RESEARCH CENTRE

Inria Paris Centre

2023
ACTIVITY REPORT

Project-Team
ALMANACH

Automatic Language Modelling and Analysis & Computational Humanities

DOMAIN

Perception, Cognition and Interaction

THEME

Language, Speech and Audio

Contents

Project-Team ALMANACH

Creation of the Project-Team: 2019 July 01

Keywords

Computer sciences and digital sciences

A3.1.1. – Modeling, representation

A3.1.7. – Open data

A3.1.8. – Big data (production, storage, transfer)

A3.1.11. – Structured data

A3.2.2. – Knowledge extraction, cleaning

A3.2.5. – Ontologies

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

A5.1. – Human-Computer Interaction

A5.7.3. – Speech

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

B9.6.1. – Psychology

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

1 Team members, visitors, external collaborators

Research Scientists

- Benoît Sagot [Team leader, Inria, Senior Researcher, HDR]
- Rachel Bawden [Inria, Researcher]
- Justine Cassell [Inria, Senior Researcher, from Nov 2023]
- Chloé Clavel [Inria, Senior Researcher, from Oct 2023, HDR]
- Thibault Clérice [Inria, Starting Research Position, from May 2023]
- Djamé Seddah [Inria, Associate Professor Detachement]
- Éric Villemonte de La Clergerie [Inria, Researcher]

Post-Doctoral Fellows

- Aina Garisoler [Télécom Paris, Post-Doctoral Fellow, from Oct 2023]
- Emer Gilmartin [Inria, Post-Doctoral Fellow, from Nov 2023]

PhD Students

- Alafate Abulimiti [Inria, from Oct 2023]
- Wissam Antoun [Inria, from Mar 2023]
- Alisa Barkar [Télécom Paris, from Oct 2023]
- Roman Castagné [Inria, until Oct 2023, Resigned from his PhD position]
- Alix Chagué [Inria, Université de Montréal (Canada), Joint supervision with Université de Montréal (Canada)]
- Lucie Chenain [Université Paris Cité, from Oct 2023]
- Cyril Chhun [Telecom-Paris, from Oct 2023]
- Floriane Chiffoleau [Université du Mans]
- Nicolas Dahan [Inria, from Oct 2023]
- Rasul Jasir Dent [Inria, from Nov 2023]
- Paul-Ambroise Duquenne [META, CIFRE]
- Matthieu Futeral-Peter [Inria, Co-supervised with the WILLOW project-team. Visited Google from Sept 2023-Dec 2023 (gap in contract)]
- Nathan Godey [Inria]
- Yanzhu Guo [École Polytechnique, from Oct 2023]
- Chadi Helwé [Télécom-Paris, from Oct 2023]
- Francis Kulumba [MINARM]
- Simon Meoni [Arkhn, CIFRE]
- Biswesh Mohapatra [Inria, from Nov 2023]

- Tú Anh Nguyen [META, CIFRE]
- Lydia Nishimwe [Inria]
- Arij Riabi [Inria]
- José Rosales Núñez [LISN, CNRS, until Jul 2023]
- Hugo Scheithauer [Inria, from Nov 2023]
- Lionel Tadonfouet Tadjou [Orange, CIFRE, until Mar 2023]
- Rian Touchent [Inria]
- Lorraine Vanel [Télécom Paris, from Oct 2023]
- Armel Zebaze Dongmo [Inria, from Nov 2023]
- You Zuo [qatent, CIFRE, from Feb 2023]

Technical Staff

- Julien Abadji [Inria, Engineer, until Aug 2023]
- Wissam Antoun [Inria, Engineer, until Feb 2023]
- Lauriane Aufrant [Inria, Engineer, from Nov 2023]
- Seth Aycock [Inria, Engineer, from Aug 2023 until Nov 2023]
- Niyati Sanjay Bafna [Inria, Engineer, until Jun 2023]
- Sarah Bénière [Inria, Engineer, from Oct 2023]
- Anna Chepaikina [Inria, Engineer, until Sep 2023]
- Cecilia Graiff [Inria, Engineer, from Dec 2023]
- Juliette Janès [Inria, Engineer, from Oct 2023]
- Jade Jenkins [Inria, Engineer, from Nov 2023 until Nov 2023]
- Tanti Kristanti Nugraha [Inria, Engineer, until Apr 2023]
- Marius Le Chapelier [Inria, Engineer, from Nov 2023]
- Menel Mahamdi [Inria, Engineer]
- Rua Mohamed Abdalla Ismail [Inria, Engineer, until Mar 2023]
- Virginie Mouilleron [Inria, Engineer]
- Oriane Nédey [Inria, Engineer, from Dec 2023]
- José Rosales Núñez [Inria, Engineer, from Aug 2023]
- Hugo Scheithauer [Inria, Engineer, until Oct 2023]
- Lionel Tadonfouet Tadjou [Inria, Engineer, from Mar 2023 until Sep 2023]
- Deepak Yadav [Inria, Engineer, from Apr 2023 until May 2023]

Interns and Apprentices

- Sarah Bénière [Inria, Intern, from Apr 2023 until Jul 2023]
- Patricia Camus [Inria, Intern, from Nov 2023]
- Léo Labat [Inria, Intern, from Nov 2023]
- Ilyas Lebleu [Inria, Intern, from Dec 2023]
- Samuel Scalbert [Inria, Intern, from Apr 2023 until Jul 2023]

Administrative Assistants

- Meriem Guemair [Inria, until Oct 2023]
- Christelle Rosello [Inria, from Nov 2023]

Visiting Scientist

- You Zuo [N/A, until Jan 2023]

External Collaborator

- Laurent Romary [Inria, DCIS, HDR]

2 Overall objectives

The ALMAnaCH project-team (Automatic Language Modelling and Analysis & Computational Humanities) [1] is a pluridisciplinary team focusing at the crossroads computer science, linguistics, statistics, and the humanities, focusing on **natural language processing**, **computational linguistics** and **digital and computational humanities and social sciences**.[2]

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,

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

[2] This section has not been changed since the creation of the ALMAnaCH project-team. It is obviously somewhat outdated. The project-team will be evaluated by Inria in 2024, which will give us the opportunity to update our overall objectives.

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-axes, are the following:

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

3 Research program

3.1 Research strands

ALMAnaCH's scientific programme is organised around three research axes.[3] 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

[3] This section has not been changed since the creation of the ALMAnaCH project-team. It is obviously somewhat outdated. The project-team will be evaluated by Inria in 2024, which will give us the opportunity to update our research programme. This will be particularly important given the recent arrival of Chloé Clavel and Justine Cassell in the team, which extends the scope of our research in a consistent yet significant way.

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 [84, 144] and one of the best for robust multilingual parsing as shown by our results in various shared tasks over the years [139, 135, 145], [100]. 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 [135].

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[4] and to Extrinsic Parsing Evaluation Shared Task[5].

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

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

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

features (namely information about the speaker and dialogue moves) was beneficial [103]. In the case of UGC, working across sentence boundaries was explored for instance, with limited success, by [82] for document-wise parsing and by [125] for POS tagging.

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

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

3.2.3 Information and knowledge extraction

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

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

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

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

3.3 Computational Modelling of Linguistic Variation

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

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

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

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

3.3.1 Theoretical and empirical synchronic linguistics

Permanent members involved: all

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

Recent advances in morphology rely on quantitative and computational approaches and, sometimes, on collaboration with descriptive linguists—see for instance the special issue of the *Morphology* journal on “computational methods for descriptive and theoretical morphology”, edited and introduced by [79]. 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 [97] 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 [118] 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 [80] and [99]. 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. [132]), 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 [119]. 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. [6]

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 [96, 133], 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 [75]). Lexical simplification, another aspect of text simplification [104, 120], 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[7] 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.

[6] 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).

[7] Site internet de GROBID.

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

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

3.4.1 Construction, management and automatic annotation of Text Corpora

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

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

ALMAnaCH pays a specific attention to the re-usability[8] 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

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

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

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

3.4.2 Development of Lexical Resources

ALPAGE, the Inria predecessor of ALMAnaCH, has put a strong emphasis in the development of morphological, syntactic and wordnet-like semantic lexical resources for French as well as other languages (see for instance [131, 90]). Such resources play a crucial role in all NLP tools, as has been proven among other tasks for POS tagging [127, 129, 145] 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 [146]), 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 [126, 101, 128], while using and developing (semi)automatic lexical information extraction techniques based on existing corpora [130, 134]. 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 [101, 128].

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) [123] 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, 122, 137, 108]) 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 [89]. 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)
- Language modelling
- Text generation, text simplification, automatic summarisation
- Spelling correction (writing aid, post-OCR, normalisation of noisy/non-canonical texts)
- Machine translation
- Chatbots, conversational agents, question answering systems
- Medical applications (analysis of medical documents, early diagnosis, language-based medical monitoring, etc.)
- Applications in linguistics (modelling languages and their evolution, sociolinguistic studies, etc.)
- Digital humanities (exploitation of text documents, for instance in historical research)

5 Social and environmental responsibility

5.1 Footprint of research activities

Given recent interest into the energy consumption and carbon emission of machine learning models, and specifically of those of language models [136, 78], we have decided to report the power consumption[9] and carbon footprint of all our experiments conducted on the Jean Zay supercomputer during 2023. For this

[9] Jean Zay documentation

Project	Type	GPU hours	Real hours	Power Consumption (kWh)	CO_2 Emissions (kg)
AD011013680R1	V100	11016.00	2754.00	5023.30	251.16
AD011013680R1	A100	1238.00	154.75	687.09	34.35
AD011013900R1	V100	5501.40	1375.35	2508.64	125.43
AD011012254R2	V100	30701.00	7675.25	13999.66	699.98
AD011012254R2	A100	6273.00	784.13	3481.52	174.08
AD010614012	A100	36232.00	3755.00	6849.12	342.46
AD011014393	V100	17.00	4.25	7.75	0.39
AD011013674R1	V100	11,553	2888,25	5,268.17	263.41
AD011013908	V100	37430.00	9357.50	17068.08	853.40
AD011012254	V100	23809.00	5952.25	10856.90	542.85
AD011014232	V100	9565.00	2391.25	4361.64	218.08
Total				83,371.50	4,168.57

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 associated with the team.

report, we follow the approach of [143]. While the ALMAnaCH team uses other computing clusters and infrastructures such as CLEPS[10] and NEF,[11] 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 limited compared 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. This could therefore somewhat compensates the non-reported consumption on both NEF and CLEPS.

Node infrastructure: We have access to two types of GPU node in Jean Zay:[12]

1. Nodes comprising 4 GPU Nvidia Tesla V100 SXM2 32GB, 192GB of RAM, and two Intel Cascade Lake 6248 processors. One Nvidia Tesla V100 card is rated at around 300W,[13] while the Intel Cascade Lake processor is rated at 150W.[14] For the DRAM we can use the work of [85] to estimate the total power draw of 192GB of RAM at approximately 20W. The total power draw of one Jean Zay node at peak use therefore adds up to around 1520W.

2. Nodes comprising 8 GPU Nvidia A100 SXM4 80GB, 512GB of RAM, and two AMD Milan EPYC 7543 processors. One Nvidia A100 card is rate at around 400W.[15] while the AMD Milan processor is rated at 225W.[16] Following [85], we estimate the total power draw of 512GB of RAM at approximately 50W. The total power draw of one A100 node at peak use therefore adds up to around 3700W.

With this information, we use the formula proposed by [143] 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

[10] CLEPS documentations

[11] NEF documenation

[12] Jean Zay architecture description

[13] Nvidia Tesla V100 specification

[14] Intel Xeon Gold 6248 specification

[15] Nvidia Tesla A100 specification

[16] AMD Milan EPYC 7543 specification

reported by 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 for V100 nodes and 8 for A100 nodes, 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 with the ALMAnaCH team during 2023.[17] 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 France, which were around 50g/kWh on average for 2023.[18] The total CO_2 emissions in kg for one single model can therefore be computed as:

$$\mathrm{CO_2e} = 0.05 p_t \tag{2}$$

All emissions are also reported in Table 1. The total emission estimate for the team is 4,168.57kg of CO_2. The carbon footprint of a single passenger on a round trip from Paris to New York (Boeing 787), flying economy, amounts to around 2300kg of CO_2.[19] This means that our computing emissions from Jean Zay for 2023 amount to just above one Paris-New York trip, and therefore the largest source of emissions from the team are from flights to conferences, which are a necessary part of communicating about our research to the international community.

6 Highlights of the year

6.1 Awards and recognition

- Benoît Sagot was elected to hold the annual chair in informatics and digital sciences at the Collège de France for the academic year 2023-2024.
- November 2023: Thibault Clérice and Alix Chagué won the Research Data Open Science Young Researchers prize for their resource HTR-United, a catalogue of metadata for transcription and segmentation datasets for HTR (hand-written text recognition).
- Lydia Nishimwe won the best paper award at the 2023 RECITAL student conference for her article entitled “Normalisation lexicale de contenus générés par les utilisateurs sur les réseaux sociaux” [47].

7 New software, platforms, open data

7.1 New software

7.1.1 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.

[17] This includes one project (AD010614012), which was joint with other research laboratoires and therefore its use cannot be solely attributed to experiments run by ALMAnaCH members.

[18] According to EDF’s website.

[19] co2.myclimate.org estimates for 2023.

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

7.1.2 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

7.1.3 HTR-United

Keywords: HTR, OCR

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

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

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

Release Contributions: First version.

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

Contact: Alix Chague

7.1.4 VGAMT

Name: Visually Guided and Adapted Machine Translation system

Keyword: Machine translation

Functional Description: Machine translation model that lets the user adding an image as input in addition to the sentence to be translated.

Publication: hal-03977982

Contact: Matthieu Futeral-peter

7.1.5 CamemBERTa

Name: a DeBERTa v3-based French language model

Keywords: Language model, French

Functional Description: CamemBERTa was initially evaluated on five distinct downstream tasks for French: part-of-speech (POS) tagging, dependency parsing, named entity recognition (NER), FLUE (French Language Understanding Evaluation), including natural language inference (NLI). It improves the state of the art for most tasks compared to previous monolingual and multilingual approaches, which again confirms the effectiveness of large pretrained language models for French. CamemBERTa is particularly effective for high-level tasks.

URL: `https://gitlab.inria.fr/almanach/CamemBERTa`

Contact: Wissam Antoun

7.1.6 CamemBERT-bio

Keywords: Language model, Deep learning, NLP, Transformer

Functional Description: CamemBERT-bio is a state-of-the-art french biomedical language model built using continual-pretraining from camembert-base. It was trained on a french public biomedical corpus of 413M words containing scientific documments, drug leaflets and clinical cases extrated from theses and articles. It shows significant improvement on multiple biomedical named entity recognition tasks compared to camembert-base.

URL: `http://camembert-bio-model.fr/`

Contact: Rian Touchent

7.1.7 CoMMuTE

Name: Contrastive multilingual and multimodal translation evaluation

Keywords: Machine translation, Evaluation, Image analysis

Functional Description: CoMMuTE is a contrastive evaluation dataset designed to assess the ability of multimodal machine translation models to exploit images in order to disambiguate the sentence to be translated. In other words, given a sentence containing a word that can be translated in several ways, the additional image determines the meaning of the word to be translated. The model must then take the image into account to propose a correct translation. CoMMuTE is available from English into French, German and Czech.

URL: `https://github.com/MatthieuFP/CoMMuTE`

Contact: Matthieu Futeral-peter

7.1.8 MANTa-LM

Name: Language Model on top of a Module for Adaptive Neural TokenizAtion

Keyword: NLP

Functional Description: Language Model on top of a Module for Adaptive Neural TokenizAtion

URL: https://huggingface.co/almanach/manta-lm-base

Contact: Nathan Godey

7.1.9 RoCS-MT

Name: Robust Challenge Set for Machine Translation

Keywords: Machine translation, NLP, Evaluation, Robustness, User-generated content

Functional Description: RoCS-MT, a Robust Challenge Set for Machine Translation (MT), is designed to test MT systems' ability to translate user-generated content (UGC) that displays non-standard characteristics, such as spelling errors, devowelling, acronymisation, etc. RoCS-MT is composed of English comments from Reddit, selected for their non-standard nature, which have been manually normalised and professionally translated into five languages: French, German, Czech, Ukrainian and Russian. The challenge set was included as a test suite at the WMT 2023 conference. This repository therefore also includes automatic translations from the submissions to the general MT task.

URL: https://github.com/rbawden/RoCS-MT

Publication: hal-04300824

Contact: Rachel Bawden

7.1.10 3MT_French Dataset

Name: 3 Minutes Thesis Corpus

Keywords: Multimodal Corpus, Video annotation

Functional Description: This new resource will be useful to computer science and social science researchers working on public speaking assessment and training. It will help refine the analysis of speaking from a fresh perspective based on social-cognitive theories rarely studied in this context, such as first impressions and theories of primacy and recency.

URL: https://zenodo.org/records/7603511#.Y9O4l3CZOUk

Publication: https://hal.science/hal-04366763

Contact: Chloe Clavel

7.1.11 feats2notes

Keywords: Natural language processing, Generative AI, Oenology, Wine, Text generation

Functional Description: The aim of this software is to generate comments about an object based on a list of properties associated with a quality and intensity. It initially organises relevant structured data according to the input, transforms it into coherent sentences and ensures their quality. It was initially developed to produce oenological comments from formalised descriptions of the organoleptic properties of wines, in collaboration with the company Winespace.

URL: https://github.com/anna-chepaikina/feats2notes

Contact: Benoit Sagot

7.1.12 CATMuS Medieval

Name: Consistent Approach to Transcribing ManuScripts - Medieval model

Keyword: Handwritten Text Recognition

Functional Description: CATMuS (Consistent Approach to Transcribing ManuScript) Medieval is a model for automatically transcribing medieval manuscripts using Latin scripts, in particular Old and Middle French, Latin, Spanish (and other languages of Spain), and Italian. The model was trained on the largest and most diverse dataset known for medieval manuscripts in Latin scripts, with more than 110 000 lines of training data.

Contact: Thibault Clerice

Partners: University of Toronto, Ecole nationale des chartes, CIHAM UMR 5648, VeDPH - Ca' Foscari, Université de Genève, ENS Lyon

7.1.13 HTRomance

Keyword: Handwritten Text Recognition

Functional Description: The ground truth produced as part of the HTRomance project aims to provide diverse data, from the 12th century to the 19th century, for training handwritten text recognition models. It covers the following languages: Latin, various states of French, Spanish, Occitan and Italian.

