RESEARCH CENTRE

Paris

2020 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 team: 2017 January 01, updated into Project-Team: 2019 July 01

Keywords

Computer sciences and digital sciences

A3.2.2. – Knowledge extraction, cleaning

A3.3.2. – Data mining

A3.3.3. – Big data analysis

A3.4.1. – Supervised learning

A3.4.2. – Unsupervised learning

A3.4.6. – Neural networks

A3.4.8. – Deep learning

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

B9.6.5. – Sociology

B9.6.6. – Archeology, History

B9.6.8. – Linguistics

B9.6.10. – Digital humanities

B9.7. – Knowledge dissemination

B9.7.1. – Open access

B9.7.2. – Open data

B9.8. – Reproducibility

1 Team members, visitors, external collaborators

Research Scientists

- Benoît Sagot [Team leader, Inria, Senior Researcher, HDR]
- Rachel Bawden [Inria, Researcher, from Nov 2020]
- Laurent Romary [Inria, Senior Researcher, HDR]
- Djamé Seddah [Inria, Researcher]
- Éric Villemonte de La Clergerie [Inria, Researcher]

Faculty Members

- Kim Gerdes [Université Sorbonne Nouvelle, Associate Professor, until Aug 2020, HDR]
- Loic Grobol [Ministère de l'Education Nationale, Professor, until Oct 2020]

Post-Doctoral Fellows

- Murielle Fabre [Inria, until Aug 2020]
- Gael Guibon [Université Denis Diderot, until June 2020]

PhD Students

- Jack Bowers [Académie autrichienne des sciences, until September 2020]
- Clementine Fourrier [Inria]
- Mohamed Khemakhem [Université Grenoble Alpes, until Aug 2020]
- Louis Martin [Facebook]
- Benjamin Muller [Inria, until Aug 2020]
- Pedro Ortiz Suarez [Inria]
- Mathilde Regnault [École Normale Supérieure de Paris]
- Jean Baptiste Remy [École polytechnique, Sep 2020]
- Jose Rosales Nunez [CNRS]
- Lionel Tadonfouet [Orange, CIFRE, from Mar 2020]

Technical Staff

- Alix Chague [Inria, Engineer]
- Floriane Chiffoleau [Inria, Engineer, from Mar 2020]
- Tanti Kristanti Nugraha [Inria, Engineer]
- Arij Riabi [Inria, Engineer, from Oct 2020]
- Yves Tadjo Takianpi [Inria, Engineer, from Sep 2020]
- Lionel Tadonfouet [Inria, Engineer, until Mar 2020]
- Lucas Terriel [Inria, Engineer, from Nov 2020]

Interns and Apprentices

- Damien Biabiany [Inria, Apprentice, until Sep 2020]
- Quentin Burthier [Inria, from Sep 2020]
- Matthieu Futeral-Peter [Inria, Jul 2020]
- Lucas Terriel [Inria, from Mar 2020 until Jul 2020]

Administrative Assistant

- Meriem Guemair [Inria]

Visiting Scientists

- Marine Courtin [Sorbonne Université, until Aug 2020]
- Jean Damien Genero [Université Denis Diderot, from Mar 2020 until Sep 2020]

External Collaborators

- Patrice Lopez [Science-Miner SASU]
- Benjamin Muller [Apple, from Sep 2020, césure]

2 Overall objectives

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

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

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

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

One of the main challenges in computational linguistics is **to model and to cope with language variation**. Language varies with respect to domain and genre (news wires, scientific literature, poetry, oral transcripts…), sociolinguistic factors (age, background, education; variation attested for instance on social media), geographical factors (dialects) and other dimensions (disabilities, for instance). But

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

language also constantly evolves at all time scales. Addressing this variability is still an open issue for NLP. Commonly used approaches, which often rely on supervised and semi-supervised machine learning methods, require very large amounts of annotated data. They still suffer from the high level of variability found for instance in **user-generated content**, **non-contemporary texts**, as well as in **domain-specific documents** (e.g. financial, legal).

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

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

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

3 Research program

3.1 Research strands

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

3.1.1 Research axis 1

Our first objective is to **stay at a state-of-the-art level in key NLP tasks** such as shallow processing, part-of-speech tagging and (syntactic) parsing, which are core expertise domains of ALMAnaCH members. This will also require us to improve the **generation of semantic representations (semantic parsing)**, and to begin to explore tasks such as machine translation, which now relies on neural architectures also used for some of the above-mentioned tasks. Given the generalisation of neural models in NLP, we will also be involved in better understanding how such models work and what they learn, something that is directly related to the investigation of language variation (Research axis 2). We will also work on the **integration of both linguistic and non-linguistic contextual information** to improve automatic linguistic analysis. This is an emerging and promising line of research in NLP. We will have to identify, model and take advantage of each type of contextual information available. Addressing these issues will enable the development of new lines of research related to conversational content. Applications include improved information and knowledge extraction algorithms. We will especially focus on challenging

datasets such as domain-specific texts (e.g. financial, legal) as well as historical documents, in the larger context of the development of digital humanities. We currently also explore the even more challenging new direction of a cognitively inspired NLP, in order to tackle the possibility to enrich the architecture of state-of-the-art algorithms, such as RNNGs, based on human neuroimaging-driven data.

3.1.2 Research axis 2

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

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

3.1.3 Research axis 3

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

3.2 Automatic Context-augmented Linguistic Analysis

This first research strand is centred around NLP technologies and some of their applications in Artificial Intelligence (AI). Core NLP tasks such as part-of-speech tagging, syntactic and semantic parsing is improved by integrating new approaches, such as (deep) neural networks, whenever relevant, while preserving and taking advantage of our expertise on symbolic and statistical system: hybridisation not only couples symbolic and statistical approaches, but neural approaches as well. AI applications are twofold, notwithstanding the impact of language variation (see the next strand): (i) information and

knowledge extraction, whatever the type of input text (from financial documents to ancient, historical texts and from Twitter data to Wikipedia) and (ii) chatbots and natural language generation. In many cases, our work on these AI applications is carried out in collaboration with industrial partners. The specificities and issues caused by language variation (a text in Old French, a contemporary financial document and tweets with a non-canonical spelling cannot be processed in the same way) are addressed in the next research strand.

