We Make Language Models Better at Handling the Diversity and Variability of Natural Languages ?' Sorbonne Université, 17th Nov. 2022. URL: https://theses.hal.science/tel-03966952.

[69] P. Ortiz Suarez. 'A Data-driven Approach to Natural Language Processing for Contemporary and Historical French'. Sorbonne Université, 27th June 2022. URL: https://theses.hal.science/tel-03770337.

Reports & preprints

[70] W. Antoun, B. Sagot and D. Seddah. *Data-Efficient French Language Modeling with CamemBERTa*. 20th Jan. 2023. URL: https://hal.inria.fr/hal-03963729.

[71] R. Bawden and F. Yvon. *Investigating the Translation Performance of a Large Multilingual Language Model: the Case of BLOOM*. 3rd Mar. 2023. DOI: 10.48550/ARXIV.2303.01911. URL: https://hal.inria.fr/hal-04015863.

[72] F. Chiffoleau and A. Baillot. *Le projet DAHN : une pipeline pour l'édition numérique de documents d'archives*. 1st Apr. 2022. URL: https://hal.science/hal-03628094.

[73] T. Clérice, M. Vlachou-Efstathiou and A. Chagué. *CREMMA Medii Aevi: Literary manuscript text recognition in Latin*. 11th Jan. 2023. URL: https://hal-enc.archives-ouvertes.fr/hal-03828353.

[74] M. Futeral, C. Schmid, I. Laptev, B. Sagot and R. Bawden. *Tackling Ambiguity with Images: Improved Multimodal Machine Translation and Contrastive Evaluation*. 20th Dec. 2022. URL: https://hal.inria.fr/hal-03977982.

[75] Y. L. Liu, R. Bawden, T. Scialom, B. Sagot and J. C. K. Cheung. *MaskEval: Weighted MLM-Based Evaluation for Text Summarization and Simplification*. 30th Oct. 2022. URL: https://hal.inria.fr/hal-03834733.

[76] A. Mcmillan-Major, Z. Alyafeai, S. Biderman, K. Chen, F. de Toni, G. Dupont, H. Elsahar, C. Emezue, A. F. Aji, S. Ilić, N. Khamis, C. Leong, M. Masoud, A. Soroa, P. Ortiz Suarez, Z. Talat, D. van Strien and Y. Jernite. *Documenting Geographically and Contextually Diverse Data Sources: The BigScience Catalogue of Language Data and Resources*. 1st Feb. 2022. URL: https://hal.inria.fr/hal-03550289.

[77] 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.

[78] T. A. Nguyen, E. Kharitonov, J. Copet, Y. Adi, W.-N. Hsu, A. Elkahky, P. Tomasello, R. Algayres, B. Sagot, A. Mohamed and E. Dupoux. *Generative Spoken Dialogue Language Modeling: preprint version*. 30th Oct. 2022. URL: https://hal.inria.fr/hal-03834730.

[79] T. A. Nguyen, M. D. Seyssel, R. Algayres, P. Rozé, E. Dunbar and E. Dupoux. *Are word boundaries useful for unsupervised language learning?* 2022. DOI: 10.48550/ARXIV.2210.02956. URL: https://hal-cnrs.archives-ouvertes.fr/hal-03992291.

[80] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé et al. *BLOOM: A 176B-Parameter Open-Access Multilingual Language Model*. 12th Nov. 2022. URL: https://hal.inria.fr/hal-03850124.

[81] H. Scheithauer, A. Chagué and L. Romary. *Which TEI representation for the output of automatic transcriptions and their metadata? An illustrated proposition*. 27th Apr. 2022. URL: https://hal.inria.fr/hal-04001303.

Other scientific publications

[82] A. Chagué. 'Conditions de la mutualisation : les principes FAIR et HTR-United'. In: Humanistica 2022. Montréal, Canada, 19th May 2022. URL: https://hal.inria.fr/hal-03685731.

[83] A. Chagué. *Intelligence Artificielle et intelligence collective : des nouveaux eldorados pour rendre les textes patrimoniaux plus accessibles ?* 24th May 2022. URL: https://hal.archives-ouvertes.fr/hal-03739948.

[84] C. L. Jacobs, A. De Santo and L. Grobol. 'Online and offline processing in zeugma constructions is insensitive to argument order'. In: Human Sentence Processing 2023. Pittsburg, United States, 9th Mar. 2023. URL: https://hal.science/hal-04030325.

[85] B. Sagot, L. Romary, R. Bawden, P. Ortiz Suarez, K. Christensen, S. Gabay, A. Pinche and J.-B. Camps. *Gallic(orpor)a: Extraction, annotation et diffusion de l'information textuelle et visuelle en diachronie longue: Restitution des travaux*. 9th Dec. 2022. URL: https://hal.science/hal-04024750.

12.3 Other

Scientific popularization

[86] A. Chagué. 'Corpus, méthodes et ressources pour la transcription automatique des documents manuscrits patrimoniaux francophones contemporains'. In: 89e Congrès de l'Acfas, Section 310 - Le numérique dans les sciences humaines : édition et visualisation. Montréal, Canada, 9th May 2022. URL: https://hal.inria.fr/hal-03664788.