URL: https://htromance-project.github.io/

Contact: Thibault Clerice

Partners: VeDPH - Ca' Foscari, Ecole nationale des chartes, CIHAM UMR 5648, ENS Lyon

7.1.14 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

7.1.15 Expresso

Name: Expresso - A Benchmark and Analysis of Discrete Expressive Speech Resynthesis

Keywords: Speech processing, Speech Synthesis, Expressive rendering, Multimodal Corpus

Functional Description: Expresso est un corpus de parole expressive de haute qualité (48 kHz) qui comprend à la fois de la parole lue de manière expressive (8 styles, au format wav mono) et des dialogues improvisés (26 styles, au format wav stéréo). L'ensemble de données inclut 4 locuteurs (2 hommes, 2 femmes) et totalise 40 heures (11h de lecture, 30h d'improvisation). Les transcriptions de la parole lue sont également fournies. La tâche du Benchmark Expresso est de resynthétiser l'audio entrant en utilisant un code discret à faible débit qui a été obtenu sans supervision à partir du texte.

Contact: Tu Nguyen

Partner: META

7.1.16 SONAR

Keywords: Sentence embeddings, Natural language processing, Multimodality, Speech, Text, Machine translation, Zero-shot

Functional Description: SONAR (for Sentence-level multimOdal and laNguage-Agnostic Representations) is a multilingual and multimodal fixed-size sentence embedding space, with a full suite of speech and text encoders and decoders. It substantially outperforms existing sentence embeddings such as LASER3 and LabSE on the xsim and xsim++ multilingual similarity search tasks. Speech segments can be embedded in the same SONAR embedding space using language-specific speech encoders trained in a teacher-student setting on speech transcription data. We also provide a single text decoder, which allows us to perform text-to-text and speech-to-text machine translation, including for zero-shot language and modality combinations.

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

Publication: hal-04264028

Contact: Paul-Ambroise Duquenne

7.1.17 SpeechMatrix

Keywords: Speech processing, Parallel corpus

Functional Description: SpeechMatrix provides massive parallel speech which is mined from VoxPopuli in 17 languages: Czech (cs), German (de), English (en), Spanish (es), Estonian (et), Finnish (fi), French (fr), Croatian (hr), Hungarian (hu), Italian (it), Lithuanian (lt), Dutch (nl), Polish (pl), Portuguese (pt), Romanian (ro), Slovak (sk) and Slovenian (sl).

URL: https://github.com/facebookresearch/fairseq/tree/ust/examples/speech_matrix

Contact: Paul-Ambroise Duquenne

Partner: META

7.1.18 T-modules

Keywords: Machine translation, Natural language processing, Speech processing, Zero-shot, Multimodality

Functional Description: A new approach to zero-shot cross-modal transfer between speech and text for translation tasks. Multilingual speech and text are encoded in a joint fixed-size representation space, and then decoded to enable zero-shot translation between languages and modalities. All our models are trained without the need of cross-modal labelled translation data. Despite a fixed-size representation, we achieve very competitive results on several text and speech translation tasks, outperform the state of the art for zero-shot speech translation and introduce the first results for zero-shot direct speech-to-speech and text-to-speech translation.

Publication: hal-03834732

Contact: Paul-Ambroise Duquenne

7.2 Open data

ALMANaCH's general policy is to release all data and software under open-source licences when possible. This currently includes the following list. Please see the 'New Software' sections of this report and previous activity reports or the BIL for more detailed information.

- 3MT_French Dataset: 3 Minutes Thesis Corpus
- ACCESS (CC-BY-NC): Controllable Text Simplification Model
- Alexina: Morphological (and sometimes syntactic) lexicons (including the Le*fff*)
- ASSET (CC-BY-NC): Text Simplification Evaluation Dataset
- CamemBERT (MIT): Neural BERT-like language model for French
- CamemBERTa (MIT): A DeBERTa v3-based French language model
- CamemBERT-bio (MIT): Neural BERT-like language model for the French biomedical domain
- CCASS-sim: Similarity detection tool for legal texts from the Cour de Cassation
- CharacterBERT-UGC (CC-BY-SA): A CharacterBERT language model for North-African Arabizi and French user-generated content
- CoMMuTE (CC-BY-SA-4.0): A contrastive evaluation dataset for multimodal (text-image) machine translation.
- FSMB (CC-BY-NC-SA 4.0): French social media bank
- D'AlemBERT (Apache 2): Neural BERT-like language model for Early Modern French
- D'AlemBERT POS (Apache 2): POS tagger for Early Modern French
- D'AlemBERT NER (Apache 2): NER model for Early Modern French
- DESIR-CodeSprint-TrackA-TextMining: A tool for extracting scholarly documents and visualizing the results on PDF files using GROBID.
- DiaBLa (CC BY-SA 4.0): Parallel dataset of English-French bilingual dialogues
- DiscEvalMT (CC-BY-SA-4.0): Contrastive test sets for the evaluation of discourse phenomena in English-to-French machine translation
- DyALog (GPL-3.0): Environment for building tabular parsers and programs
- ELMoLex: Neural parsing system developed for ALMAnaCH's submission to the CoNLL-18 multilingual parsing shared task
- EtymDB (CC BY-SA 4.0): Etymological database extracted from wiktionary

- FreEM-corpora: Corpora and NLP tools for Early Modern French (16th-18th c.)
- FRMG (LGPL-3.0): A large-coverage meta-grammar for French
- grobid-medical-report: GROBID module for extracting and restructuring medical reports from PDF documents into encoded XML/TEI documents
- HTR-United (CC-BY): HTR-United is an open Github ecosystem designed to share training data for HTR and OCR tasks
- MANTa-LM (MIT): A robust T5-like model based on a neural tokenizer
- dyalog-sr (GPL-3.0): Transition-based parser built on top of DyALog
- EASSE (GPL-3.0): Text Simplification Evaluation Library
- entity-fishing: Entity recognition and disambiguation
- eScriptorium Documentation (CC-BY): Open and collaborative documentation for eScriptorium
- feats2notes (LGPL): Génération de commentaires à partir des données structurées
- FQB (CC-BY-NC-SA): Multi-layered treebank made of questions for French
- FrELMo: ELMo language model for French
- goclassy (Apache-2.0): asynchronous concurrent pipeline for classifying Common Crawl
- GROBID (Apache-2.0): Library for extracting, parsing and re-structuring raw documents
- GROBID-Dictionaries: GROBID module for structuring digitised lexical resources and entry-based documents
- KaMI-Lib (MIT): KaMI-lib is an HTR and OCR engine agnostic Python package for evaluating transcription models
- MElt (CeCILL-C): Statistical part-of-speech tagger
- Mgwiki: Linguistic Wiki for FRMG
- ModFr-norm (CC-BY-SA-4.0): Normalisation of Modern (17th c.) French
- MRELMo: ELMo language models for 5 mid-resource languages (Bulgarian, Catalan, Danish, Finnish, Indonesian)
- nerdKid: nerdKid is a tool for grouping Wikidata entities into 27 classes (e.g., ANIMAL, LOCATION, MEDIA, PERSON).
- OFrLex-modifier (AGPL-3.0)
- OSCAR (CC-BY): Huge multilingual web-based corpus
- PAGnol (MIT): Neural GPT-based language model for French
- PFSMB (CC-BY-NC-SA-4.0): FR-EN parallel corpus of noisy user-generated content
- PMUMT (CC-BY-NC-SA): FR-EN Annotated parallel corpus of noisy user-generated content
- RoCS-MT (CC-BY-NC): Robust Challenge Set for Machine Translation
- Sequoia corpus (LGPL-LR): French corpus with surface and deep syntactic annotations
- SSK (CC-BY): Collection of research use case scenarios illustrating best practices in Digital Humanities and Heritage research

- SxPipe (CeCILL-C): shallow language pipeline
- SYNTAX (CeCILL-C): lexical and syntactic parser generator
- tseval (CC-BY-NC): Text Simplification Evaluation Library
- UDLexicons: Multilingual collection of morphological lexicons
- Ungoliant (Apache-2.0): Asynchronous concurrent pipeline for classifying Common Crawl
- VerDI project release
- VGAMT (Apache-2.0): A multimodal machine translation model
- WikiCremma (CC-BY): Dataset for HTR training on Contemporary French
- WOLF (CeCILL-C): Free Wordnet for French
- CATMuS Medieval (CC-BY): Handwritten Text Recognition model for medieval manuscripts- in Latin scripts
- HTRomance: Ground-truth for training HTR models
- SONAR: SONAR is a multilingual and multimodal fixed-size sentence embedding space, with a full suite of speech and text encoders and decoders
- SpeechMatrix (CC-BY-SA): Speech parallel corpus mined from VoxPopuli
- T-modules: Approach to cross-modal transfer between speech adn text for translation tasks
- Expresso (CC-BY-SA): A Benchmark and Analysis of Discrete Expressive Speech Resynthesis

8 New results

8.1 Large Corpus Creation: From OSCAR to COLaF

Participants: Benoît Sagot, Thibault Clérice, Rachel Bawden, Julien Abadji, Rua Ismail, Juliette Janès, Oriane Nédey, Rasul Jasir Dent, Matthieu Futeral-Peter.

8.1.1 OSCAR: Collecting Text Corpora from the Web

Since the introduction of large language models (LLMs) in Natural Language Processing (NLP), 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 [76] and our own large-scale multilingual corpus OSCAR [116].[20] However, they have quality issues, especially for low-resource languages, an issue investigated in a large-scale study we were involved in and whose initial publication in 2021 [102] was followed by a publication in the *Transactions of the Association for Computational Linguistics* [102].

Recreating and updating these corpora is very complex. Since 2021, we have been developing Ungoliant [73, 72], a 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 used it to create several new versions of OSCAR that are larger and based on more recent data [73, 72], 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 (38% of ROOTS is based on OSCAR) [70].

[20] OSCAR website.

In addition to the improved pipeline, we also have also added additional information to OSCAR. Since version 22.01, OSCAR has included 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 (amongst other aspects) large-scale improvement of the language identification mechanism, additional annotations. Ungoliant is released under an open-source licence and we publish the corpus under a research-only licence.

In early 2023 we published OSCAR 23.01, which is based on the November/December 2022 dump of Common Crawl. While being quite similar to OSCAR 22.01, it contains several new features, including KenLM-based [94] adult content detection, precomputed Locality-Sensitive Hashes for near deduplication, and blocklist-based categories. OSCAR 23.01 has also moved from gzip to Zstandard compression.

In 2023 we have also initiated the development of a new, multimodal version of OSCAR, in collaboration with engineers working at GENCI on the Jean Zay supercomputer. Our goal is to collect a large number of images from the internet associated with snippets of text (such as, but not limited to, their captions), the type of data required to train models that combine text and image information (e.g. for multimodal machine translation, where image-based information is used to improve machine translation quality, for which see Section 8.4.1).

This work is still being carried out in close collaboration with Pedro Ortiz, although he left ALMAnaCH mid-2022 after a successful PhD defence [113] and in 2023 was working at DFKI Berlin. Other people outside of ALMAnaCH regularly interact with the "OSCAR team", in particular Sebastian Nagel from CommonCrawl.

8.1.2 COLaF: Collecting High-quality Text Corpora for French and Other Languages of France

In close interaction with the (informal) OSCAR project and in collaboration with our colleagues from the MULTISPEECH team in the Nancy Inria Center (Emmanuel Vincent, Slim Ouni) and the former ALMAnaCH member Laurent Romary, as well as with the Inria headquarters' support (in particular Jean-Frédéric Gerbeau), we finalised the creation of COLaF ("Corpus et Outils pour les Langues de France" — corpora and tools for the languages of France), an Inria "DEFI" (Inria-internal multi-team project) jointly led by Benoît Sagot and Slim Ouni. The goal of COLaF is to contribute to the development of free corpora and tools for French and other languages of France, in close collaboration with academic and institutional partners.

The scope of COLaF includes both:

- Text data, for which ALMAnaCH is responsible, with Benoît Sagot as COLaF's co-PI and Thibault Clérice as COLaFALMAnaCH's project manager,
- Speech and sign language data, for which MULTISPEECH is responsible.

COLaF aims to cover French and the languages of France in all its diversity:

- It aims to have as diverse a coverage as possible: French from France and elsewhere, regional languages, French-based creoles (including outside France), indigenous languages, migrant languages and French sign language.
- All aspects of variation will be studied, beyond the standard state of the language, including specialised languages, diachrony, non-standard states (user-generated content, learner language, etc.).
- Activity within the project notably covers the acquisition and structuring of texts from non-textual sources (books, audio recordings, etc.), the classification by language and linguistic variety of large volumes of texts (in close connection with the OSCAR project), the development of annotation and transformation models (translation, normalisation, voice synthesis, sign language generation) serving the development of corpora and the exploitation of newly created resources.

Since its creation in August, and even prior to that (thanks to funding from Benoît Sagot's PRAIRIE chair), we have been working in multiple directions, including:

- We have interacted with a number of relevant external partners:

  - language-specific institutions, including Académie Régionale de la Langue Picarde (Picard) and Lo Congrès (Occitan),
  - research laboratories such as LILPa (Université de Strasbourg and CNRS, Alsacian and other languages), LISA (Université de Corse and CNRS, Corsican), MoDyCo (Université Paris-Ouest and CNRS, Breton and other languages).

- We have started to work on text corpus representation based on the TEI guidelines, with a special interest in a complete and homogeneous metadata scheme.

- We have resumed our work on language identification, at the intersection between the OSCAR and COLaF project, in the context of Rasul Dent's newly started PhD (funded by COLaF).

8.2 Language Models

Participants: Benoît Sagot, Djamé Seddah, Éric de La Clergerie, Rachel Bawden, Roman Castagné, Nathan Godey, Wissam Antoun, Niyati Sanjay Bafna.

Pretrained language models are now ubiquitous in NLP. Despite their success, many early models were either trained on English data or on the concatenation of data in multiple languages [86, 105]. One of the most visible achievements of the ALMAnaCH team was the training and release of CamemBERT in 2019, a BERT-like [86] (and more specifically a RoBERTa-like) neural language model for French trained on the French section of our large-scale web-based OSCAR corpus [116] (see Section 8.1.1), together with CamemBERT variants [107] and ELMo models trained on OSCAR corpora for other languages, including French [114, 115]. In 2023, we trained and published a new language for French, an alternative for CamemBERT based on the DeBERTaV3 architecture [93]. This model, called CamemBERTa [19], further improves the state of the art for French NLP.

However the challenge posed by low-resource languages is still a major one (see Section 8.7), and is one of the main topics of interest of language-model-related research at ALMAnaCH. In particular, as part of our collaboration with the DFKI involving the joint supervision of Niyati Bafna, we pursued our work 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. We focused on five (extremely) low-resource dialects from the Indic dialect continuum (Braj, Awadhi, Bhojpuri, Magahi, Maithili), which are closely related to each other and the standard mid-resource dialect, Hindi. We evaluated a number of strategies that broadly include from-scratch pretraining and cross-lingual transfer between the dialects as well as from different kinds of off-the- shelf multilingual models; we found that a model pretrained on other mid-resource Indic dialects and languages, with extended pretraining on target dialect data, consistently outperforms other models [43]. We subsequently investigated how bilingual lexicon extraction strategies, and which ones, can further improve very low-resource language modelling for these language varieties closely related to Hindi (see Section 8.7 for more details).

Another new line of work is the detection of language-model-generated content. It is both a key challenge to avoid training future language models on an excessive amount of language-model-generated data, which would create a risk for the quality of the models, but also to help understand how such data could be generated and used to influence the public either directly or via future language models. In the context of Wissam Antoun's PhD thesis, supervised by Benoît Sagot and Djamé Seddah, we proposed a methodology for developing and evaluating ChatGPT detectors for French text, with a focus on investigating their robustness on out-of-domain data and against common attack schemes. The proposed method involves translating an English dataset into French and training a classifier on the translated data. Results show that the detectors can effectively detect ChatGPT-generated text, with a degree of robustness against basic attack techniques in in-domain settings. However, vulnerabilities are evident in out-of-domain contexts, highlighting the challenge of detecting adversarial text [42].

Another crucial research direction is gaining a better understanding of how language models actually work and how they can be improved, which is the focus of Nathan Godey and Roman Castagné's work (both supervised by Benoît Sagot and Éric de La Clergerie. In this context, we worked on the three following questions:

- What if we could pretrain masked language models such as RoBERTa using only artificially created data, then continue pretraining on languages with very few resources? Ideally, our synthetic data could be generated quickly and be scaled to extremely large sizes as well as have properties that allow the models to "learn something". Unfortunately, despite our best efforts to incorporate some key characteristics of natural language in our synthetic dataset (a structured way of generating sequences and a notion of co-occurrence of tokens), we were unable to induce biases in the Transformer network that would transfer to the pretraining on a target language with smaller resources.[21]

- Why are vector representations in Transformer-based language models very often anisotropic, sometimes to a suprising extent? Some recent works hinted towards the idea that anisotropy could be a consequence of optimising the cross-entropy loss on long-tailed distributions of tokens. In 2023, we showed that anisotropy could also be observed empirically in language models with specific objectives that should not suffer directly from the same consequences [65].[22] We also showed that the anisotropy problem extends to Transformers trained on other modalities. Our observations tend to demonstrate that anisotropy might actually be inherent to Transformers-based models.

- How can we speed up language model training? Self-supervised pre-training of language models usually consists in predicting probability distributions over extensive token vocabularies. We proposed an innovative method that shifts away from probability prediction and instead focuses on reconstructing input embeddings in a contrastive fashion via *Constrastive Weight Tying* (CWT) [64].[23] By applying this approach, we pretrain what we refer to as "headless language models" in both monolingual and multilingual contexts. Our method offers practical advantages, substantially reducing training computational requirements by up to 20 times, while simultaneously enhancing downstream performance and data efficiency. We observe a significant +1.6 GLUE score increase and a notable +2.7 LAMBADA accuracy improvement compared to classical language models within similar computing budgets.