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

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

Traditionally, a full natural language process (NLP) chain is organised as a pipeline where each stage of analysis represents a traditional linguistic field (in a *structuralism* view) from morphological analysis to purely semantic representations. The problem is that this architecture is vulnerable to error propagation and very domain sensitive: each of these stage must be compatible at the lexical and structure levels they provide. We arguably built the best performing NLP chain for French [70, 114] and one of the best for robust multilingual parsing as shown by our results in various shared tasks over the years [108, 106, 115], [77]. So we pursue our efforts on each of our components we developed: tokenisers (e.g. SxPipe), part-of-speech taggers (e.g. MElt), constituency parsers and dependency parsers (e.g. FRMG, DyALog-SR) as well as our recent neural semantic graph parsers [106].

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

Fundamentally, we want to build tools that are less sensitive to variation, more easily configurable, and self-adapting. Our short-term goal is to explore techniques such as multi-task learning (cf. already [111]) to propose a joint model of tokenisation, normalisation, morphological analysis and syntactic analysis. We also explore adversarial learning, considering the drastic variation we face in parsing user-generated content and processing historical texts, both seen as noisy input that needs to be handled at training and decoding time.

3.2.2 Integrating context in NLP systems

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

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

A more expressive and context-aware structural representation of a linguistic production is required to accurately model UGC. Let us consider for instance the case for Syntax-based Machine Translation of social media content, as is carried out by the ALMAnaCH-led ANR project Parsiti (PI: DS). A Facebook post may be part of a discussion thread, which may include links to external content. Such information is required for a complete representation of the post's context, and in turn its accurate machine translation. Even for the presumably simpler task of POS tagging of dialogue sequences, the addition of context-based features (namely information about the speaker and dialogue moves) was beneficial [80]. In the case of UGC, working across sentence boundaries was explored for instance, with limited success, by [69] for document-wise parsing and by [98] for POS tagging.

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

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

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 [67]. In this regard, ALMAnaCH members have taken part in the design of quantitative approaches to defining and measuring morphological complexity and to assess the internal structure of morphological systems (inflection classes, predictability of inflected forms…). Such studies provide valuable insights on these prominent questions in theoretical morphology. They also improve the linguistic relevance and the development speed of NLP-oriented lexicons, as also demonstrated by ALMAnaCH members. We shall therefore pursue these investigations, and orientate them towards their use in diachronic models (see section 3.3.3).

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

3.3.2 Sociolinguistic variation

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

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

However, NLP tools are designed for standard forms of language and exhibit a drastic loss of accuracy when applied to social media varieties or non-standardised historical sources. To define appropriate tools, descriptions of these varieties are needed. However, to validate such descriptions, tools are also needed. We address this chicken-and-egg problem in an interdisciplinary fashion, by working both on linguistic descriptions and on the development of NLP tools. Recently, socio-demographic variables have been shown to bear a strong impact on NLP processing tools (see for instance [74] and references therein). This is why, in a first step, jointly with researchers involved in the ANR project SoSweet (ENS Lyon and Inria project-team Dante), we will study how these variables can be factored out by our models and, in a second step, how they can be accurately predicted from sources lacking these kinds of featured descriptions.

3.3.3 Diachronic variation

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

Very few computational models of language change have been developed. Simple deterministic finite-state-based phonetic evolution models have been used in different contexts. The PIElexicon project [93] uses such models to automatically generate forms attested in (classical) Indo-European languages but is based on an idiosyncratic and unacceptable reconstruction of the Proto-Indo-European language. Probabilistic finite-state models have also been used for automatic cognate detection and proto-form reconstruction, for example by [68] and [75]. Such models rely on a good understanding of the phonetic evolution of the languages at hand.

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

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

The second application task is computational etymology and proto-language reconstruction. Our lexical diachronic evolution models will be paired with semantic resources (wordnets, word embeddings, and other corpus-based statistical information). This will allow us to formally validate or suggest etymological

or cognate relations between lexical entries from different languages of a same language family, provided they are all inherited. Such an approach could also be adapted to include the automatic detection of borrowings from one language to another (e.g. for studying the non-inherited layers in the Ancient Greek lexicon). In the longer term, we will investigate the feasibility of the automatic (unsupervised) acquisition of phonetic change models, especially when provided with lexical data for numerous languages from the same language family.

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

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

3.3.4 Accessibility-related variation

Language variation does not always pertain to the textual input of NLP tools. It can also be characterised by their intended output. This is the perspective from which we investigate the issue of text simplification (for a recent survey, see for instance [110]). Text simplification is an important task for improving the accessibility to information, for instance for people suffering from disabilities and for non-native speakers learning a given language [94]. To this end, guidelines have been developed to help writing documents that are easier to read and understand, such as the FALC ("Facile À Lire et à Comprendre") guidelines for French. [4]

Fully automated text simplification is not suitable for producing high-quality simplified texts. Besides, the involvement of disabled people in the production of simplified texts plays an important social role. Therefore, following previous works [73, 104], our goal will be to develop tools for the computer-aided simplification of textual documents, especially administrative documents. Many of the FALC guidelines can only be linguistically expressed using complex, syntactic constraints, and the amount of available "parallel" data (aligned raw and simplified documents) is limited. We will therefore investigate hybrid techniques involving rule-based, statistical and neural approaches based on parsing results (for an example of previous parsing-based work, see [66]). Lexical simplification, another aspect of text simplification [82, 95], will also be pursued. In this regard, we have already started a collaboration with Facebook's AI Research in Paris, the UNAPEI (the largest French federation of associations defending and supporting people with intellectual disabilities and their families), and the French Secretariat of State in charge of Disabled Persons.

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

3.4 Modelling and Development of Language Resources

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

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

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

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

meaningful and computationally exploitable ways to encode linguistic information within language resources constitutes the first main scientific challenge in language resource development. It requires a strong expertise on both the linguistic issues underlying the type of resource under development (e.g. on syntax when developing a treebank) and the NLP algorithms that will make use of such information.

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

3.4.1 Construction, management and automatic annotation of Text Corpora

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

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

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

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

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

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

configured from the interface. An option will be to identify pertinent emerging platforms in DH (such as Transkribus) and to propose collaborations to ensure that NLP modules can be easily integrated.