[87] A. Chagué. 'eScriptorium: a free application to automatically transcribe manuscripts'. In: *Arabesques* 107 (2nd Sept. 2022), p. 25. DOI: 10.35562/arabesques.3100. URL: https://hal.inria.fr/hal-04030514.

12.4 Cited publications

[88] 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, July 2021. DOI: 10.14618/ids-pub-10468. URL: https://hal.inria.fr/hal-03301590.

[89] 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.

[90] M. Artetxe and H. Schwenk. 'Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond'. In: *Transactions of the Association for Computational Linguistics* 7 (2019), pp. 597–610. DOI: 10.1162/tacl_a_00288. URL: https://aclanthology.org/Q19-1038.

[91] M. Bañón, P. Chen, B. Haddow, K. Heafield, H. Hoang, M. Esplà-Gomis, M. L. Forcada, A. Kamran, F. Kirefu, P. Koehn, S. Ortiz Rojas, L. Pla Sempere, G. Ramírez-Sánchez, E. Sarrías, M. Strelec, B. Thompson, W. Waites, D. Wiggins and J. Zaragoza. 'ParaCrawl: Web-Scale Acquisition of Parallel Corpora'. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 4555–4567. DOI: 10.18653/v1/2020.acl-main.417. URL: https://aclanthology.org/2020.acl-main.417.

[92] 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, Oct. 2021. URL: https://hal.science/hal-03380805.

[93] 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.

[94] 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.

[95] 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.

[96] 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.

[97] 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.

[98] 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.

[99] 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/.

[100] 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.

[101] 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, Sept. 2021. URL: https://hal.inria.fr/hal-03357080.

[102] 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*. NASSCFL. Iowa City / Virtual, United States, May 2021, p. 5. URL: https://hal.science/hal-03187097.

[103] S. Goldwater, T. L. Griffiths and M. Johnson. 'A Bayesian framework for word segmentation: Exploring the effects of context'. In: *Cognition* 112.1 (2009), pp. 21–54. DOI: https://doi.org/10.1016/j.cognition.2009.03.008. URL: https://www.sciencedirect.com/science/article/pii/S0010027709000675.

[104] S. Goldwater, T. L. Griffiths and M. Johnson. 'Contextual Dependencies in Unsupervised Word Segmentation'. In: *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*. ACL-44. Sydney, Australia: Association for Computational Linguistics, 2006, pp. 673–680. DOI: 10.3115/1220175.1220260. URL: https://doi.org/10.3115/1220175.1220260.

[105] 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.

[106] 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.

[107] 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.

[108] 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.

[109] 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.

[110] 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.

[111] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh and K.-W. Chang. 'VisualBERT: A Simple and Performant Baseline for Vision and Language'. In: *CoRR* abs/1908.03557 (2019). arXiv: 1908.03557. URL: http://arxiv.org/abs/1908.03557.

[112] 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.

[113] 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).

[114] 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.

[115] 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.

[116] B. Muller, A. Anastasopoulos, B. Sagot and D. Seddah. 'When Being Unseen from mBERT is just the Beginning: Handling New Languages With Multilingual Language Models'. In: *NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Mexico City, Mexico, June 2021. URL: https://hal.inria.fr/hal-03251105.

[117] B. Muller, B. Sagot and D. Seddah. 'Can Multilingual Language Models Transfer to an Unseen Dialect? A Case Study on North African Arabizi'. working paper or preprint. Mar. 2021. URL: https://hal.inria.fr/hal-03161677.

[118] P. Ortiz Suarez and S. Gabay. 'A Data-driven Approach to Named Entity Recognition for Early Modern French'. In: *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 3722–3730. URL: https://aclanthology.org/2022.coling-1.327.

[119] 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.

[120] 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.

[121] 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.

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

[123] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger and I. Sutskever. 'Learning Transferable Visual Models From Natural Language Supervision'. In: *CoRR* abs/2103.00020 (2021). arXiv: 2103.00020. URL: https://arxiv.org/abs/2103.00020.

[124] 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.

[125] 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.

[126] 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.

[127] 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.

[128] 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.

[129] 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.

[130] 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.

[131] 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.

[132] 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.

[133] 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.

[134] 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.

[135] 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.

[136] 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.

[137] 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.

[138] 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.

[139] 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.

[140] 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.

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

[142] 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.

[143] 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.

[144] 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.

[145] 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.

[146] 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.

[147] 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.

[148] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy and S. Bowman. 'GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding'. In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 353–355. DOI: 10.18653/v1/W18-5446. URL: https://aclanthology.org/W18-5446.

[149] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou and H. Yang. 'OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework'. In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu and S. Sabato. Vol. 162. Proceedings of Machine Learning Research. PMLR, 17–23 Jul 2022, pp. 23318–23340. URL: https://proceedings.mlr.press/v162/wang22al.html.