Finally, in 2023 we also continued our work on the BLOOM language model [70],[24], whose development heavily involved a number of ALMAnaCH members in 2021 and 2022. This year we have also been working on the evaluation of BLOOM specifically for French over a range of different NLP tasks. The project, led by François Yvon under the informal name Bloume, received advanced support from IDRIS. The results will be released to the community to provide a point of comparison for future research.[25]

8.3 Text-based Machine Translation (MT)

Participants: Rachel Bawden, Benoît Sagot, Djamé Seddah, Lydia Nishimwe, José Rosales Nuñez, Seth Aycock, Nicolas Dahan, Armel Zebaze Dongmo, Éric Villemonte de La Clergerie.

8.3.1 MT for specific domains and low-resource scenarios

One of the major challenges for the development of high quality MT models is adapting them to specific domains, or even ideally to multiple domains at once. We have made several contributions to MT for specific domains this year.

Firstly, Rachel Bawden was one of the co-organisers of two shared tasks at WMT 2023, the main conference on MT: the general MT task, which evaluates MT models over a range of domains [34], and the biomedical MT task, which evaluates MT models on their ability to translate biomedical texts [36]. ALMAnaCH is also a partner of the MaTOS ANR project (PI François Yvon, CNRS), dedicated to the MT of

[21] Roman Castagné's blog post on this work.
[22] An updated version of the paper has just been accepted for publication at EACL 2024.
[23] A significantly updated version of the paper has just been accepted for publication at ICLR 2024.
[24] This 2023 version of the paper is a minor update of the original 2022 paper.
[25] Github repository

scientific documents for English–French and French–English [44]. Two PhD students are co-supervised by Rachel Bawden jointly with François Yvon in the projet: Nicolas Dahan on the evaluation of MT for scientific documents and Ziqian Peng (recruited by the CNRS) on document-level MT. The first research carried out in the project has focused on the NLP domain, using titles and abstracts of articles from the HAL platform and postediting automatic translations from a variety of MT models. We analysed the performance of the different models and studied the differences between postedits produced by professional translators and those produced by authors in the NLP community.[26]

Adapting neural MT models to specific domains was the aim of the DadaNMT project led by Rachel Bawden and involving Jesujoba Alabi (engineer in 2022) and Seth Aycock. We explored two different strategies for adaptation: (i) the integration of matched terms from bilingual lexicons to overcome tye presence of otherwise unknown terms, and (ii) the selection of few-shot examples when translating using LLMs, guided by topic models. In the first work,[27] we explored the commonly used strategy of adding additional information inline in source sentences in a bid to offer insights into whether this can be more than a copy mechanism. We found that gains were limited using this method but that it is typically more than just a copy mechanism, and models are capable of selecting or ignoring terms amongst those given, as appropriate. In the second approach,[28] we explored several strategies of using topic models to perform domain adaptation of LLMs for MT, including appending topic-based keywords, topic labels and few-shot examples from specific topics.

This second approach is linked to the topic of Armel Zebaze's PhD on analogy for low-resource NLP, supervised by Rachel Bawden and Benoît Sagot, which began in November 2023. He has been studying the use of different language models for few-shot example selection, particularly to help the translation of low-resource language pairs. The topics of domain adaptation and low-resource processing are in fact highly linked for two main reasons: (i) many domains are low-resource and therefore many of the techniques used in low-resource MT can be used for adaptation to specific domains and (ii) adapting a model to a new language can be seen as analogous to adapting a model to a new domain, particularly in multilingual models. A further contribution to low-resource MT was the use of multilingual language models for Indian languages. In [40], in the context of Sonal Sannigrahi's 2021 internship, we studied the use of transliteration and segmentation strategies to facilitate cross-lingual transfer and few-shot transfer.

8.3.2 MT applied to non-standard texts

User-generated content (UGC) such as texts found on social media are characterised by multiple phenomena not typically present in standard edited texts and which present challenges for MT (e.g. spelling mistakes, acronyms, truncations, and contractions). It is important to develop MT models that are robust, which means they are able to translate these kinds of text just as well as if the texts had not displayed non-standard variation. In 2023, we explored both the development of robust models and the creation of evaluation dataset specific to non-standard texts.

In the context of José Rosales Nuñez's PhD thesis co-supervised by Djamé Seddah with the Laboratoire de Linguistique Formelle (Université de Paris),[29] we explored the use of variational MT for the translation of UGC and proposed an extension using mixture density networks and normalising flows, based on the hypothesis that having independent latent distributions could help model different UGC specificities [39]. Our proposed model showed results comparable or superior to standard state-of-the-art variational MT models, whilst improving performance on UGC translation. We provide analysis into the behaviour of neural learning representations when processing non-standard data. For example, our analysis of the models reveals that the representations of non-standard sentences have a higher cosine similarity with respect to their normalised versions than for the standard variational model. Finally, we confirm that the learned embeddings are more robust, since they can be used to initial the embeddings of a a vanilla transformer model to produce a more robust model.

An alternative strategy to developing more robust models is to normalise the non-standard text before applying tools trained on standard data. The lexical normalisation step is itself challenging, as it requires

[26] The paper is currently under submission at an international conference.

[27] This work is currently under submission at an international conference.

[28] This work is currently under submission at an international workshop.

[29] José defended his thesis entitled *Machine Translation of User-Generated Contents : an Evaluation of Neural Translation Systems under Zero-shot Conditions* [53] on 3rd October 2023.

determining which words require normalising and how to choose standard variants to replace them with. In the context of her PhD on robustness MT, supervised by Rachel Bawden and Benoît Sagot, Lydia Nishimwe explored the challenges faced by this task, by surveying the literature, providing preliminary experiments using various masked language models combined with levenshtein distance, and discussing the problems of evaluation for the task [47]. Her paper won the best paper award at the RECITAL 2023 conference.[30] She has since been working on building robust sentence embeddings, using the test case of Laser, and, inspired by T-modules (research carried out in the context of Paul-Ambroise Duquenne's PhD) using distillation to reduce the distance between non-standard sentences and their normalised versions.[31]

8.3.3 MT Evaluation

Research into evaluation is important for the development of models. On the subject of MT robustness and evaluation, we developed a new evaluation set designed to test MT models' capacity to handle specifically non-standard texts [20]. RoCS-MT (Robust Challenge Set for Machine Translation) is composed of non-standard English texts from the Reddit forum, which we manually normalised and had professionally translated into 5 languages, French, German, Czech, Ukrainian and Russian (See the resource in Section 7.1.9), funded by Rachel Bawden's PRAIRIE chair position. RoCS-MT was submitted as a test suite to the WMT 2023 general translation task, which meant that we could evaluate translations produced by all models submitted to the task. We found that most models still struggle with many non-standard phenomena, in particular those that alter the spelling of words.

In a follow-up to the BigScience Workshop [70], we also completed the evaluation of the BLOOM language model in collaboration with François Yvon (CNRS) [21]. We chose to study a range of domains, language pairs, resourcedness levels, as well as other aspects such as prompt choice, number of few-shot examples, the use of cross-lingual few-shot examples and the influence of linguistic context. As an extension to this work and as described in Section 8.2, we have also been working on the evaluation of BLOOM for a range of French NLP tasks.

Finally, as mentioned above, Nicolas Dahan, supervised by Rachel Bawden and François Yvon in the context of the MaTOS ANR project, began a PhD on the evaluation of MT for scientific documents. He has begun to explore the current state of the art in MT metrics and analyse how well adapted they are to evaluating document-level aspects such as lexical cohesion.

8.4 Multimodal Machine Translation

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

8.4.1 Image-enhanced MT

In the context of Matthieu Futeral-Peter's PhD thesis, co-supervised by Rachel Bawden, Benoît Sagot and Cordelia Schmid (WILLOW project-team) and in collaboration with Ivan Laptev (ex-member of WILLOW), we have continued developing new approaches to multimodal (text-image) MT. Our main focus has been on ambiguity, which presents one of the major challenges in MT, and on how to exploit visual context to resolve it. Previous work in the field has shown that obtaining improvements from the addition of 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.

Our new multimodal method, named VGAMT (Visually Guided and Adapted Machine Translation) [2] (see the model in Section 7.1.4) is based on a strong pre-trained MT model, fine-tuned using adapters on two training objectives: the multimodal MT objective (exploiting multimodal parallel sentences, each accompanied by an image) and visually conditioned masked language modelling (using English captioning data). We found that combining these two objectives was necessary to force the model to

[30] Conference page

[31] This work is currently under submission at an international conference.

effectively use visual context. The model's performance is further boosted by the use of a novel guided self-attention mechanism, by which irrelevant connections between the text and the image parts are masked, thereby guiding the model to consider the relevant connections. An important aspect of this work was also evaluation, standard metrics being ill-adapted and misleading for reporting progress on this task, due to the lack of adequate test data. We released CoMMuTE (see the resource in Section 7.1.7), a Contrastive Multilingual Multimodal Translation Evaluation dataset, composed of ambiguous sentences and their possible translations, accompanied by disambiguating images corresponding to each translation. For three language directions (English–{French, Czech and German}), we show that our method remains competitive against text-only models on standard benchmarks and performs above all other models by a large margin on CoMMuTE, whereas the previous state-of-the-art models do not beat the baseline.

8.4.2 Multimodal MT with speech

This subsection is directly inspired by the introduction of Paul-Ambroise Duquenne's PhD thesis (not yet published).

In the context of Paul-Ambroise CIFRE PhD thesis, co-supervised by Benoît Sagot and Holger Schwenk (META), we had previously investigated the use of fixed-sized sentence embeddings spaces for multiple purposes, with a focus on text-to-text, speech-to-text, text-to-speech and speech-to-speech MT. We first introduced multilingual and multimodal speech/text sentence embeddings using a teacher/student approach with the existing LASER sentence embedding space [88]. We demonstrated that we could perform semantic similarity estimation between speech and text in different languages and introduce speech mining as an extension of bi-text mining for the speech modality. We trained speech translation systems using the mined data and demonstrate significant gains with this additional data. Based on these promising results, we scaled speech-to-speech mining to 136 language pairs to introduce the SpeechMatrix corpus and trained several speech translation systems on this mined data [28].

We then explored how to efficiently decode these fixed-size representations into multiple languages and modalities, and how to perform zero-shot cross-modal machine translation in this framework. We demonstrated that we can combine independently trained encoders and decoders from different languages and modalities in a zero-shot way to perform cross-modal translation [32]. As a second step, published this year, we explore multilingual training in such modular framework, to benefit from cross-lingual learning [29].

This year, we drew conclusions from our previous work and introduced SONAR, a state-of-the-art massively multilingual speech/text sentence embedding space for both cross-lingual and cross-modal similarity search as well as decoding capabilities [63]. We recently complemented these semantic sentence representations with a modality specific representation encoding non-semantic speech properties of an audio signal.

8.5 Speech Modelling

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

The content of this subsection is partly inspired by the abstract of Robin Algayres's PhD,[33] *co-supervised by Emmanuel Dupoux, from the CoML team at Inria Paris, EHESS and ENS, and by Benoît Sagot.*

Unsupervised text-based language modelling techniques and challenges can be a source of inspiration for 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 former CoML team at Inria Paris and with specialists of speech processing at META.

Spoken word discovery is the task of uncovering words from the speech signal, a task which presents two technical challenges: (i) the learning of speech sequence embeddings (SSEs), fixed-size vectors that represent variable-size sequences of speech, and (ii) the segmentation of speech into words. To learn SSEs

[32] duquenne:hal-03834732

[33] His PhD was defended on the 26th September 2023, and his thesis is currently awaiting publication by the doctoral school.

we trained a neural network on top of self-supervised features (Wav2vec2.0) using data augmentation, contrastive learning and iterative self-training. For the challenge of speech segmentation, we developed DP-Parse [74], a Dirichlet Process non-parametric Bayesian model inspired by [91]. DP-Parse finds frequent patterns in the speech signals using density estimation on speech fragments represented with SSEs. Frequent patterns serve as anchors to perform full-sentence segmentation into word tokens. On the Zero-Resource Speech Benchmark 2017 Speech Segmentation Task [87], 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. In 2023, we published new results that further improve DP-Parse segmentation performances by fine-tuning an XLS-R model to predict DP-Parse boundaries [18].

We used our work on spoken word discovery, to tackle a second challenge of spoken language modelling. Our goal was to train a spoken language model on speech recordings segmented using DP-Parse and represented by sequences of SSEs. Instead of discretising the SSEs with clustering, which we argue is challenging task because of Zipf's Law, we decided to adapt the training and inference steps of our spoken language model to the continuous SSEs [17]. In order to evaluate the semantic and syntactic representations learned by our SLM, we introduced two ABX metrics that were sensitive to the quality of the input speech segmentation. Regarding spoken generation, our model was able to generate utterances that were on par with the state of the art in intelligibility and meaningfulness as assessed by automatic metrics as well as human raters.

In an effort to further challenge how speech sequences should be representated, 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 Benoît Sagot and Emmanuel Dupoux (this time as a META fellow), our scientific question was whether such a discrete bottleneck is necessary, given that it potentially introduces irreversible errors in the encoding of the speech signal, and whether we could 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 [111] [66]. We showed that it removes linguistically irrelevant information from the continuous features, helping to improve language modelling performance. More recently, we studied how we could improve the invariance of such discrete input representations to non-spoken augmentations for generative spoken language modeling [32]. We proposed to apply a set of signal transformations to the speech signal and to optimise the model using an iterative pseudo-labelling scheme. Our method significantly improved over the evaluated baselines when considering encoding and modelling metrics, and we also saw better results than the baselines on the speech-to-speech translation task, considering Spanish–English and French–English translations.

In collaboration with several 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 [110][14]. 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.

Finally, we also carried out work in dataset creation, with the development of EXPRESSO, a high-quality expressive speech dataset for textless speech synthesis that includes both read speech and improvised dialogues rendered in 26 spontaneous expressive styles [37]. Recent work had shown that it was possible to resynthesise high-quality speech based, not on text, but on low bitrate discrete units that have been learned in a self-supervised fashion and can therefore capture expressive aspects of speech that are hard to transcribe (prosody, voice styles, non-verbal vocalisation). The adoption of these methods was still limited by the fact that most speech synthesis datasets were read, severely limiting spontaneity and expressivity. EXPRESSO therefore fulfils this gap in the literature.

8.6 Hate Speech and Radicalisation Detection

Participants: Djamé Seddah, Arij Riabi, Wissam Antoun, Virginie Mouilleron, Menel Mahamdi, José Rosales Nuñez.

8.6.1 Analysing Zero-Shot Transfer Scenarios across Spanish Variants

In previous work on multilingual zero-shot hate-speech detection [109], we highlighted the need to alleviate the cultural gap found in cross-lingual scenarios. An example of the issue of the cultural gap is the use of the slur *puta*, which can translate to English "prostitute", as a non-offensive intensifier in Spanish, which we found to be a strong trigger for our models trained on an English misogyny-oriented dataset. While this cultural gap has been highlighted for different languages from different countries and has been under increasing scrutiny recently, the question of its impact on NLP models when applied to dialects of the same language remains. Language variation between geolects such as American English and British English, Latin-American Spanish, and European Spanish is still a problem for NLP models, which often rely on (latent) lexical information for their classification tasks. More importantly, the cultural aspect just mentioned, which is crucial for hate speech detection, is often overlooked.

Along this line of research, we therefore carried out a thorough analysis of hate speech detection models for different variants of Spanish and introduced a new dataset of hate speech directed toward immigrants from Twitter. Using mBERT [117] and Beto [81], a monolingual Spanish BERT-based language model, as the basis of our transfer learning architecture, our results indicate that hate speech detection models for a given Spanish variant are affected when different variations of the language are not considered. When compared to multilingual model such as mBERT, these differences are accentuated by the specific nature of its training data (Wikipedia, no discussions). Our intuitions were confirmed with an error analysis conducted with the SHAP interpretability framework [106], which highlighted the vulnerability of language models fine-tuned on another geolect to cultural-specific hateful terms. In an era where cross-cultural issues in NLP are becoming increasingly important [98, 112, 95], our work and methodology constitute an interesting step in this process. This work was published in a EACL2023's specialised workshop on language variation (VARDIAL 2023) [23].

8.6.2 Multilingual Radicalisation Detection in the CounteR Project

Since 2021 ALMAnaCH has been the leader of the NLP part of the CounteR H2020 European project devoted to radicalisation detection (grant agreement n°101021607). Part of this work consists in building a radicalisation classifier to be integrated in the CounteR platform, a prototype targeting Law Enforcement Agencies. The classifier is trained on two sources of data: data acquired by the consortium from a third-party covering French, English and Arabic, and data targeting six others languages (Portuguese, German, Romania, Latvian, Greek and Bulgarian) produced through the translation and adaption of the English data. A third source of data, currently used for evaluation, was produced by prompting LLMs following different user scenarios.

We based our classification architecture on the multilingual multitask framework we developed from [92] and detailed in [109]. As the basis for our multi-task classifier, we used XLMR-T, a large multilingual XLM-R model [83] fine-tuned on a masked-language modelling task on 200M tweets [77]. We then jointly trained the classifier on our radicalisation dataset and on a variety of auxiliary tasks (part-of-speech tagging, dependency parsing, named entity recognition (NER) and sentiment analysis). In order to cope with code-switching and to avoid having to run language identification, we concatenated the training data for all languages into one training set. The performance gain compared to the use of monolingual models greatly exceeded expectations (+6 points for Arabic for example, despite it being written in a different script). As our deliverables are UE restricted, we unfortunately cannot provide further details of our results in a publicly available report.

An important part of our work this year was the improvement of our model by producing and including domain-specific NER annotations for our core languages (Arabic, English, French). We took advantage of this time-consuming task to also pseudo-anonymise all entities and non-public URLs. Unlike the current state of the art, we decided to keep all socio-demographic variables that can be expressed through aliases, the topic of the URLs, hashtags and Telegram channels in order to preserve the meaning. The idea is to make sure that a classifier trained on pseudo-anonymised data would still benefit from lexical information while being able to distribute the corpus in a GDPR-compliant way. Note that no other corpus of this type exists with this granularity of annotation and in particular for this target domain. Its release will most certainly establish a notable milestone.

Regarding the potential biases of our classifiers, we are involved in a bias assessment of our models with one of the partners of the project, Ethicas Tech, an auditing firm. Based on target profiles we identified, we generated a synthetic dataset using the Vicuna Uncensored LLM,[34] for about 1800 documents for French and English and annotated them following the same guidelines used to build the initial dataset. Of course, being purely generated, there is no intent in this type of content *per se*, so analysing biases against non-existent users is arguable. However, given GDPR restrictions and the volatile nature of social media content (most of the most radical content is quickly deleted by the platforms), using such synthetic data can be considered a plausible proxy for such studies. The content we generated will also be used to train an LLM-based radical content detector using the same methodology we demonstrated in [42], where we released the first adversarial dataset for the detection of ChatGPT content.