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

3.4.2 Development of Lexical Resources

ALPAGE, the Inria predecessor of ALMAnaCH, has put a strong emphasis in the development of morphological, syntactic and wordnet-like semantic lexical resources for French as well as other languages (see for instance [8, 1]). Such resources play a crucial role in all NLP tools, as has been proven among other tasks for POS tagging [100, 102, 115] 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 [116]), especially ancient languages. Finally, semantic lexicons such as wordnets will play a crucial role in assessing lexical similarity and automating etymological research.

Therefore, an important effort towards the development of new morphological lexicons will be initiated, with a focus on ancient languages of interest. Following previous work by ALMAnaCH members, we will try and leverage all existing resources whenever possible such as electronic dictionaries, OCRised dictionaries, both modern and ancient [99, 79, 101], while using and developing (semi)automatic lexical information extraction techniques based on existing corpora [103, 105]. A new line of research will be to integrate the diachronic axis by linking lexicons that are in diachronic relation with one another thanks to phonetic and morphological change laws (e.g. XIIth century French with XVth century French and contemporary French). Another novelty will be the integration of etymological information in these lexical resources, which requires the formalisation, the standardisation, and the extraction of etymological information from OCRised dictionaries or other electronic resources, as well as the automatic generation of candidate etymologies. These directions of research are already investigated in ALMAnaCH [79, 101].

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

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

3.4.3 Development of Annotated Corpora

Along with the creation of lexical resources, ALMAnaCH is also involved in the creation of corpora either fully manually annotated (gold standard) or automatically annotated with state-of-the-art pipeline processing chains (silver standard). Annotations will either be only morphosyntactic or will cover more complex linguistic levels (constituency and/or dependency syntax, deep syntax, maybe semantics). Former members of the ALPAGE project have a renowned experience in those aspects (see for instance

[109, 96, 107, 87]) and will participate to the creation of valuable resources originating from the historical domain genre.

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

4 Application domains

4.1 Application domains for ALMAnaCH

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

- Information extraction, information retrieval, text mining (e.g. opinion surveys)
- Text generation, text simplification, automatic summarisation
- Spelling correction (writing aid, post-OCR, normalisation of noisy/non-canonical texts)
- Machine translation, computer-aided translation
- Chatbots, conversational agents, question answering systems
- Medical applications (early diagnosis, language-based medical monitoring...)
- Applications in linguistics (modelling languages and their evolution, sociolinguistic studies...)
- Digital humanities (exploitation of text documents, for instance in historical research)

5 Highlights of the year

5.1 Awards

- The DARIAH Working Group on Lexical Resources, co-chaired by Laurent Romary, has been named winner of the 2020 Rahtz Prize for TEI Ingenuity in recognition of their work on TEI Lex-0, a technical specification and a set of community-based recommendations for encoding machine-readable dictionaries. https://www.dariah.eu/2020/11/20/dariah-working-group-on-lexical-resources-wins-innovation-prize/

6 New software and platforms

6.1 New software

6.1.1 Enqi

Author: Benoît Sagot

Contact: Benoît Sagot

6.1.2 SYNTAX

Keyword: Parsing

Functional Description: Syntax system includes various deterministic and non-deterministic CFG parser generators. It includes in particular an efficient implementation of the Earley algorithm, with many original optimizations, that is used in several of Alpage's NLP tools, including the pre-processing chain Sx Pipe and the LFG deep parser SxLfg . This implementation of the Earley algorithm has been recently extended to handle probabilistic CFG (PCFG), by taking into account probabilities both during parsing (beam) and after parsing (n-best computation).

URL: http://syntax.gforge.inria.fr/

Authors: Pierre Boullier, Philippe Deschamps, Benoît Sagot

Contacts: Pierre Boullier, Benoît Sagot

Participants: Benoît Sagot, Pierre Boullier

6.1.3 FRMG

Keywords: Parsing, French

Functional Description: FRMG is a large-coverage linguistic meta-grammar of French. It can be compiled (using MGCOMP) into a Tree Adjoining Grammar, which, in turn, can be compiled (using DyALog) into a parser for French.

URL: http://mgkit.gforge.inria.fr/

Contact: Éric De La Clergerie

Participant: Eric de La Clergerie

6.1.4 MElt

Name: Maximum-Entropy lexicon-aware tagger

Keyword: Part-of-speech tagger

Functional Description: MElt is a freely available (LGPL) state-of-the-art sequence labeller that is meant to be trained on both an annotated corpus and an external lexicon. It was developed by Pascal Denis and Benoît Sagot within the Alpage team, a joint INRIA and Université Paris-Diderot team in Paris, France. MElt allows for using multiclass Maximum-Entropy Markov models (MEMMs) or multiclass perceptrons (multitrons) as underlying statistical devices. Its output is in the Brown format (one sentence per line, each sentence being a space-separated sequence of annotated words in the word/tag format).

MElt has been trained on various annotated corpora, using Alexina lexicons as source of lexical information. As a result, models for French, English, Spanish and Italian are included in the MElt package.

MElt also includes a normalization wrapper aimed at helping processing noisy text, such as user-generated data retrieved on the web. This wrapper is only available for French and English. It was used for parsing web data for both English and French, respectively during the SANCL shared task (Google Web Bank) and for developing the French Social Media Bank (Facebook, twitter and blog data).

URL: https://team.inria.fr/almanach/melt/

Contact: Benoît Sagot

6.1.5 dyalog-sr

Keywords: Parsing, Deep learning, Natural language processing

Functional Description: DyALog-SR is a transition-based dependency parser, built on top of DyALog system. Parsing relies on dynamic programming techniques to handle beams. Supervised learning exploit a perceptron and aggressive early updates. DyALog-SR can handle word lattice and produce dependency graphs (instead of basic trees). It was tested during several shared tasks (SPMRL'2013 and SEMEVAL'2014). It achieves very good accuracy on French TreeBank, alone or by coupling with FRMG parser. In 2017, DyALog-SR has been extended into DyALog-SRNN by adding deep neuronal layers implemented with the Dynet library. The new version has participated to the evaluation campaigns CONLL UD 2017 (on more than 50 languages) and EPE 2017.

Contact: Éric De La Clergerie

6.1.6 FSMB

Name: French Social Media Bank

Keywords: Treebank, User-generated content

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

Contacts: Djamé Seddah, Benoît Sagot

6.1.7 DyALog

Keyword: Logic programming

Functional Description: DyALog provides an environment to compile and execute grammars and logic programs. It is essentially based on the notion of tabulation, i.e. of sharing computations by tabulating traces of them. DyALog is mainly used to build parsers for Natural Language Processing (NLP). It may nevertheless be used as a replacement for traditional PROLOG systems in the context of highly ambiguous applications where sub-computations can be shared.

URL: http://dyalog.gforge.inria.fr/

Contact: Eric de La Clergerie

Participant: Eric de La Clergerie

6.1.8 SxPipe

Keyword: Surface text processing

Scientific Description: Developed for French and for other languages, Sx Pipe includes, among others, various named entities recognition modules in raw text, a sentence segmenter and tokenizer, a spelling corrector and compound words recognizer, and an original context-free patterns recognizer, used by several specialized grammars (numbers, impersonal constructions, quotations...). It can now be augmented with modules developed during the former ANR EDyLex project for analysing unknown words, this involves in particular (i) new tools for the automatic pre-classification of unknown words (acronyms, loan words...) (ii) new morphological analysis tools, most notably automatic tools for constructional morphology (both derivational and compositional), following the results of dedicated corpus-based studies. New local grammars for detecting new types of entities and improvement of existing ones, developed in the context of the PACTE project, will soon be integrated within the standard configuration.

Functional Description: SxPipe is a modular and customizable processing chain dedicated to applying to raw corpora a cascade of surface processing steps (tokenisation, wordform detection, non-deterministic spelling correction...). It is used as a preliminary step before ALMAnaCH's parsers (e.g., FRMG) and for surface processing (named entities recognition, text normalization, unknown word extraction and processing...).

URL: http://lingwb.gforge.inria.fr/

Authors: Benoît Sagot, Pierre Boullier

Contact: Benoît Sagot

Participants: Benoît Sagot, Djamé Seddah, Eric de La Clergerie

6.1.9 Mgwiki

Keywords: Parsing, French

Functional Description: Mgwiki is a linguistic wiki that may used to discuss linguistic phenomena with the possibility to add annotated illustrative sentences. The work is essentially devoted to the construction of an instance for documenting and discussing FRMG, with the annotations of the sentences automatically provided by parsing them with FRMG. This instance also offers the possibility to parse small corpora with FRMG and an interface of visualization of the results. Large parsed corpora (like French Wikipedia or Wikisource) are also available. The parsed corpora can also be queried through the use of the DPath language.

URL: http://alpage.inria.fr/frmgwiki/

Contact: Eric de La Clergerie

Participant: Eric de La Clergerie

6.1.10 WOLF

Name: WOrdnet Libre du Français (Free French Wordnet)

Keywords: WordNet, French, Semantic network, Lexical resource

Functional Description: The WOLF (Wordnet Libre du Français, Free French Wordnet) is a free semantic lexical resource (wordnet) for French.

The WOLF has been built from the Princeton WordNet (PWN) and various multilingual resources.

URL: http://alpage.inria.fr/~sagot/wolf-en.html

Contact: Benoît Sagot

6.1.11 vera

Keyword: Text mining

Functional Description: Automatic analysis of answers to open-ended questions based on NLP and statistical analysis and visualisation techniques (vera is currently restricted to employee surveys).

Authors: Benoît Sagot, Dimitri Tcherniak

Contact: Benoît Sagot

Participants: Benoît Sagot, Dimitri Tcherniak

Partner: Verbatim Analysis

6.1.12 Alexina

Name: Atelier pour les LEXiques INformatiques et leur Acquisition

Keyword: Lexical resource

Functional Description: Alexina is ALMAnaCH's framework for the acquisition and modeling of morphological and syntactic lexical information. The first and most advanced lexical resource developed in this framework is the Lefff, a morphological and syntactic lexicon for French.

URL: http://gforge.inria.fr/projects/alexina/

Contact: Benoît Sagot

Participant: Benoît Sagot

6.1.13 FQB

Name: French QuestionBank

Keyword: Treebank

Functional Description: The French QuestionBanks is a corpus of around 2000 questions coming from various domains (TREC data set, French governmental organisation, NGOs, etc..) it contains different kind of annotations - morpho-syntactic ones (POS, lemmas) - surface syntaxe (phrase based and dependency structures) with long-distance dependency annotations.

The TREC part is aligned with the English QuestionBank (Judge et al, 2006).

Contacts: Djamé Seddah, Marie Candito

6.1.14 Sequoia corpus

Keyword: Treebank

Functional Description: The Sequoia corpus contains French sentences, annotated with various linguistic information: - parts-of-speech - surface syntactic representations (both constituency trees and dependency trees) - deep syntactic representations (which are deep syntactic dependency graphs)

Contacts: Djamé Seddah, Marie Candito, Bruno Guillaume

6.2 New platforms

7 New results

7.1 New results on text simplification

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

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

In 2017 we initiated a collaboration with the Facebook Artificial Intelligence Research (FAIR) lab in Paris and with the UNAPEI, the federation of French associations helping people with mental disabilities and their families. The objective of this collaboration is to develop tools to help the simplification of texts aimed at mentally disabled people. More precisely, the aim is to develop a computer-assisted text simplification platform (as opposed to an automatic TS system). In this context, a CIFRE PhD thesis was initiated in collaboration with the FAIR on the TS task. We first dedicated important efforts to the problem of the evaluation of TS systems, which remains an open challenge. Since the task has common points with machine translation (MT), TS is often evaluated using MT metrics such as BLEU [91]. However, such metrics require high quality reference data, which is rarely available for TS. TS has the advantage over MT of being a monolingual task, which allows for direct comparisons to be made between the simplified text and its original version. We compared multiple approaches to reference-less quality estimation of sentence-level TS systems, based on the dataset used for the QATS 2016 shared task [113]. We distinguished three different dimensions: grammaticality, meaning preservation and simplicity. We showed that n-gram-based MT metrics such as BLEU and METEOR [81] correlate the most with human judgements of grammaticality and meaning preservation, whereas simplicity is best evaluated by basic length-based metrics [84]. This year, our implementations of several metrics were made easily accessible and described in a demo paper in collaboration with the University of Sheffield [65].