8.6.3 Enriching the Narabizi Treebank

One of the target languages of the CounteR project is Arabic in all its forms, including dialectal Arabic. Given our expertise on handling user-generated content, we focused on the North-African Arabic dialect, which displays significant variability across regions and predominantly exists in spoken form and lacks standardised spelling when written. Many Arabic speakers use the Latin script to write their dialects online, using digits and symbols for phonemes not easily mapped to Latin letters [138]. This written form, called ARABIZI and its North African variant, NARABIZI, often showcases code-switching, with French with up to 36% of French words in the NARABIZI dataset [138].

In order to increase the usefulness of the NARABIZI treebank in a multitask framework where task transfer is better realised through higher level tasks (e.g. NER, Sentiment analysis) than lower level ones (e.g. morpho-syntactic analysis via part-of-speech tagging and dependency parsing) [109, 121], we added two more annotation layers to the treebank: named-entities and offensive language annotation. We also conducted an extensive consistency checking that led to re-annotating many linguistic phenomena. In particular, we switched from the light tokenisation inspired by the French Social Media Bank [140] to a more classic morphology-based tokenisation. This step was made necessary by the inclusion of the named entity annotation layer, as we wanted to avoid creating multiple entities differing only in the the use of determiners. This step enabled us to conduct (surprisingly) the first experiments on NER with non-gold tokenisation. This work was published in a specialised workshop (Linguistic Annotation Workshop) colocated with ACL 2023 [38].

8.7 Low-resource NLP

Participants: Arij Riabi, Djamé Seddah, Benoit Sagot, Niyati Sanjay Bafna, Rachel Bawden, Armel Zebaze Dongmo.

State-of-the-art NLP models typically require large quantities of training data. Multiple research topics during 2023 were characterised by a lack of such data, presenting a challenge to the development of high-quality models and requiring specific strategies to compensante.

Some of these works have been mentioned in previous sections, particularly those concerning MT. As described in Section 8.3, (i) multilingual Indian-language MT models were used for the translation of low-resource language pairs [40], bilingual lexicons were integrating into MT models in order to adapt to new domains, and (iii) we also explored the selection of few-shot examples to help adapt to new domains and to under-resourced language pairs. For speech translation (Section 8.4.2), we contributed to the development of SONAR, a highly multilingual and fixed-size sentence embedding space for both speech and text [63] and to SpeechMatrix, a large-scaled and multilingual dataset of mined speech translations

[34] /cognitivecomputations/Wizard-Vicuna-13B-Uncensored on HuggingFace

[28]. In terms of resource creation, as described in Section 8.6, we also enriched the NArabizi treebank with two additional annotation layers (NER and offensive language detection) [38]. Finally, as described in Section 8.2, we also explored optimal strategies for cross-lingual LM-learning strategies for extremely low-resource Indian languages [43].

In addition to these, an additional contribution can be cited for 2023, namely an unsupervised approach to bilingual lexicon induction (BLI) applied to extremely low-resource languages of the Indic dialect continuum [55].[35] We introduced a simple, iterative strategy requiring inference from a BERT language model for a higher-resource, related language (in our case Hindi). This accommodates for the fact that state-of-the-art approaches, which typically require good embeddings models for all languages concerned, show near-zero performance in this low-resource settings. We worked with two low-resource languages (<5M monolingual tokens), Bhojpuri and Magahi and repeated our experiments on Marathi and Nepali, two higher-resource Indic languages, to compare approach performances by resource range. We released automatically produced bilingual lexicons for five languages of the Indic dialect continuum, for which lexicons have not been previously been available.

8.8 Biomedical NLP

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

In the context of the BPI-funded project ONCOLAB, we explored the application of NLP techniques to the medical domain. The medical domain is a very specialised one, and clinical documents (EHR – Electronic Health Records) are very sensitive data and not readily available, especially for French. Therefore, during his Master's thesis, Rian Rouchent tried several approaches to collect a dataset of French biomedical documents (biomed-fr) and used it to continue the pretraining of the CamemBERT language model [107] [7], leading to CamemBERT-bio [49], one of the first French language models specialised for the biomedical domain. Continual pre-training ensured a very low environmental impact (compared to a pre-training from scratch) and CamemBERT-bio exhibits state-of-the-art performance on several French biomedical benchmarks. However, as shown by its performance in the DEFT 2023 shared task for multiple choice question answering, coupling a language model with specialised medical knowledge seems necessary. This coupling is being investigated as part of Rian's PhD, supervised by Éric de La Clergerie, in particular via the potential use of augmented LLMs. The lack of biomedical data and especially annotated biomedical data also led Simon Meoni to explore the use of LLMs (in particular Instruct-GPT) to automaticaly annotate medical entities in the E3C corpus (without any prior fine-tuning)in the context of CIFRE PhD (with the company Arkhn), supervised by Éric de La Clergerie for ALMAnaCH. He used the annotated data to fine-tune a BERT-like model and was able to get good performance, which was further improved by combining LLM-generated annotations with dictionary-based annotations [35, 45]. To overcome the issue of the necessary confidentiality of medical data, which leads to a lack of any freely available datasets, he is also currently exploring the generation of synthetic documents, which are close to real documents without containing any confidential information.

8.9 NLP for Patents

Participants: You Zuo, Benoît Sagot, Éric Villemonte de La Clergerie.

As part of You Zuo's CIFRE PhD, in collaboration with the start-up qatent, supervised by Benoît Sagot and Éric de La Clergerie for ALMAnaCH and Kim Gerdes for qatent (see Section 9), we have been carrying out research into patent classification. Patent classification is a particularly difficult task, because of the complex taxonomies of domains necessary for defining patent scope and for patent law. For example, the International Patent Classification (IPC) has over 70,000 leaf nodes and five levels of hierarchy, and the task is further complicated by the fact that it is regularly updated, and that variations across

[35] This work is currently under submission at an international conference.

countries and markets makes the transfer of models or data across languages and domains difficult. We proposed a new approach to the classification of French patents, published in [50], relying on data-centric strategies. We compared different strategies for the two deepest levels of the IPC hierarchy, and showed that while simple ensemble strategies work for shallower levels, deeper levels require more sophisticated techniques such as data augmentation, clustering, and negative sampling. Our experiments highlighted the importance of language-specific features and data-centric strategies for accurate and reliable French patent classification.

In recently submitted (and therefore currently unpublished) work, we also performed a comparative evaluation and analysis of language models for the task of patent generation, focusing on two different tasks: (i) the generation of abstract from claims and (ii) the generation of claims given previous ones. We developed a benchmark, PatentEval, and a comprehensive error typology for both tasks. We manually annotated various models, including those specific to the patent domain as well as general-purpose language models. In addition, we designed and evaluated automatic metrics to approximate human judgments.

8.10 Email Thread Constitution

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

Lionel Tadonfouet Tadjou concluded his PhD work, co-supervised by Éric de La Clergie and Laurent Romary, on conversation disentanglement in the context of e-mail discussions, his defense taking place in October 2023.[36] Strongly inspired by discourse theories and in particular dialogue acts and communication acts, he designed, trained, and evaluated several neural architectures with the final goal of detecting pairs of related dialogue acts (such as suggestion/agreement or disagreement) in e-mail threads, in order to confirm that the mails took place in an ongoing conversation (possibly in complement to other features based on metadata or semantic similarity) but also potentially providing some hints about the current status of a conversation (for instance, a non-answered request). One of the difficulties was the lack of annotated for such a task, for English and even more for French, leading to an important effort adjusting existing corpora to his needs.

8.11 Automatic generation of oenological descriptions

Participants: Anna Chepaikina, Benoît Sagot.

We continued our collaboration with the start-up company Winespace, with the aim of developing a wine recommendation system using information extraction from wine descriptions. This represents the second phase in the project, following an initial collaboration in 2020.

In particular, we improved the coherence of how the relevant concepts (e.g. acidity, red fruits, etc.) are mentioned in the texts generated, and we also integrated the ability to take into account intensity and quality for each concept. Finally, we worked on the integration of our algorithms into the Winespace architecture.

The success of the collaboration led to Anna Chepaikina, the engineer hired to work on the project, being recruited by the start-up at the end of her contract at Inria.

8.12 Information Extraction from Specialised Collections

[36] The first 3 years of the PhD were carried out as a CIFRE PhD with Orange and the last 6 months were carried out fully at ALMAnaCH. Lionel's thesis will be publicly available from 19th April 2024.

Participants: Alix Chagué, Floriane Chiffoleau, Hugo Scheithauer, Tanti Kristanti, Sarah Bénière, Thibault Clérice, Juliette Janès, Cecilia Graiff, Menel Mahamdi, Éric de La Clergerie.

In the context of DataCatalogue [16, 51], a project with the Bibliothèque nationale de France (BnF) and the Institut national d'histoire de l'art (INHA), we opened a new PhD position to explore the expansibility of the content extraction pipeline. In tandem with the DEFI COLaF, Hugo Scheithauer (PhD candidate) and Sarah Bénière actively contributed to an annotation campaign of digitised catalogs, seeking alternatives to Grobid [124] that might offer broader applicability. Building on the developments of the preceding year, we continued the normalisation and modelling of the structure of such catalogues.

Our collaboration within the international EHRI project led to an augmented publication workflow for digital editions [61]. The current iteration of the pipeline facilitates the extraction of text from digitised documents (both manuscripts and typed materials) and transforms them into a standardised representation in XML-TEI. Subsequent post-processing enables NER annotations. This facet aligns with our engagement in the broader NER4Archives initiative, and we are sustaining collaboration with the Archives Nationales.

Within the context of the DAHN project, a revamped version of DiScholEd was launched at discholed.huma-num.fr. The application is designed to facilitate the publication of digital editions adhering to the XML-TEI standard, featuring the display of rich annotations, such as NER annotations, integration with digital facsimile, and support for document translation.

Building on the success of a DARIAH grant, our involvement extended to the creation of documentation and outreach materials, including videos and tutorials. This collaborative initiative, spearheaded by the Institut Historique Allemand (IHA), proved to be highly successful.

In the context of the LIRIAe project with the Ministry of Ecological Transition (MTE), Menel Mahamdi and Éric de La Clergerie explored the processing of very large PDF documents (containing several hundreds of pages) describing project proposals with potential environmental impacts (such as the establishment of a windmill field). The motivation is to enrich the documents with annotations such as named entities and pertinent concepts in order to speed up reading (inside a web-based PDF reader) and to provide a more powerful and accurate search engine. Work has been done on the design of the reader, the extraction of entities and concepts, with a preliminary phase of knowledge acquisition from a corpus and the identification of existing ontologies to identify pertinent (lexicalised) concepts. Given the complexity of the PDF document, it was also necessary to develop a new piece of software, named PDFStruct, which is still in progress, in order to retrieve the documents' logical structure (sections, sub-sections, tables, figures, etc.) and textual content.

8.13 Automatic Text Recognition for Historical Documents

Participants: Thibault Clérice, Alix Chagué, Floriane Chiffoleau, Hugo Scheithauer, Juliette Janès, Sarah Bénière.

2023 saw several of our projects with handwritten text recognition (HTR) come to fruition, and our partnership with various projects and institutions led to several new research outputs. Specifically, we were involved in a large-scale study [69] of the provenance of data in the domain of automatic text recognition (ATR), through the role of both Alix Chagué et Thibault Clérice in the standardisation of ground truth cataloguing with HTR-United [1]. HTR-United has continued to grow, with 80 registered datasets, including more than 43 million characters and 1 million lines, spanning over 21 different languages and 7 scripts. Alix and Thibault were awarded the Research Data Open Science Young Researchers prize for HTR-United in November 2023.

Through several workshops and presentations [26, 57, 58, 60]), we have participated in the efforts of bringing up to date with the latest tooling for textual data acquisition and publication of their ground truth. This was done in part in the context of our continued participation in the development of eScriptorium, albeit mostly as an advisor entity in 2023.

In 2023, our work on both contemporary and medieval handwritten text recognition resulted in a few new research products, both on the model side [25, 68] and on the dataset side [59, 13]. We played an essential role in the release of the two biggest open models for Latin scripts, specifically for the medieval era with the first model built on a multilingual, large-scale, diachronic dataset, including data for five main languages (Latin, Spanish, Italian, Old French and Occitan) and less represented languages as of 2023 (Navarese, Venitian, etc.). It has been shown that it can be successfully fine-tuned with very little data on new languages not included in its training set (mainly Old English and Middle Dutch). This model benefitted from our work in the context of the grant of the Bibliothèque nationale de France: the HTRomance project, involving a dozen research teams from France and Italy. The CATMuS model broadened this partnership with the University of Toronto (for Latin manuscripts), Princeton and Antwerp Universities (for Middle Dutch manuscripts).

We were also involved in generative artificial intelligence initiatives for HTR, specifically on medieval manuscripts, to overcome issues with the cost of data production [52].

As a result of the CREMMACall published in 2022, we granted access to more than 30 projects over the course of a year, and managed to start the migration of the Traces6 servers to the CREMMA infrastructure, which was documented in late 2023 [71].

In the context of the DEFI COLaF and our ongoing work on improving the segmentation of documents with digital proto-edition in mind as a target application, we starting developing the LADaS dataset, focusing on a varied set of documents, both historical and contemporary, with the aim of extracting clean text for language models. This works builds on [12] and we are currently involved in implementing its vocabulary as a base vocabulary for the eScriptorium software while expanding the original controled and publish vocabulary for contemporary data. A partnership with the Persée digital academic library started in late 2023 and will probably come to fruition in 2024. The ability to treat Persée's data could lead to the production of an academic dataset of more than 10 million pages of academic papers in a standardised and clean format.

Finally, since June 2023, we have been involved in a collaboration with the CARAMBA team in Nancy and the Université de Picardie for the automatic acquisition of encrypted historical documents from the 16th century. The current state of historical diciperhing relies mostly on manual work by researchers. We have started to prototype a workflow in which we are able to acquire a transcription of an ad-hoc script (i.e. not alphabetical but more symbolic) for the CARAMBA team to post-process with decryption. This led in early 2024 to the submission of an Action Exploratoire with Thibault Clérice as one of the three main researchers of the project.

8.14 NLP for Historical and Literary Sources

Participants: Thibault Clérice.

This research area is new for the team, with the arrival of Thibault Clérice. Unlike automatic text recognition and content extraction, NLP for historical and literary sources focuses on bringing NLP techniques to humanities debates, bringing another point of view to questions such as authorship of old texts and historical trends.

In the context of CHR 2023 [27], Thibault Clérice proposed a methodological paper evaluating the use of Siamese networks in authorship verification for Ancient Greek texts and evaluated it against former research to validate the approach. The approach differs from traditional approaches used for authorship verification in computational humanities, which tend to use SVMs or similar methods for the purpose of explainability. It also differs from current approaches in NLP by using extracted features such as the ones used in computational humanities. This paper was a first step in ongoing work on the attribution of authorship of certain texts whose authorship is debated to John Chrysostom, an author with allegedly more than 1000 wrongly attributed texts over the course of four centuries. Current work focuses on a specific text, the *Quod Christus Sit Deus*, the authorship of which the community has been debating for the past 10 years. Validating a new method was important for further research.

Aside from authorship verification, a collaboration on large-scale corpus production and mining for historical trends analysis was at the centre of another work [22]. This work is done in combination with an ongoing project on large corpus mining for Latin manuscripts, with a new 120M-token corpus from manuscripts. The aims are to provide new elements for research in the coming years with manuscript mining. One of the first areas of research will be the continuation of sexual content detection in Latin [62].

9 Bilateral contracts and grants with industry

Participants: Benoît Sagot, Rachel Bawden, Djamé Seddah, Éric Villemonte de La Clergerie, Lionel Tadonfouet Tadjou, Tu Anh Nguyen, Paul-Ambroise Duquenne, You Zuo, Anna Chepaikina, Chloé Clavel, Justine Cassell.

9.1 Bilateral contracts with industry

Verbatim Analysis

Participants: Benoît Sagot.

Partner type: Inria start-up

Leader for ALMAnaCH: Benoît Sagot.

Dates: 1 Jul 2009–31 Dec 2024

Description: Verbatim Analysis is an Inria start-up co-created in 2009 by enoît 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. Its activities have been progressively taken over by opensquare (see below), and the company will be discontinued at the end of Jan 2024.

opensquare

Participants: Benoît Sagot.

Partner type: Inria start-up

Leader for ALMAnaCH: Benoît Sagot.

Dates: 1 Dec 2016–present

Description: Opensquare was co-created in December 2016 by Benoît Sagot with 2 senior specialists of human resources (HR) consulting. It is dedicated to designing, carrying out and analysing employee surveys as well as HR consulting based on these results. It uses a new employee survey analysis tool, enqi, which is still under development. 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 Inria based on its turnover. Benoît Sagot currently contributes to opensquare, under the "Concours scientifique" scheme.

META AI

Participants: Benoît Sagot, Tú Anh Nguyen, Paul-Ambroise Duquenne.

Partner type: Company

Leader for ALMAnaCH: Benoît Sagot.

Dates: 1 Jan 2018–present

Funding received: 210,000€

Description: Our collaboration with META AI is centered around the joint supervision of CIFRE PhD theses. A first collaboration (Louis Martin's PhD thesis), co-supervised by Benoît Sagot, Éric de La Clergerie and Antoine Bordes (META) was dedicated to text simplification ("français Facile À Lire et à Comprendre", FALC), in collaboration with UNAPEI. This collaboration was part of a larger initiative called Cap'FALC involving (at least) these three partners as well as the relevant ministries. Louis defended his PhD in 2022. Two other joint PhD theses started in 2021 and will be defended in 2024. Paul-Ambroise Duquenne's PhD, co-supervised by Benoît Sagot and Holger Schwenk (META), is dedicated to sentence embeddings for massively multilingual speech and text processing. Tú Anh Nguyen's PhD, co-supervised by Benoît Sagot and Emmanuel Dupoux (META), is dedicated to the unsupervised learning of linguistic representations from speech data, with a focus on textless dialogue modelling and speech generation.

In addition, Benoît Sagot was one of the very few academic researchers who received a $50,000 gift grant from META AI.

Winespace

Participants: Benoît Sagot, Anna Chepaikina.

Partner type: Start-up

Leader for ALMAnaCH: Benoît Sagot.

Dates: 1 Sept 2019–30 Sep 2023

Funding received: 32,391€

Description: The collaboration with this start-up company, dedicated to information extraction from wine descriptions to develop a wine recommendation system, was carried out in 2020 following previous discussions, in collaboration with Inria Bordeaux's "InriaTech" structure. In 2021, we designed and prepared a second step for this collaboration, which involved the hiring of Anna Chepaikina, a dedicated research engineer who started at the end of March 2022. After twelve months working in the framework of a "plan de relance" scheme, she was hired at Inria for a further six months fully funded by Winespace in the context of a new bilateral agreement. Since then, she has been hired by Winespace as an NLP expert.

INPI

Participants: You Zuo, Benoît Sagot, Éric de La Clergerie.

Partner type: Public institution

Leader for ALMAnaCH: Benoît Sagot.

Dates: 1 Oct 2021–30 Nov 2023

Funding received: 99,999.99€

Description: A collaboration with the Institut National de la Propriété Industrielle (France's patent office) started in 2021. A research engineer, You Zuo, was hired for a one-year contract to work on patent classification, later extended to 14 months. This project also involved an informal collaboration with the qatent startup.

Qatent

Participants: You Zuo, Benoît Sagot, Éric de La Clergerie.

Partner type: Inria start-up

Leader for ALMAnaCH: Benoît Sagot.

Dates: 1 Jan 2021–present

Description: Qatent is a startup supported by the Inria Startup Studio and ALMAnaCH that applies NLP to help write better patents faster. It follows the 18-month stay at ALMAnaCH of Kim Gerdes, one of the three founders of the company, and benefits from ALMAnaCH's scientific expertise and the Inria Startup Studio's counselling and financial support. You Zuo's CIFRE PhD thesis, co-supervised by Benoît Sagot and Kim Gerdes (now at qatent), is the latest development of this collaboration.

Orange

Participants: Éric de La Clergerie, Lionel Tadjou Tadonfouet.

Partner type: Company

Leader for ALMAnaCH: Éric de La Clergerie.

Dates: 1 Mar 2020–28 Feb 2023

Funding received: 30,000€

Description: The collaboration between ALMAnaCH and Orange is centered around a joint CIFRE PhD thesis dedicated to conversation disentanglement. The CIFRE convention ended in February 2023, and Lionel Tadjou defended his PhD in September 2023.

9.2 Active collaborations without a contract

AXA ReV

Participants: Djamé Seddah.

Partner type: Company

Leader for ALMAnaCH: Djamé Seddah.

Dates: 1 Oct 2020–present

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

LightON

Participants: Djamé Seddah.

Partner type: Start-up

Leader for ALMAnaCH: Djamé Seddah.

Dates: 22 Sept 2020–present

Description: LightON builds Optical Processor Units, a specialised line of processor able to outperform GPUs on certains tasks. We are working with them to see if we can use their technology to speed up the training of large language models. This informal collaboration has already resulted in the design, training and publication of the PAGnol generative language model for French.

zaion

Participants: Chloé Clavel, Lorraine Vanel.

Partner type: Company

Leader for ALMAnaCH: Chloé Clavel.

Dates: 1 Feb 2022–1 Mar 2025

Funding received: 16,000€

Description: CIFRE PhD thesis between Telecom-Paris[37] and Zaion in order to develop conversational systems integrating socio-emotional strategies in an explicit way. The CIFRE contract being with TelecomParis, it is mentioned in this section rather than in the previous one.

10 Partnerships and cooperations

10.1 European initiatives

10.1.1 H2020 projects

H2020 EHRI "European Holocaust Research Infrastructure"

Participants: Hugo Scheithauer, Floriane Chiffoleau, Alix Chagué, Sarah Bénière.

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)

[37] This partnership was set up before Chloé Clavel joined Inria. Part of the budget (16k euros) will be transferred from Telecom-Paris to Inria.

- 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.

CounteR

Participants: Djamé Seddah, Arij Riabi, Wissam Antoun, Mouilleron Virginie, Menel Mahamdi, Syrielle Montariol, Deepak Yadav, José Carlos Rosales Núñez, Galo Castillo Lopez.

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 radicalisation and thus prevent future terrorist attacks from taking place, the CounteR 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 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 in the psychology of radicalization processes to provide a complete solution for law enforcement authorities to understand the when, where and why of radicalization in the community.

10.2 National initiatives

ANR BASNUM

Participants: Benoît Sagot.

Duration: 1 Oct 2018–30 Jun 2023.

PI: Geoffrey Williams (Université de Grenoble).

Coordinator for ALMAnaCH: Laurent Romary.

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

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

ANR MaTOS

Participants: Rachel Bawden, Éric de La Clergerie, Nicolas Dahan.

Duration: 1 Jan 2023–31 Dec 2026.

PI: François Yvon.

Coordinator for ALMAnaCH: Rachel Bawden.

Partners:
- Université de Paris
- CNRS

Summary: The MaTOS (Machine Translation for Open Science) project aims to develop new methods for the machine translation (MT) of complete scientific documents, as well as automatic metrics to evaluate the quality of these translations. Our main application target is the translation of scientific articles between French and English, where linguistic resources can be exploited to obtain more reliable translations, both for publication purposes and for gisting and text mining. However, efforts to improve MT of complete documents are hampered by the inability of existing automatic metrics to detect weaknesses in the systems and to identify the best ways to remedy them. The MaTOS project aims to address both of these issues.

ANR TraLaLaM

Participants: Rachel Bawden, Benoît Sagot, Thibault Clérice, Juliette Janès.

Duration: 1 Oct 2023–30 Sept 2026.

PI: Josep Crego (SYSTRAN).

Coordinator for ALMAnaCH: Rachel Bawden.

Partners:

- SYSTRAN
- CNRS

Summary: The aim of TraLaLaM is to explore the use of large language models (LLMs) for machine translation, by asking two main questions: (i) in what scenarios can contextual information be effectively used via prompting? and (ii) for low-resource scenarios (with a focus on dialects and regional languages), can LLMs be effectively trained without any parallel data?

10.2.1 Competitivity Clusters and Thematic Institutes

3IA PRAIRIE

Participants: Benoît Sagot, Rachel Bawden, Roman Castagné, Nathan Godey, Lydia Nishimwe, Matthieu Futeral-Peter, Julien Abadji, Rua Ismail, Alafate Abulimiti.

Duration: 1 Oct 2019–31 Dec 2023.

PI: Isabelle Ryl.

Coordinators for ALMAnaCH: Benoît Sagot, Rachel Bawden and Justine Cassell.

Partners:

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

- Vertex

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

Benoît Sagot holds a PRAIRIE chair and Rachel Bawden holds a junior PRAIRIE chair.

LabEx EFL

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

Duration: 1 Oct 2010–30 Sept 2024.

PI: Barbara Hemforth (LLF).

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

Summary: Empirical foundations of linguistics, including computational linguistics and natural language processing. ALMAnaCH's predecessor team ALPAGE was one of the partner teams of this LabEx, which gathers a dozen of teams within and around Paris whose research interests include one aspects of linguistics or more. Several ALMAnaCH members are now "individual members" of the LabEx EFL. B. Sagot serves as deputy head (and former head) of one of the scientific strands of the LabEx, namely strand 6 dedicated to language resources. Benoît Sagot and D; Seddah are (co-)heads of a number of scientific "operations" within strands 6, 5 ("computational semantic analysis") and 2 ("experimental grammar"). Main collaborations are related to language resource development (strands 5 and 6), syntactic and semantic parsing (strand 5, especially with LIPN [CNRS and U. Paris 13]) and computational morphology (strands 2 and 6, especially with CRLAO [CNRS and Inalco] and LLF [CNRS and Paris-Diderot]).

GDR LiLT

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

Duration: 1 Jan 2019–present.

Summary: Linguistic issues in language technology.

10.2.2 Other National Initiatives

Informal initiative Cap'FALC

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

Duration: 1 Jan 2018–present.

Coordinator for ALMAnaCH: Benoît Sagot.

Partners:
- UNAPEI
- FAIR

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

Convention (MIC, Archives Nationales) NER4archives

Participants: Cecilia Graiff.

Duration: 1 Jan 2020–27 Nov 2024.

PI: Laurent Romary.

Coordinator for ALMAnaCH: Laurent Romary.

Partners:
- Ministère de la culture
- Archives Nationales

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

TGIR Huma-Num

Participants: Benoît Sagot, Thibault Clérice.

Duration: 1 Jan 2013–present.

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

DIM Matériaux Anciens et Patrimoniaux

Participants: Alix Chagué, Thibault Clérice.

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.

Convention (MIC) DataCatalogue

Participants: Hugo Scheithauer, Sarah Bénière.

Duration: 12 Aug 2021–31 Oct 2024.

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.

Sorbonne Emergence DAdaNMT

Participants: Rachel Bawden, Seth Aycock.

Duration: 1 Feb 2022–31 Dec 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.

Contrat PIA (AMI santé numérique) OncoLab

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

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.

BNF Datalab HTRomance

Participants: Thibault Clérice, Alix Chagué.

Duration: 1 Jan 2023–31 Dec 2023.

PIs: Thibault Clérice, Alix Chagué and Laurent Romary.

Coordinators for ALMAnaCH: Thibault Clérice and Alix Chagué.

Summary: The HTRomance project is based on handwriting recognition (HTR). In particular, it proposes to evaluate and improve the capabilities of this technology when applied to literary manuscripts and public and private archives, in Latin and Romance languages, from the 11th to the 19th century, kept at the French National Library. The main objective of the project is the production of training data and transcription models resistant to changes in handwriting and language. It also intends to produce language models applicable to documents in ancient languages or to ancient language states. The development of training corpora will be accompanied and consolidated by the development and implementation of a novel process for evaluating the readability of output texts and the costs of producing new training data for HTR. HTRomance is complementary to editing or data mining projects: the models produced are likely to be used to obtain the textual data needed for editing or text mining.

DEFI Inria COLaF

Participants: Benoît Sagot, Thibault Clérice, Rachel Bawden, Juliette Janès, Rasul Dent, Oriane Nédey.

Duration: 1 Aug 2023–31 Jul 2027.

PIs: Benoît Sagot and Slim Ouni.

Coordinator for ALMAnaCH: Benoît Sagot.

Partner:
- MULTISPEECH (Inria Nancy)

Summary: The Inria DEFI COLaF (Corpus and Tools for the Languages of France) aims to strengthen the ecosystem of automatic text and speech processing for the languages and speakers of France. To do this, it aims to create open datasets and use them to develop open-source models and tools.

10.2.3 Regional Initiatives

Framework agreement with Inria AP-TAL

Participants: Éric de La Clergerie, Benoît Sagot.

Duration: 1 Apr 2020–present.

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

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

Partner:
- APHP

Summary: Within the AP-TAL and HopiTAL projects, ALMAnaCH is involved in collaborative work with APHP and other Inria teams whose goal is 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.

11 Dissemination

Participants: Rachel Bawden, Benoit Sagot, Djame Seddah, Eric Villemonte De La Clergerie, Chloé Clavel, Justine Cassell, Thibault Clerice, Alix Chagué, Hugo Scheithauer, Lydia Nishimwe, Wissam Antoun, Floriane Chiffoleau, Nathan Godey, Rian Touchent, Lauriane Aufrant, Niyati Sanjay Bafna, Sarah Benière, Samuel Scalbert, Tú Anh Nguyen, Lauriane Aufrant.

11.1 Promoting scientific activities

Participants: Benoît Sagot, Rachel Bawden, Justine Cassell, Djamé Seddah, Chloé Clavel, Thibault Clerice, Floriane Chiffoleau, Alix Chague, Hugo Scheithauer, Nathan Godey, Niyati Sanjay Bafna, Tú Anh Nguyen, Lauriane Aufrant.

11.1.1 Scientific events: organisation

Member of the organizing committees

- Floriane Chiffoleau: Member of the organising committee for *From Source to Full Text: Workshop on Using Automatic Character Recognition (ATR)*.
- Rachel Bawden: Member of the organising committee of the WMT general shared task and the WMT biomedical shared task.

Inria-internal events

- Niyati Bafna and Wissam Antoun: Organisers of the ALMAnaCH reading group (Niyati until 06/2023).
- Rachel Bawden: Organiser of the PRAIRIE colloquium (2 talks) and ALMAnaCH seminar series (10 ALMAnaCH seminars, 3 streamed as part of B. Sagot's invited professorship at the Collège de France).

11.1.2 Scientific events: selection

Reviewer and Member of the Conference Program Committees

- Alix Chagué: Reviewer for DH2024.
- Chloé Clavel: Senior area chair for EACL and NAACL.
- Justine Cassell: Reviewer for ACL Rolling Reviews.
- Nathan Godey: Reviewer for ACL 2023 and EMNLP 2023.
- Rachel Bawden: Reviewer for ACL Rolling Reviews, ACL SRW 2023 (Student research workshop), EAMT 2023 and WMT 2023. Area chair for ACL 2023 ("Resources and Evaluation" track) and EMNLP 2023 ("Machine Translation" track).
- Benoît Sagot: Senior area chair for EMNLP 2023.
- Djamé Seddah: Senior area chair for ACL 2023. Reviewer for CAWL 2023, C3NLP and LAW-XVII 2023.
- Thibault Clérice: Reviewer for DH2024.
- Tú Anh Nguyen: Reviewer for ASRU Workshop on Speech Foundation Models.

11.1.3 Journal

Member of the editorial boards

- Chloé Clavel: Member of the editorial board for *IEEE Transactions of Affective Computing*.
- 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)*.
- Hugo Scheithauer: Reviewer for *Humanités numériques*.
- Rachel Bawden: Reviewer for *Journal of Specialised Translation (JoSTrans)* (Special issue on Translation Automation and Sustainability), *Journal of Language Engineering* (Special issue on the Role of Context in Neural Machine Translation Systems and its Evaluation), *Transactions on Audio, Speech and Language Processing* and *Revue TAL* (Special thematic issue on “Robustesse et limites des modèles de traitement automatique des langues”).
- Djamé Seddah: Reviewer for *Computational Linguistics*.
- Thibault Clérice: Reviewer for *JDMDH (Journal of Data Mining & Digital Humanities)*.

11.1.4 Invited talks

- Rachel Bawden:
  - Laboratoire Informatique de Grenoble, Université de Grenoble, France (28 Mar 2023): “From Linguistic to Visual Context in Machine Translation”.
  - ANITI-PRAIRIE workshop, Toulouse, France (27 Jun 2023): “From Linguistic to Visual Context in Machine Translation”.
- Benoît Sagot:
  - Journée d’étude “La littérature au prisme des humanités numériques” (One-day workshop organised by the ObTIC - Sorbonne Université) (16 Mar 2023): “Une approche computationnelle pour l’étude scriptométrique du français du XVIIème siècle”.
  - European Masters Program “Language & Communication Technologies” Annual Meeting 2023 (19 Jun 2023): “Language models and their training data: experiments and challenges”.
  - PRAIRIE colloquium, PariSanté Campus (16 Mar 2023): “Combining modalities: two experiments on multimodal NLP”. Website
  - George Mason University, Fairfax, Virginia, USA (21 Apr 2023): “NLP beyond text”. GMU Linguistics Colloquium Series
- Hugo Scheithauer:
  - Journée d’étude “Traitements automatique pour les humanités numériques” - Université Paris-Nanterre (15 Mar 2023): “DataCatalogue : un projet pour la restructuration automatique des catalogues de vente”. Website
- Chloé Clavel:
  - Seminar LORIA Nancy (17 Nov 2023): “Socio-conversational AI: Modelling the socio-emotional component of interactions using neural models”. Website
  - Amphi21 Sciences Po (30 Nov 2023): “Nouveaux outils, nouveaux usages. Quels enjeux pour une intelligence artificielle transparente et responsable ?”. Website

- Alix Chagué and Floriane Chiffoleau:
  - Institut historique allemand (DHIP/IHA), Paris (7 Sept 2023): "ATR: What can eScriptorium do for you?". From Source to Full Text: Workshop on Using Automatic Character Recognition (ATR)
  - Institut historique allemand (DHIP/IHA), Paris (8 Sept 2023): "What can you do next? Choice of output and reuse of your transcription". From Source to Full Text: Workshop on Using Automatic Character Recognition (ATR)
- Alix Chagué:
  - Institut historique allemand (DHIP/IHA), Paris (8 Sept 2023): "Image Acquisition and Layout Analysis". Joint with Hippolyte Souvay. From Source to Full Text: Workshop on Using Automatic Character Recognition (ATR)
  - Atelier Antologie Grecque, Sorbonne Univ. Paris (6 Mar 2023): "EScriptorium and automatic text recognition". Collaborative edition of the Greek Anthology - practical workshops
- Justine Cassell:
  - MBZ University of Artificial Intelligence, Abu Dhabi (12 Dec 2023): "What future for human-computer interaction in the era of AI".
  - Grenoble Institute Polytechnique (28 Nov 2023): "Sociality in Language and Thought".
  - World Economic Forum, Davos (16 Dec 2023): "Empowering Humans and Machines in Industry". Various talks in the context of the Davos meeting
  - INSEI (Suresnes): Laboratoire Graphes (17 Nov 2023): "Programmable Virtual Peers for Autism".
  - Canopé: Réseaux de Formation des Enseignants (15 Nov 2023): "L'IA Generative et l'Education".
- Thibault Clérice:
  - Institut Élie Cartan de Lorraine, Nancy (21 Nov 2023): "Par-delà le mot: détecter l'expression de la sémantique sexuelle dans un corpus latin en diachronie longue". Journée Fédération Charles Hermite : Représentations du Langage
  - Centre de Recherche Inter-Universitaire en Humanités Numériques, Montréal, Canada (13 Oct 2023): "Détection de la sémantique sexuelle en langue latine". Colloque du CRIHN 2023
- Djamé Seddah:
  - Jelinek Summer Workshop on Speech and Language Technology, JSALT, Le Mans Universités (26 Jun 2023): "Faut-il avoir peur du grand méchant GPT ? Démystifions les modèles de langues". Website
- Floriane Chiffoleau:
  - Institut historique allemand (DHIP/IHA), Paris (8 Sept 2023): "Text recognition and correction". Joint with Ariane Pinche, Hippolyte Souvay. From Source to Full Text: Workshop on Using Automatic Character Recognition (ATR)
  - University of Oslo, Norway (online) (21 Sept 2023): "TEI Publisher, a platform for sustainable digital editions". Workshop Digital Scholarly Editions in the Nordic Region - Where are we now? Where will we be in 10 years?
- Lauriane Aufrant:
  - Bibliothèque nationale de France (24 Nov 2023): "Fiabiliser sans entraver, comprendre pour utiliser : les normes en soutien à la réglementation européenne sur l'IA".
- Éric de La Clergerie:
  - Université d'Orléans (14 Dec 2023): "Grands modèles de langues, revisite d'une discipline ? (ou grand chamboulement ?)". Intervention dans le master SDL

11.1.5 Scientific expertise

- Alix Chagué:
  - Expert for the Fondation des arts et des sciences.
  - Member of Comité de Perfectionnement du Master Technologies Numériques Appliquées à l'Histoire de l'Ecole nationale des chartes (Paris).
- Rachel Bawden:
  - Reviewer of thematic committee for Genci projects (Reviewing projects for the allocation of computation resources).
- Benoît Sagot:
  - Member of the scientific advisory board of ERIC CLARIN.
- Djamé Seddah:
  - Expert for the Cascade funding call published for the EU project UTTER.