In 2019, we investigated an important issue inherent to the TS task. Although it is often considered an all-purpose generic task where the same simplification is suitable for all, multiple audiences can benefit from simplified text in different ways. We therefore introduced a discrete parametrisation mechanism providing explicit control for TS systems based on Seq2Seq neural models; users can condition the simplifications returned by a model on parameters such as length and lexical complexity. We also showed that carefully chosen values of these parameters allow out-of-the-box Seq2Seq neural models to outperform their standard counterparts on simplification benchmarks. Our best parametrised model improves over the previous state-of-the-art performance, as shown by our 2020 publication on the topic [86].

In 2020, we have been involved in the development of a new text simplification corpus, in collaboration with the University of Sheffield. In order to simplify a sentence, human editors perform multiple rewriting transformations: splitting it into several shorter sentences, paraphrasing (i.e. replacing complex words or phrases by simpler synonyms), reordering components and/or deleting information deemed unnecessary. Despite the vast range of possible text alterations, current models for automatic sentence simplification are evaluated using datasets that focus on single transformations, such as paraphrasing or splitting. This makes it impossible to understand the ability of simplification models in more abstractive and realistic settings. This is what motivated the development of ASSET, a new dataset for assessing sentence simplification in English [18]. ASSET is a crowdsourced multi-reference corpus in which each simplification was produced by executing several rewriting transformations. Through quantitative and qualitative experiments, we have shown that simplifications in ASSET are better at capturing characteristics of simplicity when compared to other standard evaluation datasets for the task. Furthermore, we motivated the need to develop better methods of automatic evaluation using ASSET, since we show that current popular metrics may not be suitable for assessment when multiple simplification transformations

are performed.

In 2020 we also extended the scope of TS to multiple languages using a totally unsupervised method and with state-of-the-art results, even compared to previous supervised models [49]. We use a large-scale mining pipeline to extract one billion sentences from the web using Common Crawl. These sentences are used to find millions of paraphrases in three languages: English, French and Spanish. This allows us to train our controllable models [86] in multiple languages, which we then adapt for the TS task. These models push the state-of-the-art further in all three languages and benchmarks. Generated simplifications are considered more fluent, and simpler than previous models. This research enables us to exploit our previous works and apply them to the Cap'FALC project, whose aim is to assist the simplification of French documents by people with intellectual disabilities.

7.2 Neural language modelling

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

Pretrained language models are now ubiquitous in Natural Language Processing. Despite their success, most available models have either been trained on English data or on the concatenation of data in multiple languages [71, 83]. This makes practical use of such models—in all languages except English—very limited. In 2019, one of the most visible achievements of the ALMAnaCH team was the training and release of CamemBERT, a BERT-like [71] (and more specifically a RoBERTa-like) neural language model for French trained on the French section of our large-scale web-based OSCAR corpus [90], together with CamemBERT variants [85]. Our goal was to investigate the feasibility of training monolingual Transformer-based language models for other languages, taking French as an example and evaluating our language models on part-of-speech tagging, dependency parsing, named entity recognition and natural language inference tasks. We showed that the use of web-crawled data (such as found in OSCAR) to train such language models is preferable to the use of Wikipedia data, because of the homogeneity of Wikipedia data. More surprisingly, we also showed that a relatively small web crawled dataset (4GB of randomly extracted text from the French section of OSCAR) leads to results that are as good as those obtained using larger datasets (130+GB, i.e. the whole French section of OSCAR) [29, 37]. CamemBERT allowed us to reach or improve the state of the art in all four downstream tasks.

In parallel, we used OSCAR to train ELMo language models [92], i.e. LSTM-based monolingual contextualised language models, for five mid-resource languages (Catalan, Finnish, Indonesian, Bulgarian and Danish). We showed that, despite the noise in the web-based OSCAR data, embeddings trained on OSCAR perform much better than monolingual embeddings trained on Wikipedia [33]. They actually equal or improve the current state of the art in tagging and parsing for all five languages. In particular, they also improve over multilingual Wikipedia-based contextual embeddings (multilingual BERT), which almost always constitute the previous state of the art, thereby showing that the benefit of a larger, more diverse corpus surpasses the cross-lingual benefit of multilingual embedding architectures, a result in line with our findings on CamemBERT.

7.3 Named entity recognition

Participants Pedro Ortiz Suárez, Tanti Kristanti, Laurent Romary.

Two teams involving ALMAnaCH members participated in the CLEF-HIPE 2020 challenge on Named Entity (NE) Processing on old newspapers, using different approaches and competing on different tasks within the challenge.

One team, a collaboration with the LATTICE (Sorbonne Université), competed under the name SinNer on Named Entity Recognition in French and German texts. The best system we proposed ranked third for these two languages. It uses FastText embeddings and ELMo language models, which we trained ourselves (FrELMo and German ELMo). We show that combining several word representations enhances

the quality of the results for all NE types and that the segmentation in sentences has an important impact on the results [31].

The other team competed in the NERC-COARSE-LIT and "Entity Linking only" (EL-ONLY) tasks for English and French [28] using two systems jointly developed by Patrice Lopez and ALMAnaCH: 1) DeLFT, a Deep Learning framework for text processing; 2) entity-fishing, generic NE recognition and disambiguation service. The main goal of this participation was to assess the performance level of these tools, rather than to do everything possible to perform optimally in the competition. Despite this, we ranked second for the NERC coarse English strict (literal sense) task and first for the EL-ONLY task for English.

7.4 Cognate prediction

Participants Clémentine Fourrier, Benoît Sagot.

In 2020 we continued our experiments to investigate whether and under which conditions neural networks can be used to learn sound correspondences between two related languages, i.e. for the prediction of cognates of source language words in a related target language. In particular, we investigated the learnability of sound correspondences between a proto-language and daughter languages for two machine-translation-inspired models, one statistical, the other neural [21]. We first carried out our experiments on plausible artificial languages, without noise, in order to study the role of each parameter on the algorithms' respective performance under almost perfect conditions. We then studied real languages, namely Latin, Italian and Spanish, using our EtymDB 2.0 etymological database [22], to see if the performance generalises well. We showed that both model types can learn sound changes despite data sparsity, although the best performing model type depends on multiple parameters such as the size of the training data, the ambiguity, and the prediction direction.