11.1.6 Research administration

- Benoît Sagot:
  - Member of the scientific board of Inria Paris's Comité des Projets (Inria Paris research centre's Bureau du Comité des Projets).
  - Member of the board of the Société de Linguistique de Paris (Administrateur).
- Rachel Bawden:
  - Member of the board of the Société de Linguistique de Paris (Administratrice).

11.2 Teaching - Supervision - Juries

Participants: Benoît Sagot, Rachel Bawden, Eric Villemonte De La Clergerie, Justine Cassell, Djamé Seddah, Chloé Clavel, Thibault Clerice, Floriane Chiffoleau, Alix Chague, Hugo Scheithauer, Arij Riabi, Rian Touchent, Lauriane Aufrant.

11.2.1 Teaching

- Hugo Scheithauer:
  - Master's course (M2) as part of the Master "Documentation et Humanités Numériques". Méthodologie de la recherche et préprofessionalisation (18 hours), coorganised with Françoise Dalex. École du Louvre, France.
- Arij Riabi:
  - Master's course (M1) as part of the Cycle d'ingénieur. Applied statistics project (20 hours). ENSAE.
- 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 Robin Algayres. ENS Paris-Saclay, France.

  - Computational Linguistics 1 (Spring 2023), Language and Computers, invited lecture, George Mason University, Fairfax, Virginia, USA. Introduction to NLP for historical linguistics (3 hours). George Mason University, Fairfax, Virginia, USA. 20 Apr 2023.
  - Classes for the general public as part of the Chaire annuelle "Informatique et sciences numériques". Apprendre les langues aux machines (9 hours). Collège de France. "Leçon inaugurale" (Inaugural lecture) on the 30th Nov 2023

- Alix Chagué:
  - Bachelor; Master; Doctorate's course (1er, 2e et 3e cycles) as part of the Faculté des Arts et des Sciences. HNU-6059: Humanités Numériques: Langages de Programmation (15 hours). Université de Montréal, Canada.

- Nathan Godey:
  - Bachelor's course (L2) as part of the Licence. Python. Université Pierre et Marie Curie.
  - Master's course (M2) as part of the Master SCIA. Advanced NLP (22 hours). EPITA.

- Rian Touchent:
  - Eq. Bachelor's course (1st year) as part of the Cycle ingénieur. Information systems and programming (22.5 hours). CentraleSupélec.

- Chloé Clavel:
  - Master's course (M2) as part of the Artificial Intelligence & Advanced Visual Computing Master. Sentiment analysis, speech emotion recognition, speech synthesis and conversational systems. (12 hours). Polytechnique.

- Lauriane Aufrant:
  - Master's course (M2) as part of the Master "Linguistique informatique". Information extraction and automated knowledge graph construction (12 hours). Université Paris Cité.
  - Executive education's course (Executive education) as part of the Formation continue. Traitement automatique du langage naturel (NLP) et des données textuelles (text-mining) (7 hours). Telecom Executive Education.

- Thibault Clérice, Alix Chagué and Hugo Scheithauer:
  - Workshop on HTR-United for ADHO's DH2023 conference, University of Graz, RESOWI, Graz, Austria. Workshop HTR-United: metadata, quality control and sharing process for HTR training data (4 hours). University of Graz, RESOWI, Graz, Austria. 12 Jul 2023.

11.2.2 Supervision

PhD

- José Carlos Rosales Núñez: "Machine translation for user-generated content" (1 Jun 2018–31 Jul 2023). Primary affiliation: LISN, CNRS. Supervised by Guillaume Wisniewski and Djamé Seddah. PhD defended on 3 Oct 2023.

- Robin Algayres: "Unsupervised Automatic Speech Recognition in low resource conditions" (1 Oct 2019–30 Sept 2023). Primary affiliation: CoML Inria/ENS project-team. Supervised by Emmanuel Dupoux and Benoît Sagot. PhD defended on 26 Sept 2023.

- Lionel Tadjou Tadonfouet: "Conversation Disentanglement" (1 Mar 2020–30 Sept 2023). CIFRE PhD with Orange. Supervised by Laurent Romary, Éric de La Clergerie and Fabrice Bourge (CIFRE advisor). PhD defended on 19 Oct 2023.

- Alafate Abulimiti: "The Role of Socio-conversational strategies in Task-Oriented Dialogues in the case of Peer-Tutoring Interactions: A Focus on Off-Task Talk and Hedges" (1 Nov 2023–15 Dec 2023). Secondary affiliation: ENS/PSL. Supervised by Justine Cassell and Chloé Clavel. PhD defended on 15 Dec 2023.

- Tú 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 2021–30 Sept 2023). 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 Laurent Romary and Djamé Seddah.

- Cyril Chhun: "Story Generation and Evaluation" (1 Oct 2023–present). Primary affiliation: Télécom Paris. Supervised by Chloé Clavel and Fabian Suchanek.

- 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.

- Alisa Barkar: "Interpretable textual features, public speeches, multimodal systems" (1 Nov 2022–present). Primary affiliation: Telecom Paris. Supervised by Chloé Clavel, Beatrice Biancardi and Mathieu Chollet.

- Rian Touchent: "Information Extraction on French Electronic Health Records" (1 Dec 2022–present). Supervised by Laurent Romary and Éric de La Clergerie.

- Simon Meoni: "Exploration of adaptation methods for neural models in the French clinical domain" (1 Dec 2022–present). CIFRE PhD with Arkhn. Supervised by Laurent Romary and Éric de La Clergerie.

- Wissam Antoun: "Detecting Dataset Manipulation and Weaponisation of NLP Models" (1 Mar 2023–present). Supervised by Benoît Sagot and Djamé Seddah.

- You Zuo: "Patent representation learning for innovation generation and technical trend analysis" (1 Mar 2023–present). CIFRE PhD with qatent. Supervised by Benoît Sagot, Éric de La Clergerie and Kim Gerdes (CIFRE advisor).

- Nicolas Dahan: “Evaluation of the machine translation of scientific documents” (1 Oct 2023–present). Secondary affiliation: CNRS/ISIR. Supervised by François Yvon and Rachel Bawden.
- Ziqian Peng: “Machine translation of scientific documents” (1 Oct 2023–present). Primary affiliation: CNRS/ISIR. Supervised by François Yvon and Rachel Bawden.
- Yanzhu Guo: “Language model evaluation, argument mining, computational social science” (1 Oct 2023–present). Primary affiliation: Ecole Polytechnique. Supervised by Michalis Vazirgiannis and Chloé Clavel.
- Lucie Chenain: “Speech Emotion Recognition for Huntington’s Disease risky behaviour” (1 Oct 2023–present). Primary affiliation: Paris Cité. Supervised by Anne-Catherine Bachoud Levy and Chloé Clavel.
- Lorraine Vanel: “Conversational AI, Social/emotional Dialogue Generation” (1 Oct 2023–present). Primary affiliation: Télécom Paris. CIFRE PhD with Zaion. Supervised by Chloé Clavel and Alya Yacoubi (CIFRE advisor).
- Biswesh Mohapatra: “Improving chatbot dialogue systems through collaborative grounding” (1 Oct 2023–present). Supervised by Justine Cassell and Laurent Romary.
- Chadi Helwé: “Evaluating and Improving Reasoning Abilities of Large Language Model” (1 Oct 2023–present). Primary affiliation: Télécom Paris. Supervised by Chloé Clavel and Fabian Suchanek.
- Hugo Scheithauer: “Acquisition, integration and redistribution of structured data in GLAMs: harmonising practices” (1 Nov 2023–present). Supervised by Laurent Romary.
- Armel Zebaze: “Analogy for Multilingual Natural Language Processing” (1 Nov 2023–present). Supervised by Benoît Sagot and Rachel Bawden.
- Rasul Dent: “Large-scale language identification (numerous languages, massive data, distinction between closely related varieties) with a focus on the languages of France and French-based creoles.” (1 Nov 2023–present). Supervised by Benoît Sagot, Thibault Clérice and Pedro Ortiz.

Interns

- Samuel Scalbert: “Reflections and development of techniques for implementing and displaying (new) digital editions on the DiScholEd publishing platform” (3 Apr 2023–31 Jul 2023). Supervised by Floriane Chiffoleau.
- Sarah Bénière: “Processing for the digital edition of the EHRI collection, in particular within the DiScholEd publishing platform” (3 Apr 2023–31 Jul 2023). Supervised by Floriane Chiffoleau.
- Leo Labat: “Extraction of knowledge graphs by combining and adapting tools for a number of NLP tasks, such as named entity recognition, named entity linking, coreference resolution, relation extraction, relation clustering, document-level event extraction and slot filling.” (25 Sept 2023–present). Supervised by Lauriane Aufrant.
- Ilyas Lebleu: “Exploring the reasoning capabilities of transformers in language models based on the design of families of prompts” (4 Dec 2023–present). Supervised by Éric de La Clergerie.

Engineers

- Tanti Kristanti Nugraha: “Entity fishing for scholarly literature in the humanities” (1 Nov 2017–30 Apr 2023). Supervised by Laurent Romary.
- Julien Abadji: “Large-scale multilingual corpus development and extension (OSCAR corpus)” (1 Apr 2021–31 Aug 2023). Supervised by Benoît Sagot.
- Hugo Scheithauer: “Training segmentation models for sales catalogues with GROBID” (1 Oct 2021–31 Oct 2023). Supervised by Laurent Romary.

- Rua Ismail: “Language identification for large-scale multilingual raw corpus development” (17 Jan 2022–31 Mar 2023). Supervised by Benoît Sagot.
- Wissam Antoun: “Language models for languages displaying high variabilty, in particular Arabic dialects used on social media” (1 Mar 2022–28 Feb 2023). Supervised by Djamé Seddah and Benoît Sagot.
- Anna Chepaikina: “Automatic generation of oenological descriptions” (31 Mar 2022–30 Sept 2023). 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 Sept 2022–31 Aug 2023). Supervised by Éric de La Clergerie.
- Menel Mahamdi: “Data set annotation, synthetic data generation, conversational data sets” (1 Sept 2023–present). Supervised by Djamé Seddah.
- Niyati Bafna: “Linguistically inspired language models for closely related languages” (1 Oct 2022–2 Jun 2023). Secondary affiliation: DFKI. Supervised by Benoît Sagot, Rachel Bawden, Josef van Genabith and Cristina España-Bonet.
- Mouilleron Virginie: “Correction and annotation of the Alien vs Predator dataset, Prompt Tuning and Data extraction from LLMs” (1 Dec 2022–present). Supervised by Djamé Seddah.
- Deepak Yadav: “Exploration of neural classifiers for very low resource languages and domains” (1 Apr 2023–31 May 2023). Supervised by Djamé Seddah.
- Seth Aycock: “Domain adaptation for neural machine translation in low-resource settings” (1 Aug 2023–6 Nov 2023). Supervised by Rachel Bawden.
- Juliette Janès: “Recovery, encoding, maintenance, and publication of textual data on French and other languages of France produced within the framework of the DEFI COLaF” (1 Oct 2023–present). Supervised by Benoît Sagot and Thibault Clérice.
- Sarah Bénière: “Automatic analysis of digitized sales catalogs” (1 Oct 2023–present). Supervised by Laurent Romary.
- Samuel Scalbert: “Detection of software in HAL articles using GROBID and Softcite in the context of the GrapOS project.” (1 Oct 2023–present). Supervised by Laurent Romary.
- Marius Le Chapelier: “Developing the SARA (Socially Aware Robot Assistant) dialogue system to be able to build social bonds (rapport) with users in order to improve performance.” (1 Nov 2023–present). Supervised by Justine Cassell.
- Oriane Nédey: “Data collection and translation models for a regional language of France.” (1 Dec 2023–present). Supervised by Rachel Bawden, Thibault Clérice and Benoît Sagot.
- Cecilia Graiff: “Named entity disambiguation for the National Archives of France in the context of the NER4Archives project.” (1 Dec 2023–present). Supervised by Laurent Romary.

Postdocs

- Aina Garí Soler: “Word Meaning Representation in Neural Language Models: Lexical Polysemy and Semantic Relationships” (6 Sept 2021–present). Primary affiliation: Télécom-Paris. Supervised by Chloé Clavel.
- Emer Gilmartin: “Collaboration with researchers in Korea to understand and model the effects of interlocutor personality on dialogue, and the effects of conversational behaviors on personality. This is leading to a new model of ‘interpersonality’, how personality related behaviours of each participant in a conversation affect the conversation as a whole and, vice-versa, how conversational behaviors affect perceptions of personality.” (1 Oct 2022–present). Supervised by Justine Cassell.

- José Carlos Rosales Núñez: "Radicalisation detection, robust UGC processing and machine translation." (1 Aug 2023–present). Supervised by Djamé Seddah.

11.2.3 Juries

PhD

- Rachel Bawden
  - Member of the PhD committee as examiner for Lorenzo Lupo at Université Grenoble Alpes on 28 Mar 2023. Title: *Challenges and Remedies for Context-Aware Neural Machine Translation.*
- Benoît Sagot
  - Member of the PhD committee as reviewer for David Adelani at Saarland University on 27 Jun 2023. Title: *Natural language processing for African languages.*
  - Member of the PhD committee as reviewer and president for Liam Cripwell at Université de Lorraine, CNRS/LORIA on 10 Nov 2023. Title: *Controllable and document-level simplification systems.*
  - Member of the PhD committee as co-director for Robin Algayres at ENS on 26 Sept 2023. Title: *Unsupervised word discovery in speech data.*
  - Member of the PhD committee as examiner for Lucence Ing at École Nationale des Chartes - PSL on 29 Sept 2023. Title: *L'obsolescence lexicale en français médiéval. Philologie et linguistique computationnelles sur le Lancelot en prose.*
  - Member of the PhD committee for Hugo Boulanger at Universtité Paris-Saclay on 30 Mar 2023. Title: *Data Augmentation and Generation for Natural Language Processing.*
- Justine Cassell
  - Member of the PhD committee as examiner for Juan Vazquez at Université Grenoble Alpes on 27 Nov 2023. Title: *Multimodal Transformers for Emotion Recognition.*
- Chloé Clavel
  - Member of the PhD committee as president for Juliette Faille at LORIA on 17 Nov 2023. Title: *Data-Based Natural Language Generation: Evaluation and Explainability.*
  - Member of the PhD committee as reviewer for Lionel Tadonfouet Tadjou at Inria Paris on 19 Oct 2023. Title: *Constitution de fils de discussion cohérents à partir de conversations issues d'outils professionnels de communication et de collaboration.*
  - Member of the PhD committee as reviewer for Liu Yang at ISIR on 8 Dec 2023. Title: *Modelling interruptions in human-agent interaction.*
  - Member of the PhD committee as director for Marc Hulcelle at Telecom-Paris on 22 Dec 2023. Title: *Automatic analysis of trust over the course of a human-robot interaction using multimodal features and recurrent neural architectures.*
  - Member of the PhD committee as co-director for Alafate Abulimiti at Inria on 14 Dec 2023. Title: *The Role of Socio-conversational strategies in Task-Oriented Dialogues in the case of Peer-Tutoring Interactions: A Focus on Off-Task Talk and Hedges.*
- Éric de La Clergerie
  - Member of the PhD committee as co-director for Lionel Tadonfouet Tadjou at Inria Paris on 19 Oct 2023. Title: *Constitution de fils de discussion cohérents à partir de conversations issues d'outils professionnels de communication et de collaboration.*