Since the experiments published in [21], we have improved both our way to extract data from EtymDB and our neural systems in multiple ways, leading to improved results that should be published in 2021.

7.5 Detecting Word Variability in User-Generated Content

Participants Djamé Seddah, Ganesh Jawahar, Benjamin Muller.

The problem of comparing two bodies of text and searching for words that differ in their usage between them arises often in digital humanities and computational social science and is becoming more and more crucial and is one of the keys to detecting emerging community that may use a specific, almost self-identifying language. This is why as part of the ANR project SoSweet and the PHC Maimonide projects (in collaboration with Bar Ilan University for the latter), ALMAnaCH has invested a lot of effort since 2018 into studying language variation within user-generated content (UGC), taking into account two main interrelated dimensions: how language variation is related to socio-demographic and dynamic network variables, and how UGC language evolves over time. Taking advantage of the SoSweet corpus (600 million tweets) and of the Bar Ilan Hebrew Tweets (180M tweets) both collected over the last 5 years, we have been addressing the problem of studying semantic changes via the use of dynamic word embeddings (i.e. embeddings evolving over time). We devised a novel attention model, based on Bernouilli word embeddings, which are conditioned on contextual extra-linguistic features such as network, spatial and socio-economicvariables, which can be inferred from Twitter users metadata, as well as topic-based features. We posit that these social features provide an inductive bias that is likely to help our model overcome the narrow time-span regime problem. Our extensive experiments reveal that, as a result of being less biased towards frequency cues, our proposed model is able to capture subtle semantic shifts and therefore benefits from the inclusion of a reduced set of contextual features. Our model therefore fits the data better than current state-of-the-art dynamic word embedding models and therefore is a

promising tool to study diachronic semantic changes over short time periods. These ideas and results are published in [78].

In addition to detecting semantic word variations over time, in the same Maimonide project, we explored the question of detecting word usage change by proposing a simple (yet effective) method that does not rely on complex word embeddings reformulation and a rich feature set. A common approach to the task of detecting word usage change is to train word embeddings on each corpus, to align the vector spaces, and then look for words whose cosine distance in the aligned space is large. However, these methods often require extensive filtering of the vocabulary to perform well, and they result in unstable and therefore less reliable results. We proposed an alternative approach that does not use vector space alignment and instead considers the neighbours of each word. The method is simple, interpretable and stable. We demonstrated its effectiveness in nine different setups, considering different corpus splitting criteria (age, gender and profession of tweet authors as well as the time of the tweet) and different languages (English, French and Hebrew). This work was published at ACL 2020, the most prestigious NLP/CL venue [24].

7.6 Processing non-standard languages in extremely low-resource scenarios: towards efficient transfer learning in cross-lingual scenarios

Participants Djamé Seddah, Benoît Sagot, Benjamin Muller, Jose Rosales Nuñez.

Pushing the boundaries of noisy user-generated content (UGC) processing and, given the increasing importance of being able to tackle low-resource dialects available through social media, last year, we released the first dataset for a non-formalised North-African dialect, Arabizi, written in Latin script and which is often code-mixed with French. The resulting annotated dataset and its accompanying unlabelled data constitutes the first UGC dataset for an Arabic-related language and therefore constitutes an important milestone for the community [34]. All parsing issues raised for morphologically rich languages, UGC and noisy content are present in this dataset, which makes it a perfect *crash test* for current deep learning approaches to NLP. We are currently exploring many transfer-learning techniqes involving neural contextualised language models to address the challenges caused by this kind of highly variable dialect.

Building NLP systems for such highly variable and low-resource languages is a difficult challenge. The recent success of large-scale multilingual pretrained neural language models (including our CamemBERT language model for French) provides us with new modelling tools to tackle it. We have studied the ability of the multilingual version of BERT to model the same unseen dialect cited above (Arabizi).We have shown in different scenarios that multilingual language models are able to transfer to such an unseen dialect, specifically in two extreme cases: across scripts (Arabic to Latin) and from Maltese, a related language written in the Arabic script, unseen during pretraining. The first results have already been published [112, 89] and have led to fruitful collaborations with Yanai Elazar from Bar Illan University and Antonios Anastasopoulos from George Mason University. These works, expanding from North-African Arabic to many other languages that share similar properties, focus on the objective of understanding the inner workings of large multilingual neural language models when fine-tuned on cross-language scenarios with various degres of resource availability (from large to extremely scarce). Conducted in 2020, those works led to the submission of two papers [50, 88]. The second article was accepted to EACL 2021 and the first is currently in submission at NAACL-HLT 2021.

7.7 NLP for Old French

Participants Mathilde Regnault, Gaël Guibon, Éric Villemonte de La Clergerie, Benoît Sagot.

In the context of the ANR project PROFITEROLE, we have been studying the diachronic evolution of Old French, in particular at the lexical and syntactic levels, through the development of linguistic

resources.

In 2020, we published a new version of OFrLex, resulting from an automatic approach to the enrichment of lexical entries [26]. We also updated LEMA, an application to update lexical entries and to validate automatically enriched ones. To study the lexical diachronic evolution of French, we made available an ELMO-like neural language model trained over the PROFITEROLE corpus, and we carried out some preliminary experiments with it. Still in relation to diachronic language models, we focused on learning meta word embeddings for diachronic French lexemes, by combining Ofrlex and the PROFITEROLE corpus.

At the syntactic level, Mathilde Regnault pursued the development of MetaMOF, a meta-grammar for Old French derived from FRMG, our large coverage meta-grammar for modern French. Improvements addressed in particular clitic raising, sentence modifiers and extraposed modifiers, and of course a freer word ordering than found in modern French. Coupling with Ofrlex lexicon has also been strengthened, through an in-depth revision of closed lexical categories (such as determiners and pronouns) but also by the constitution of a temporary list of verbal entries enriched with their valency (to be incorporated properly within Ofrlex). Thanks to these evolutions, the latest experiments run in December showed a promising coverage of 83% on the PROFITEROLE corpus.