Master

- Benoît Sagot
  - Member of the Master's committee as tutor for Ali Charara at MVA on 5 Oct 2023. Title: *Génération automatique de "briefs" pour la vérification d'informations.*
  - Member of the Master's committee as tutor for Mohamed Mohamed Abdallahi at MVA & EDF on 3 Oct 2023. Title: *Key Information Extraction.*
  - Member of the Master's committee as tutor for Zeki Topçu at MVA & Université Paris Cité on 15 Sept 2023. Title: *Exploration of press articles related to Covid-19 at the European level within the Covid-19 Museum.*
  - Member of the Master's committee as tutor for Armel Randy Zebaze at MVA & Hugging Face on 13 Sept 2023. Title: *Instruction fine-tuning of Code Large Language Models.*
- Floriane Chiffoleau
  - Member of the Master's committee as examiner for Samuel Scalbert at Ecole nationale des chartes, Paris, France on 25 Sept 2023. Title: *Les CMS et le low-code au service des humanités numériques: l'exemple de DiScholEd, une application TEI Publisher.*
  - Member of the Master's committee as examiner for Sarah Bénière at Ecole nationale des chartes, Paris, France on 25 Sept 2023. Title: *De l'encodage à la publication: Les éditions en ligne de la European Holocaust Research Infrastructure et leur chaîne éditoriale.*
- Hugo Scheithauer
  - Member of the Master's committee as examiner for Miranda Cerino Galindo at École du Louvre, Paris, France on 14 Sept 2023. Title: *Créer une documentation pour l'atelier de la Chalcographie du Louvre dans les Ateliers d'Art de la Réunion des musées nationaux – Grand Palais: Contenus, Acteurs, Outils et Objectifs.*
  - Member of the Master's committee as examiner for Christophe Carini-Siguret at École du Louvre, Paris, France on 19 Sept 2023. Title: *Cartographier l'incertitude: Des référentiels d'indexation à la visualisation des données, à partir d'un corpus photographique du musée du quai Branly - Jacques Chirac.*
  - Member of the Master's committee as examiner for Louise Glodt-Chauchoy at École du Louvre, Paris, France on 20 Sept 2023. Title: *De l'inventaire à la valorisation de dessins botaniques: processus d'appropriation patrimoniale de dessins inédits des sœurs Vesque par la documentation au Muséum national d'Histoire naturelle.*
  - Member of the Master's committee as examiner for Emma Fourgeaud at École du Louvre, Paris, France on 2 Oct 2023. Title: *École du Louvre, Paris, France.*
  - Member of the Master's committee as examiner for Pauline Heleine at École du Louvre, Paris, France on 25 Oct 2023. Title: *Mémoires de restauration: les dossiers de restauration des objets mobiliers protégés au titre des Monuments historiques. Nature, cadre législatif et diffusion. La spécificité de la Médiathèque du Patrimoine et de la Photographie, dépositaire des archives des Monuments Historiques..*
  - Member of the Master's committee as examiner for Manon Tauziet at École du Louvre, Paris, France on 28 Jun 2023. Title: *La place de la documentation et de la bibliothèque dans un projet de centre de recherche. Étude de cas au Musée national Picasso Paris pour le projet d'un Centre d'étude Picasso..*
- Tú Anh Nguyen
  - Member of the Master's committee as co-supervisor for Paul Fotso Kaptue at MVA on 7 Nov 2023. Title: *Spoken Language Modeling with Soft Speech Units.*

CSD

- Rachel Bawden
  - Member of the CSD committee for Antoine Yang at Inria Paris on 27 Jun 2023. Title: *Multimodal video representation with cross-modal learning*.
  - Member of the CSD committee for Tom Calamai at Inria on 30 Jun 2023. Title: *Détection automatique d'argument fallacieux*.
- Benoît Sagot
  - Member of the CSD committee for Nathanaël Beau at Université Paris Cité & onespace on 21 Sept 2023. Title: *Génération de code métier Python à partir d'une description en langage naturel*.
  - Member of the CSD committee for Gautier Izacard at ENS & META AI on 21 Dec 2023. Title: *Improving natural language understanding by learning better language models*.
- Éric de La Clergerie
  - Member of the CSD committee for Shrey Mishra at ENS on 6 Jul 2023. Title: *Extraction of Proofs and Theorems in the Scientific Literature*.
  - Member of the CSD committee for Cyril Bruneau at Université Paris Nanterre on 3 Jul 2023. Title: *Transmettre des valeurs à l'école: développement d'un outillage informatique appliqué aux manuels scolaires d'histoire (1870-2020)*.
  - Member of the CSD committee for Maya Sahraoui at Sorbonne Université, Paris, France on 13 Oct 2023. 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.
  - Member of the CR-TH hiring committee at Inria (National). CR recrutement for applicants with disabilities.
- Benoît Sagot:
  - Member of the AER hiring committee at Inria (Sophia Centre).
- Thibault Clérice:
  - External Member of the Engineer hiring committee at PSL (École nationale des Chartes). ERC Engineers.

11.3 Popularization

Participants: Benoît Sagot, Rachel Bawden, Chloé Clavel, Sarah Bénière, Samuel Scalbert, Floriane Chiffoleau, Alix Chagué, Lydia Nishimwe, Hugo Scheithauer.

11.3.1 Articles and contents

Authored article

- Sarah Bénière
  - for the Digital Intellectuals Blog (Outreach article), "Writing an ODD for the EHRI Online Editions — Preparatory Work". Online, 5 Jun 2023.
  - for the Digital Intellectuals Blog (Outreach article), "Writing an ODD for the EHRI Online Editions — Specifications and Documentation". Online, 28 Jul 2023.
- Samuel Scalbert
  - for the Digital Intellectuals Blog (Outreach article), "Overcoming challenges in DiScholEd's development: a journey of problem-solving and design enhancements". Online, 29 Jun 2023.

Article with citation

- Rachel Bawden
  - cited in an article by www.letudiant.fr (Media article), "Faut-il bannir ChatGPT du monde de la recherche ?". Online, 3 Apr 2023.

Media interview

- Benoît Sagot
  - interviewed as part of France TV (France 2, Journal de 20h) (National news), "Intelligence artificielle – Des robots rédigent des dissertations". TV + Online replay, 4 Jan 2023.
  - interviewed as part of Radio France (France Inter, Interception) (Radio programme), "Les deux visages de l'intelligence artificielle". Radio broadcast + Online, 12 Mar 2023.

11.3.2 Education

- Rachel Bawden:
  - participated in the *Rendez-Vous des Jeunes Mathématiciennes et Informaticiennes Inria 2023* (Presentation on NLP to high school girls), "Traitement Automatique des Langues (TAL). Que se cache-t-il derrière les modèles de langue ?" Inria Paris, 24 Oct 2023.
- Benoît Sagot:
  - gave a talk at Colloque « L'IA et ses défis » (Colloque ouvert au public du Campus de l'Innovation pour les Lycées), "L'apprentissage profond, au cœur de l'IA moderne". Collège de France, Paris, 28 Sept 2023.
- Chloé Clavel:
  - gave a talk at Super Demain, Le grand laboratoire de l'éducation aux médias numériques, www.superdemain.fr (Seminar), "Robot et émotion : la reconnaissance des manifestations émotionnelles". Lyon, 24 Nov 2023.
- Floriane Chiffoleau, Hugo Scheithauer and Lydia Nishimwe:
  - were mentors for work shadowing (Seconday "3ème" school students), Inria Paris, 19 Dec 2023.
- Alix Chagué and Lydia Nishimwe:
  - ran a workshop at the *Rendez-vous des Jeunes Mathématiciennes et Informaticiennes*, Inria Paris, 24-25 Oct 2023.

11.3.3 Interventions

- Benoît Sagot:
  - gave a talk at PRAIRIE evening (Professional workshop), "A few words on BLOOM (and ChatGPT)". Inria Paris, 17 Jan 2023.
  - participated in FD3, co-organised by PRAIRIE and France Digitale (Professional workshop), "Round table on "Is Generative AI a revolution and what to expect next?"". Station F, 29 Mar 2023.
  - gave a talk at Dauphine Digital Days (Professional workshop), "Quelles données pour entraîner les grands modèles de langue ?", Opening keynote for the round table "Quels régimes de régulation des données pour entraîner les intelligences artificielles ?". Université Paris Dauphine, 20 Nov 2023.
  - gave a talk at Inria Alumni Meeting (Professional workshop), "Une brève introduction à ChatGPT". Inria Paris, 23 Nov 2023.
- Hugo Scheithauer (with EHRI):
  - participated in Yad Vashem (EHRI General Partneer Meeting), Jerusalem, Israël, 29 May 2023.

12 Scientific production

12.1 Major publications

[1] A. Chagué and T. Clérice. '"I'm here to fight for ground truth": HTR-United, a solution towards a common for HTR training data'. In: Digital Humanities 2023: Collaboration as Opportunity. Graz, Austria, 2023. URL: https://inria.hal.science/hal-04094233.

[2] M. Futeral, C. Schmid, I. Laptev, B. Sagot and R. Bawden. 'Tackling Ambiguity with Images: Improved Multimodal Machine Translation and Contrastive Evaluation'. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vol. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada, 3rd July 2023, pp. 5394–5413. DOI: 10.18653/v1/2023.acl-long.295. URL: https://inria.hal.science/hal-03977982.

[3] 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.

[4] 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://inria.hal.science/hal-03479757.

[5] 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.

[6] L. Martin, A. Fan, E. 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://inria.hal.science/hal-03834719.

[7] 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.

[8] 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, 6th June 2021. URL: https://inria.hal.science/hal-03251105.

[9] 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.

[10] 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.* 20th Nov. 2023. URL: https://inria.hal.science/hal-03850124.

[11] 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.

12.2 Publications of the year

International journals

[12] T. Clérice. 'You Actually Look Twice At it (YALTAi): using an object detection approach instead of region segmentation within the Kraken engine'. In: *Journal of Data Mining and Digital Humanities* Historical Documents and... (26th Dec. 2023). DOI: 10.46298/jdmdh.9806. URL: https://enc.hal.science/hal-03723208.

[13] T. Clérice, M. Vlachou-Efstathiou and A. Chagué. 'CREMMA Medii Aevi: Literary manuscript text recognition in Latin'. In: *Journal of Open Humanities Data* 9 (12th Apr. 2023), p. 4. DOI: 10.5334/johd.97. URL: https://enc.hal.science/hal-03828353.

[14] 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: *Transactions of the Association for Computational Linguistics* 11 (14th Mar. 2023), pp. 250–266. DOI: 10.1162/tacl_a_00545. URL: https://inria.hal.science/hal-03985368.

National journals

[15] S. Gabay, P. Gambette, R. Bawden and B. Sagot. 'Old or Modern? Towards a Computational Graphematic Analysis of 17th Century French Texts'. In: *Linx* 85 (20th Feb. 2023). DOI: 10.4000/linx.9346. URL: https://hal.science/hal-04110764.

Invited conferences

[16] H. Scheithauer. 'DataCatalogue : Un projet pour la restructuration automatique de catalogues de vente'. In: Traitements automatiques pour les humanités numériques - corpus d'histoire de l'art, d'enseignement, d'urbanisme. Nanterre, France, 15th May 2023. URL: https://inria.hal.science/hal-04265312.

International peer-reviewed conferences

[17] R. Algayres, Y. Adi, T. A. Nguyen, J. Copet, G. Synnaeve, B. Sagot and E. Dupoux. 'Generative Spoken Language Model based on continuous word-sized audio tokens'. In: The 2023 Conference on Empirical Methods in Natural Language Processing. Singapore, Singapore, 6th Dec. 2023. URL: https://inria.hal.science/hal-04402373.

[18] R. Algayres, P. Diego-Simon, B. Sagot and E. Dupoux. 'XLS-R fine-tuning on noisy word boundaries for unsupervised speech segmentation into words'. In: Findings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Singapore, Singapore, 6th Dec. 2023. URL: https://inria.hal.science/hal-04398496.

[19] W. Antoun, B. Sagot and D. Seddah. 'Data-Efficient French Language Modeling with CamemBERTa'. In: Findings of the Association for Computational Linguistics: ACL 2023. Toronto, Canada, 9th July 2023, pp. 5174–5185. DOI: 10.18653/v1/2023.findings-acl.320. URL: https://inria.hal.science/hal-03963729.

[20] R. Bawden and B. Sagot. 'RoCS-MT: Robustness Challenge Set for Machine Translation'. In: *Proceedings of the Eighth Conference on Machine Translation*. WMT23 - Eighth Conference on Machine Translation. Proceedings of the Eighth Conference on Machine Translation. Singapore, Singapore, 2023, pp. 198–216. URL: https://hal.science/hal-04300824.

[21] R. Bawden and F. Yvon. 'Investigating the Translation Performance of a Large Multilingual Language Model: the Case of BLOOM'. In: *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*. EAMT 2023 - 24th Annual Conference of the European Association for Machine Translation. Tampere, Finland, 3rd Mar. 2023. DOI: 10.48550/ARXIV.2303.01911. URL: https://inria.hal.science/hal-04015863.

[22] J.-B. Camps, N. Baumard, P.-C. Langlais, O. Morin, T. Clérice and J. Norindr. 'Make Love or War? Monitoring the Thematic Evolution of Medieval French Narratives'. In: *Proceedings of the Computational Humanities Research Conference 2023*. Computational Humanities Research (CHR 2023). CEUR Workshop Proceedings. Paris, France, 2023. URL: https://enc.hal.science/hal-04250657.

[23] G. Castillo-López, A. Riabi and D. Seddah. 'Analyzing Zero-Shot transfer Scenarios across Spanish variants for Hate Speech Detection'. In: Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023). Dubrovnik, Croatia: Association for Computational Linguistics, 5th May 2023, pp. 1–13. DOI: 10.18653/v1/2023.vardial-1.1. URL: https://inria.hal.science/hal-04243810.

[24] A. Chagué and T. Clérice. '"I'm here to fight for ground truth": HTR-United, a solution towards a common for HTR training data'. In: Digital Humanities 2023: Collaboration as Opportunity. Graz, Austria, 2023. URL: https://inria.hal.science/hal-04094233.

[25] A. Chagué, T. Clérice, J. Norindr, M. Humeau, B. Davoury, E. V. Kote, A. Mazoue, M. Faure and S. Doat. 'Manu McFrench, from zero to hero: impact of using a generic handwriting recognition model for smaller datasets'. In: Digital Humanities 2023: Collaboration as Opportunity. Graz, Austria, Nov. 2022. URL: https://inria.hal.science/hal-04094241.

[26] T. Clérice, A. Chagué and H. Scheithauer. 'Workshop HTR-United: metadata, quality control and sharing process for HTR training data'. In: DH 2023 - Digital Humanities Conference: Collaboration as Opportunity. Graz, Austria, Nov. 2022. URL: https://inria.hal.science/hal-04094235.

[27] T. Clérice and A. Glaise. 'Twenty-One* Pseudo-Chrysostoms and more: authorship verification in the patristic world'. In: *Proceedings of the Computational Humanities Research Conference 2023*. Computational Humanities Research Conference 2023. Proceedings of the Computational Humanities Research Conference 2022. Paris, France, 2023. URL: https://inria.hal.science/hal-04211176.

[28] P.-A. Duquenne, H. Gong, N. Dong, J. Du, A. Lee, V. Goswami, C. Wang, J. Pino, B. Sagot and H. Schwenk. 'SpeechMatrix: A Large-Scale Mined Corpus of Multilingual Speech-to-Speech Translations'. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. ACL 2023 - 61st Annual Meeting of the Association for Computational Linguistics. Toronto, Canada, 9th Aug. 2023. DOI: 10.18653/v1/2023.acl-long.899. URL: https://inria.hal.science/hal-04264040.

[29] P.-A. Duquenne, H. Schwenk and B. Sagot. 'Modular Speech-to-Text Translation for Zero-Shot Cross-Modal Transfer'. In: *Proceedings of INTERSPEECH 2023*. INTERSPEECH 2023. Dublin, Ireland, 20th Aug. 2023. DOI: 10.21437/Interspeech.2023-2484. URL: https://hal.science/hal-04264023.

[30] A. Elkahky, W.-N. Hsu, P. Tomasello, T. A. Nguyen, R. Algayres, Y. Adi, J. Copet, E. Dupoux and A. Mohamed. 'Do Coarser Units Benefit Cluster Prediction-Based Speech Pre-Training?' In: 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023). ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Ixia-Ialyssos, Greece: IEEE, 2023. DOI: 10.1109/ICASSP49357.2023.10096788. URL: https://cnrs.hal.science/hal-04208427.

[31] M. Futeral, C. Schmid, I. Laptev, B. Sagot and R. Bawden. 'Tackling Ambiguity with Images: Improved Multimodal Machine Translation and Contrastive Evaluation'. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vol. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada, 3rd July 2023, pp. 5394–5413. DOI: 10.18653/v1/2023.acl-long.295. URL: https://inria.hal.science/hal-03977982.

[32] I. Gat, F. Kreuk, T. Anh Nguyen, A. Lee, J. Copet, G. Synnaeve, E. Dupoux and Y. Adi. 'Augmentation Invariant Discrete Representation for Generative Spoken Language Modeling'. In: 20th International Conference on Spoken Language Translation (IWSLT 2023). Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 465–477. DOI: 10.18653/v1/2023.iwslt-1.46. URL: https://cnrs.hal.science/hal-04208443.

[33] 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.

[34] T. Kocmi, E. Avramidis, R. Bawden, O. Bojar, A. Dvorkovich, C. Federmann, M. Fishel, M. Freitag, T. Gowda, R. Grundkiewicz, B. Haddow, P. Koehn, B. Marie, C. Monz, M. Morishita, K. Murray, M. Nagata, T. Nakazawa, M. Popel, M. Popović, M. Shmatova and J. Suzuki. 'Findings of the 2023 Conference on Machine Translation (WMT23): LLMs Are Here But Not Quite There Yet'. In: *Proceedings of the Eighth Conference on Machine Translation*. WMT23 - Eighth Conference on Machine Translation. Singapore, Singapore, 2023, pp. 198–216. URL: https://hal.science/hal-04300702.

[35] S. Meoni, T. Ryffel and E. Villemonte de La Clergerie. 'Large Language Models as Instructors: A Study on Multilingual Clinical Entity Extraction'. In: *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*. The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks. Toronto, Canada: Association for Computational Linguistics, 13th July 2023, pp. 178–190. DOI: 10.18653/v1/2023.bionlp-1.15. URL: https://hal.science/hal-04394012.

[36] M. Neves, A. Jimeno Yepes, A. Névéol, R. Bawden, G. M. D. Nunzio, R. Roller, P. Thomas, F. Vezzani, M. V. Navarro, L. Yeganova, D. Wiemann and C. Grozea. 'Findings of the WMT 2023 Biomedical Translation Shared Task: Evaluation of ChatGPT 3.5 as a Comparison System'. In: *Proceedings of the Eighth Conference on Machine Translation*. WMT23 - Eighth Conference on Machine Translation. Singapore, Singapore, 2023, pp. 43–54. URL: https://hal.science/hal-04300785.

[37] T. A. Nguyen, W.-N. Hsu, A. d'Avirro, B. Shi, I. Gat, M. Fazel-Zarani, T. Remez, J. Copet, G. Synnaeve, M. Hassid, F. Kreuk, Y. Adi and E. Dupoux. 'Expresso: A Benchmark and Analysis of Discrete Expressive Speech Resynthesis'. In: INTERSPEECH 2023 - 24th Annual Conference of the International Speech Communication Association. Dublin, Ireland: ISCA, 20th Aug. 2023, pp. 4823–4827. DOI: 10.21437/Interspeech.2023-1905. URL: https://cnrs.hal.science/hal-04208441.