This work on MetaMOF, in particular on reusing FRMG, has triggered a new interest in meta-grammars, and, driven by Mathilde Regnault, led to the creation of Demeter, a new working group within GdR Lift. The Demeter initiative has vocation to promote the use of meta-grammars for computational linguists and field linguists, through the diffusion of resources and the development of new tools. We organised a first session during the last GdR days (December 10th and 11th 2020), during which Mathilde Regnault and Éric Villemonte de la Clergerie presented MetaMOF, FRMG, but also SMG, the underlying meta-grammar formalism.

Still at the syntactic level, Mathilde Regnault was involved in neural parsing experiments for Old French, with the submission of a paper to ACL-IJCNLP 2021.

7.8 Open science

Participants Laurent Romary, Tanti Kristanti, Lucas Foppiano.

Like many other fields, research in computational linguistics requires scholarly information and in particular research data to be shared widely within communities so that experiments and results can be compared and, when possible, reproduced. We have therefore explored the nature of the underlying scholarly tenets and the ways forward to improve the global landscape in the domain of open science.

With Jennifer Edmond from Trinity College and current director general of the EU infrastructure DARIAH, we contributed [41] in the debate on the general evolution of scholarly publishing in relation to the digital turn. Further to a survey carried out within Inria, we also addressed [64] the situation in the domain of research data and contributed to the definition of the corresponding policy within our institution.

We also analysed the importance of standards in the reusability of research data in language processing research [58], following the experience gained in our various involvements within ISO committee TC 37 (Language and terminology). These ideas have been experimented in a large-scale experiment involving the extraction of named entities in an open corpus of digital book publications in relation to the EU infrastructure OPERAS [14]. We also had the opportunity to provide a global vision of the Text Encoding Guidelines as an open science endeavour during an invited talk at the academy of sciences in Vienna [59].

Finally we addressed the issues related to new mechanisms for assessing scholarly research beyond the traditional peer review processes as we know them. We explored in particular the various aspects of post-publication peer review [57] and its implementation in the form of the Episciences journal management platform [63].

7.9 Models for the representation of lexical content

Participants Jack Bowers, Mohamed Khemakhem, Laurent Romary, Benoit Sagot.

For several years, the ALMAnaCH team has taken a leadership role in defining standards for representing lexical content, either as a result of digitising legacy dictionaries or through the creation of new lexical resources serving as a basis for computational linguistics processes. In this respect, some significant progress has been made in 2020. First, the second part of the ISO 24613 portfolio dedicated to machine readable dictionaries [76] was published and two other parts have been moved forward to the last step of the ISO process: part 3 on the modelling of etymological process and part 4 (published in January 2021), providing a reference subset of the TEI guidelines, encompassing the features standardised in parts 1 to 3. The bridge between the general ISO standardisation work and the community-based TEI guidelines has been further developed in the context of the DARIAH working group on lexical resources, whose objective consists in identifying univocal TEI-based constructs for a variety of lexical features encountered in machine readable dictionaries. The quality of the work carried out by the group (see `https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html`) has been recognised by the TEI community as a whole with the Rahtz Prize for TEI Ingenuity and implemented in several major dictionary endeavours worldwide as exemplified in [27]. Through the PhD thesis of Jack Bowers, we have also had the opportunity to demonstrate how such standardisation principles could also be part of a wider logic for representing linguistic resources in the context of a language documentation project on Mixtepec-Mixtec [42].

7.10 Information extraction from specialised collections

Participants Alix Chagué, Mohamed Khemakhem, Laurent Romary, Tanti Kristanti, Lucas Terriel, Yves Tadjo, Éric Villemonte de La Clergerie, Benoît Sagot, Floriane Chiffoleau.

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

In the continuity of Mohamed Khemakhem's PhD thesis [44], which has demonstrated the possibility to parse lexical entries from legacy dictionary content in a very fine-grained way, we investigated the applicability of the same architecture on catalogue-like textual objects as demonstrated in [38].

The occurrence of the COVID pandemic in Spring 2020 has also been an opportunity to start a project with the Parisian hospital network (APHP). One of the goals of the project (and the action that ended up being the main focus of our attention) was to see how the GROBID suite could be further expanded to parse the variety of medical reports and documents associated with a patient so that doctors can trace the precise relevant information (anamnesis, symptoms, treatments etc.). The excellent results obtained so far have led us to build up a stable collaboration with the APHP to make the GROBID workflow an essential building block of their document processing chain. Other actions initiated in Spring, such as the automatic detection of the date of the first symptoms in emergency service records, as well as the exploitation of parsing technologies to improve information extraction from medical reports, proved less immediately operational.

Following the provision of a direct grant from the Ministry of Higher Education and Research (MESRI - DAHN project) and benefiting from a collaborative framework with the French National Archives (Lectaurep project [62]), we have explored how our experience in extracting information from legacy documents could be made part of wider understanding of the components of a generic digitisation workflow of documents initially available as images. In particular, in collaboration with colleagues from the ÉPHÉ (École Pratique des Hautes Études), we contributed to the further development of the eScriptorium platform, which aims to provide an annotation, training and recognition environment for hand-written recognition tasks. We have made contributions in three complementary directions:

- We have provided a step-by-step overview of the necessary articulation between existing *standards*, whether they come from the cultural heritage institution domain (e.g. EAD), or adopted by scholarly

communities (e.g. TEI). We have also explored the role of recent endeavours such as IIIF for providing access to image corpora in repositories;

- We have defined methods for the coherent management and sharing of ground truth, i.e reference annotated data with potential interest across a variety of time periods and domains. To this end, the HTR United (https://htr-united.github.io) project was launched to gather various annotated corpora for the creation of HTR models;
- Finally, based on our experience in setting up services [16] and infrastructures [40] at the European level we started joining efforts with various parisian institution to provide a joint technical deployment of eScriptorium that would be available to a large research community (CREMMA project).

8 Bilateral contracts and grants with industry

8.1 Bilateral contracts with industry

Ongoing contracts:

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

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

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

Winespace The collaboration with this start-up company, dedicated to information extraction from wine descriptions to develop a wine recommendation system, was finally carried out in 2020, in collaboration with Inria Bordeaux's "InriaTech" structure. Funding received: 2,391 euros (not including the engineering effort, for which Inria Bordeaux received the funding)

Active collaborations without a contract:

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

Hyperlex A collaboration was initiated in 2018 on NLP and information extraction from raw legal documents (mostly PDF format), involving especially Éric de La Clergerie, who has served as a part-time employee of the company until November 2020.