[38] A. Riabi, M. Mahamdi and D. Seddah. 'Enriching the NArabizi Treebank: A Multifaceted Approach to Supporting an Under-Resourced Language'. In: Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII). Toronto, Canada: Association for Computational Linguistics, 13th July 2023, pp. 266–278. DOI: 10.18653/v1/2023.law-1.26. URL: https://inria.hal.science/hal-04243832.

[39] J. C. Rosales Núñez, D. Seddah and G. Wisniewski. 'Multi-way Variational NMT for UGC: Improving Robustness in Zero-shot Scenarios via Mixture Density Networks'. In: NoDaLiDa 2023 - 24th Nordic Conference on Computational Linguistics. Torshavn, Faroe Islands, 22nd May 2023. URL: https://hal.science/hal-04384748.

[40] S. Sannigrahi and R. Bawden. 'Investigating Lexical Sharing in Multilingual Machine Translation for Indian Languages'. In: *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*. EAMT 2023 - 24th Annual Conference of the European Association for Machine Translation. Tampere, Finland, 4th May 2023. URL: https://hal.science/hal-04093966.

[41] V. Taillandier, D. Hupkes, B. Sagot, E. Dupoux and P. Michel. 'Neural Agents Struggle to Take Turns in Bidirectional Emergent Communication'. In: ICLR 2023 - 11th International Conference on Learning Representation. Kigali, Rwanda, 1st May 2023. URL: https://inria.hal.science/hal-04264045.

National peer-reviewed Conferences

[42] W. Antoun, V. Mouilleron, B. Sagot and D. Seddah. 'Towards a Robust Detection of Language Model-Generated Text: Is ChatGPT that easy to detect?' In: *Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : travaux de recherche originaux – articles longs*. 18e Conférence en Recherche d'Information et Applications – 16e Rencontres Jeunes Chercheurs en RI – 30e Conférence sur le Traitement Automatique des Langues Naturelles – 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues. Paris, France: ATALA, 2023, pp. 14–27. URL: https://hal.science/hal-04130146.

[43] N. Bafna, C. España-Bonet, J. van Genabith, B. Sagot and R. Bawden. 'Cross-lingual Strategies for Low-resource Language Modeling: A Study on Five Indic Dialects'. In: *Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : travaux de recherche originaux – articles longs*. 18e Conférence en Recherche d'Information et Applications – 16e Rencontres Jeunes Chercheurs en RI – 30e Conférence sur le Traitement Automatique des Langues Naturelles – 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues. Paris, France: ATALA, 2023, pp. 28–42. URL: https://hal.science/hal-04130175.

[44] M. Bénard, A. Mestivier, N. Kubler, L. Zhu, R. Bawden, E. De La Clergerie, L. Romary, M. Huguin, J.-F. Nominé, Z. Peng and F. Yvon. 'MaTOS : Machine Translation for Open Science'. In: *Actes de CORIA-TALN 2023. Actes de l'atelier "Analyse et Recherche de Textes Scientifiques" (ARTS)@TALN 2023*. 18e Conférence en Recherche d'Information et Applications – 16e Rencontres Jeunes Chercheurs en RI – 30e Conférence sur le Traitement Automatique des Langues Naturelles – 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues. Paris, France: ATALA, 2023, pp. 8–15. URL: https://hal.science/hal-04131594.

[45] S. Meoni, T. Ryffel and E. De La Clergerie. 'Annotation d'entités cliniques en utilisant les Larges Modèles de Langue'. In: *Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : travaux de recherche originaux – articles longs*. 18e Conférence en Recherche d'Information et Applications – 16e Rencontres Jeunes Chercheurs en RI – 30e Conférence sur le Traitement Automatique des Langues Naturelles – 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues. Paris, France: ATALA, 2023, pp. 190–203. URL: https://hal.science/hal-04130197.

[46] S. Meoni, R. Touchent and E. De La Clergerie. 'Passe ta pharma d'abord !' In: *Actes de CORIA-TALN 2023. Actes du Défi Fouille de Textes@TALN2023*. 18e Conférence en Recherche d'Information et Applications – 16e Rencontres Jeunes Chercheurs en RI – 30e Conférence sur le Traitement Automatique des Langues Naturelles – 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues. Paris, France: ATALA, 2023, pp. 68–76. URL: https://hal.science/hal-04131583.

[47] L. Nishimwe. 'Lexical normalisation of user-generated content on social media'. In: *Actes de CORIA-TALN 2023. Actes des 16e Rencontres Jeunes Chercheurs en RI (RJCRI) et 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL)*. 18e Conférence en Recherche d'Information et Applications – 16e Rencontres Jeunes Chercheurs en RI – 30e Conférence sur le Traitement Automatique des Langues Naturelles – 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues. Vol. Actes des 16e Rencontres Jeunes Chercheurs en RI (RJCRI) et 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL). Paris, France: ATALA, 2023, pp. 160–183. URL: https://hal.science/hal-04130239.

[48] L. Tadonfouet Tadjou, E. De La Clergerie, F. Bourge and T. Marie. 'Constitution de sous-fils de conversations d'emails'. In: *Actes de CORIA-TALN 2023. Actes de la 18e Conférence en Recherche d'Information et Applications (CORIA)*. 18e Conférence en Recherche d'Information et Applications – 16e Rencontres Jeunes Chercheurs en RI – 30e Conférence sur le Traitement Automatique des Langues Naturelles – 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues. Paris, France: ATALA, 2023, pp. 157–171. URL: https://hal.science/hal-04131559.

[49] R. Touchent, L. Romary and E. De La Clergerie. 'CamemBERT-bio : Un modèle de langue français savoureux et meilleur pour la santé'. In: *Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : travaux de recherche originaux – articles longs*. 18e Conférence en Recherche d'Information et Applications
16e Rencontres Jeunes Chercheurs en RI
30e Conférence sur le Traitement Automatique des Langues Naturelles
25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues. Paris, France: ATALA, 2023, pp. 323–334. URL: https://hal.science/hal-04130187.

[50] Y. Zuo, K. Gerdes, H. Mouzoun, S. Ghamri Doudane and B. Sagot. 'Exploring Data-Centric Strategies for French Patent Classification: A Baseline and Comparisons'. In: *Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : travaux de recherche originaux – articles longs*. 18e Conférence en Recherche d'Information et Applications – 16e Rencontres Jeunes Chercheurs en RI – 30e Conférence sur le Traitement Automatique des Langues Naturelles – 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues. Paris, France: ATALA, 2023, pp. 349–365. URL: https://hal.science/hal-04130188.

Conferences without proceedings

[51] H. Scheithauer, S. Bénière, J.-P. Moreux and L. Romary. 'DataCatalogue : rétro-structuration automatique des catalogues de vente'. In: Webinaire Culture-Inria. Paris, France, 29th Nov. 2023. URL: https://hal.science/hal-04360229.

[52] C. Vidal-Gorène, J.-B. Camps and T. Clérice. 'Synthetic lines from historical manuscripts: an experiment using GAN and style transfer'. In: Visual Processing of Digital Manuscripts: Workflows, Pipelines, Best Practices. ICIAP 2023 Workshops. ICIAP 2023. Udine, Italy, 11th Sept. 2023. URL: https://inria.hal.science/hal-04178550.

Doctoral dissertations and habilitation theses

[53] J. Rosales Núñez. 'Machine Translation of User-Generated Contents : an Evaluation of Neural Translation Systems under Zero-shot Conditions'. Université Paris-Saclay, 3rd Oct. 2023. URL: https://theses.hal.science/tel-04301123.

Reports & preprints

[54] W. Antoun, D. Seddah and B. Sagot. *From Text to Source: Results in Detecting Large Language Model-Generated Content*. 23rd Sept. 2023. URL: https://inria.hal.science/hal-04264050.

[55] N. Bafna, C. España-Bonet, J. van Genabith, B. Sagot and R. Bawden. *A Simple Method for Unsupervised Bilingual Lexicon Induction for Data-Imbalanced, Closely Related Language Pairs*. 23rd May 2023. URL: https://inria.hal.science/hal-04264052.

[56] B. Biancardi, M. Chollet and C. Clavel. *Introducing the 3MT_French Dataset to Investigate the Timing of Public Speaking Judgements*. 4th Oct. 2022. DOI: 10.21203/rs.3.rs-2122814/v1. URL: https://hal.science/hal-04366763.

[57] A. Chagué and F. Chiffoleau. *ATR: What can eScriptorium do for you?* 8th Sept. 2023. URL: https://hal.science/hal-04247827.

[58] A. Chagué and F. Chiffoleau. *What can you do next? Choice of output and reuse of your transcription*. 8th Sept. 2023. URL: https://hal.science/hal-04247966.

[59] A. Chagué and T. Clérice. *Données ouvertes, données propres, et autres vies : Testaments de Poilus et CREMMA*. Dec. 2023. URL: https://inria.hal.science/hal-04347066.

[60] A. Chagué and H. Souvay. *Image Acquisition and Layout Analysis*. Sept. 2023. URL: https://inria.hal.science/hal-04254223.

[61] F. Chiffoleau. *TEI Publisher, a platform for sustainable digital editions*. 21st Sept. 2023. URL: https://hal.science/hal-04247980.

[62] T. Clérice. *Detecting Sexual Content at the Sentence Level in First Millennium Latin Texts*. 22nd Sept. 2023. URL: https://inria.hal.science/hal-04214375.

[63] P.-A. Duquenne, H. Schwenk and B. Sagot. *SONAR: Sentence-Level Multimodal and Language-Agnostic Representations*. 29th Oct. 2023. URL: https://inria.hal.science/hal-04264028.

[64] N. Godey, E. Villemonte de La Clergerie and B. Sagot. *Headless Language Models: Learning without Predicting with Contrastive Weight Tying*. 15th Sept. 2023. URL: https://inria.hal.science/hal-04264051.

[65] N. Godey, E. Villemonte de La Clergerie and B. Sagot. *Is Anisotropy Inherent to Transformers?* 13th June 2023. URL: https://inria.hal.science/hal-04264026.

[66] 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://cnrs.hal.science/hal-03992291.

[67] Y. Parmentier, S. Pogodalla, R. Bawden, M. Labeau and I. Eshkol-Taravella. *Procédure de diffusion des publications de l'ATALA sur les archives ouvertes*. ATALA, Sept. 2023, p. 17. URL: https://hal.science/hal-04258177.

[68] A. Pinche, T. Clérice, A. Chagué, J.-B. Camps, M. Vlachou-Efstathiou, M. Gille Levenson, O. Brisville-Fertin, F. Boschetti, F. Fischer, M. Gervers, A. Boutreux, A. Manton, S. Gabay, P. O'Connor, W. Haverals, M. Kestemont and C. Vandyck. *CATMuS-Medieval: Consistent Approaches to Transcribing ManuScripts: A generalized set of guidelines and models for Latin scripts from Middle Ages (8th–16th century)*. Dec. 2023. URL: https://inria.hal.science/hal-04346939.

[69] C. A. Romein, T. Hodel, F. Gordijn, J. Zundert, A. Chagué, M. V. Lange, H. S. Jensen, A. Stauder, J. Purcell, M. Terras et al. *Exploring Data Provenance in Handwritten Text Recognition Infrastructure: Sharing and Reusing Ground Truth Data, Referencing Models, and Acknowledging Contributions. Starting the Conversation on How We Could Get It Done*. 2023. DOI: 10.5281/zenodo.7267244. URL: https://hal.science/hal-04244372.

[70] 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*. 20th Nov. 2023. URL: https://inria.hal.science/hal-03850124.

12.3 Other

Scientific popularization

[71] A. Chagué and T. Clérice. *017 - Deploying eScriptorium online: notes on CREMMA's server specifications*. 22nd Dec. 2023. URL: https://inria.hal.science/hal-04362085.

12.4 Cited publications

[72] 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. 12 pages, 6 figures, 2 tables. Marseille, France, June 2022. URL: https://inria.hal.science/hal-03536361.

[73] 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.

[74] 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 (2022). Ed. by B. Roark and A. Nenkova, pp. 1051–1065. DOI: 10.1162/tacl_a_00505. URL: https://aclanthology.org/2022.tacl-1.61.

[75] 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.

[76] 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.

[77] F. Barbieri, L. Espinosa-Anke and J. Camacho-Collados. 'A Multilingual Language Model Toolkit for Twitter'. In: *arXiv preprint arXiv:2104.12250*. 2021.

[78] 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.

[79] 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.

[80] 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.

[81] J. Canete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang and J. Pérez. 'Spanish pre-trained bert model and evaluation data'. In: *Pml4dc at iclr* 2020 (2020), p. 2020.

[82] 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.

[83] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer and V. Stoyanov. 'Unsupervised Cross-lingual Representation Learning at Scale'. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by D. Jurafsky, J. Chai, N. Schluter and J. Tetreault. Online: Association for Computational Linguistics, July 2020, pp. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747. URL: https://aclanthology.org/2020.acl-main.747.

[84] 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.

[85] 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.

[86] J. Devlin, M. 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/.

[87] E. Dunbar, X. N. Cao, J. Benjumea, J. Karadayi, M. Bernard, L. Besacier, X. Anguera and E. Dupoux. 'The Zero Resource Speech Challenge 2017'. In: *Proceedingsn of the IEEE Workshop on Automatic Speech Recognition and Understanding*. Okinawa, Japan, 2017. URL: https://hal.science/hal-01664586.

[88] P.-A. Duquenne, H. Gong and H. Schwenk. 'Multimodal and Multilingual Embeddings for Large-Scale Speech Mining'. In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang and J. W. Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 15748–15761. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/8466f9ace6a9acbe71f75762ffc890f1-Paper.pdf.

[89] Y. Fang and M. 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.

[90] 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://inria.hal.science/hal-01174492.

[91] 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.

[92] R. van der Goot, A. Üstün, A. Ramponi, I. Sharaf and B. Plank. 'Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP'. In: *arXiv preprint arXiv:2005.14672* (2020).

[93] P. He, J. Gao and W. Chen. 'DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing'. 2023. arXiv: 2111.09543 [cs.CL].

[94] K. Heafield. 'KenLM: Faster and Smaller Language Model Queries'. In: *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Ed. by C. Callison-Burch, P. Koehn, C. Monz and O. F. Zaidan. Edinburgh, Scotland: Association for Computational Linguistics, July 2011, pp. 187–197. URL: https://aclanthology.org/W11-2123.

[95] D. Hershcovich, S. Frank, H. Lent, M. de Lhoneux, M. Abdou, S. Brandl, E. Bugliarello, L. Cabello Piqueras, I. Chalkidis, R. Cui, C. Fierro, K. Margatina, P. Rust and A. Søgaard. 'Challenges and Strategies in Cross-Cultural NLP'. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by S. Muresan, P. Nakov and A. Villavicencio. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 6997–7013. DOI: 10.18653/v1/2022.acl-long.482. URL: https://aclanthology.org/2022.acl-long.482.

[96] 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.

[97] 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.

[98] D. Hovy and D. Yang. 'The Importance of Modeling Social Factors of Language: Theory and Practice'. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty and Y. Zhou. Online: Association for Computational Linguistics, June 2021, pp. 588–602. DOI: 10.18653/v1/2021.naacl-main.49. URL: https://aclanthology.org/2021.naacl-main.49.

[99] 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.

[100] 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.

[101] 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.

[102] 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 (2022). Ed. by B. Roark and A. Nenkova, pp. 50–72. DOI: 10.1162/tacl_a_00447. URL: https://aclanthology.org/2022.tacl-1.4.

[103] 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.

[104] 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.

[105] 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).

[106] S. M. Lundberg and S.-I. Lee. 'A unified approach to interpreting model predictions'. In: *Advances in neural information processing systems* 30 (2017).

[107] 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.

[108] 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.

[109] 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: *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*. Accepted to Findings of AACL-IJCNLP 2022. Online, France, Nov. 2022. URL: https://inria.hal.science/hal-03840070.

[110] 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'. working paper or preprint. Oct. 2022. URL: https://inria.hal.science/hal-03834730.

[111] 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* (Aug. 2022). URL: https://inria.hal.science/hal-03831707.

[112] D. Nozza. 'Exposing the limits of Zero-shot Cross-lingual Hate Speech Detection'. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Ed. by C. Zong, F. Xia, W. Li and R. Navigli. Online: Association for Computational Linguistics, Aug. 2021, pp. 907–914. DOI: 10.18653/v1/2021.acl-short.114. URL: https://aclanthology.org/2021.acl-short.114.

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

[114] 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.

[115] 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.

[116] 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.

[117] T. Pires, E. Schlinger and D. Garrette. 'How Multilingual is Multilingual BERT?' In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 4996–5001. DOI: 10.18653/v1/P19-1493. URL: https://aclanthology.org/P19-1493.

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

[119] 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.

[120] 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.

[121] A. Riabi, S. Montariol and D. Seddah. 'Tâches Auxiliaires Multilingues pour le Transfert de Modèles de Détection de Discours Haineux (Multilingual Auxiliary Tasks for Zero-Shot Cross-Lingual Transfer of Hate Speech Detection)'. French. In: *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*. Ed. by Y. Estève, T. Jiménez, T. Parcollet and M. Zanon Boito. Avignon, France: ATALA, June 2022, pp. 413–423. URL: https://aclanthology.org/2022.jeptalnrecital-taln.41.

[122] 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.

[123] 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.

[124] L. Romary and P. Lopez. 'GROBID - Information Extraction from Scientific Publications'. In: *ERCIM News*. Scientific Data Sharing and Re-use 100 (Jan. 2015). URL: https://hal.inria.fr/hal-01673305.

[125] 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.

[126] 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.

[127] 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.

[128] 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.

[129] 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.

[130] 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.

[131] B. Sagot. 'The Lefff, a Freely Available and Large-coverage Morphological and Syntactic Lexicon for French'. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Ed. by N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner and D. Tapias. Valletta, Malta: European Language Resources Association (ELRA), May 2010. URL: http://www.lrec-conf.org/proceedings/lrec2010/pdf/701_Paper.pdf.

[132] B. Sagot and É. de la Clergerie. 'Error Mining in Parsing Results'. In: *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Ed. by N. Calzolari, C. Cardie and P. Isabelle. Sydney, Australia: Association for Computational Linguistics, July 2006, pp. 329–336. DOI: 10.3115/1220175.1220217. URL: https://aclanthology.org/P06-1042.

[133] 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.

[134] 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.

[135] 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.

[136] 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.

[137] 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.

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

[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] 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.

[144] 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.

[145] 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.

[146] 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.