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

INPI Patent classification (project will start in 2021)

Cour de cassation (in the context of the LabIA): retrieval of relevant jurisprudence (project started in 2021)

9 Partnerships and cooperations

9.1 European initiatives

9.1.1 FP7 & H2020 Projects

H2020 EHRI "European Holocaust Research Infrastructure"

Duration: 1 May 2015–31 Aug 2024.

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

Coordinator for ALMAnaCH: Laurent Romary.

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

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

9.1.2 Collaborations in European Programs, except FP7 and H2020

ERIC DARIAH

Duration: 1 Sep 2014–31 Aug 2034.

Coordinator for ALMAnaCH: Laurent Romary.

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

COST enCollect

Duration: 7 Mar 2017–1 May 2020.

PI: Lionel Nicolas (European Academy of Bozen/Bolzano).

Coordinator for ALMAnaCH: Éric de La Clergerie.

Summary: Combining language learning and crowdsourcing for developing language teaching materials and more generic language resources for NLP.

9.1.3 ANR

ANR ParSiTi

Duration: 1 Oct 2016–31 Mar 2022.

PI: Djamé Seddah.

Coordinator for ALMAnaCH: Djamé Seddah.

Partners:
- LISN (ex-LIMSI)
- LIPN

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

ANR SoSweet

Duration: 1 Oct 2015–31 Dec 2020.

PI: Jean-Philippe Magué (ICAR).

Coordinator for ALMAnaCH: Djamé Seddah.

Partners:
- ICAR (ENS Lyon, CRNS)
- Dante (Inria)

Summary: Studying sociolinguistic variability on Twitter, comparing linguistic and graph-based views on tweets.

ANR BASNUM

Duration: 1 Oct 2018–30 Jun 2023.

PI: Geoffrey Williams (Université de Grenoble).

Coordinator for ALMAnaCH: Laurent Romary.

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

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

ANR Profiterole

Duration: 1 Oct 2017–31 Jan 2021.

PI: Sophie Prévost (LaTTICe).

Coordinator for ALMAnaCH: Éric de La Clergerie.

Partners:
- LaTTICe
- LLF
- IRHIM

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

ANR TIME-US

Duration: 1 Oct 2016–31 Dec 2021.

PI: Manuela Martini (LARHRA).

Coordinator for ALMAnaCH: Éric de La Clergerie.

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

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

9.1.4 Competitivity Clusters and Thematic Institutes

3IA PRAIRIE

Duration: 1 Oct 2019–30 Sep 2024.

PI: Isabelle Ryl.

Coordinator for ALMAnaCH: Benoît Sagot.

Partners:
- Inria, CNRS
- Institut Pasteur
- PSL
- Université de Paris
- Amazon
- Google DeepMind
- Facebook
- faurecia
- GE Healthcare
- Google
- Idemia
- Janssen

- Microsoft
- Naver Labs
- Nokia
- Pfizer
- PSA
- Uber
- 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.

LabEx EFL

Duration: 1 Oct 2010–30 Sep 2024.

PI: Barbara Hemforth (LLF).

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

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

GDR LiLT

Duration: 1 Jan 2019–present.

Summary: Linguistic issues in language technology.

9.1.5 Other National Initiatives

Informal initiative Cap'FALC

Duration: 1 Jan 2018–present.

Coordinator for ALMAnaCH: Benoît Sagot.

Partners:
- UNAPEI
- FAIR

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

Convention (MIC, Archives Nationales) LECTAUREP

Duration: 1 Jan 2018–4 Nov 2021.

PI: Laurent Romary.

Coordinator for ALMAnaCH: Laurent Romary.

Partners:
- ÉPHÉ
- Archives Nationales
- Ministère de l'information et de la communication

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

Convention (MIC, Archives Nationales) DAHN

Duration: 1 Jun 2019–30 Apr 2022.

PI: Laurent Romary.

Coordinator for ALMAnaCH: Laurent Romary.

Partners:
- ÉPHÉ
- Université du Mans, Ministère de l'information et de la communication

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

Convention (MIC, Archives Nationales) NER4archives

Duration: 1 Jan 2020–23 Sep 2021.

PI: Laurent Romary.

Coordinator for ALMAnaCH: Laurent Romary.

Partners:
- Ministère de l'information et de la communication
- Archives Nationales

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

TGIR Huma-Num

Duration: 1 Jan 2013–present.

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

PIA OPALINe

Duration: 10 Jan 2017–31 Mar 2020.

PI: Laurent Romary.

Coordinator for ALMAnaCH: Laurent Romary.

Partners:
- BrailleNet
- ERDLab
- FeniXX

Summary: Development of tools for the accessibility of digital books for visually impaired people.

9.1.6 Regional Initiatives

Framework agreement with Inria AP-TAL

Duration: 1 Apr 2020–present.

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

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

Partner:
- APHP

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

DIM Matériaux Anciens et Patrimoniaux

Duration: 1 Jan 2017–present.

PI: Étienne Anheim, Loïc Bertrand, Isabelle Rouget.

Coordinator for ALMAnaCH: Laurent Romary.

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

10 Dissemination

10.1 Promoting Scientific Activities

10.1.1 Scientific Events: Selection

Chair of Conference Program Committees

- Djamé Seddah: Programme chair for EACL 2021 (Demo track) and IWPT 2020 Enhanced Dependency Shared Task.

Reviewer and Member of the Conference Program Committees

- Clémentine Fourrier: Reviewer for ACL 2020.
- Éric de La Clergerie: Reviewer for AAAI, COLING 2020, LREC 2020, STAIRS 2020 and EMNLP 2020.
- Louis Martin: Reviewer for CHI 2020 and EMNLP 2020 (Outstanding Reviewer).
- Pedro Javier Ortiz Suárez: Subreviewer for ACL 2020 and COLING 2020.
- Laurent Romary: Reviewer for AACL 2020, COLING 2020, ACL 2020, CMLC 8, ElPub 2020, MWE-ELEX 2020 (Joint Workshop on Multiword Expressions and Electronic Lexicons), LDL 2020 (7th Workshop on Linked Data in Linguistics: Building tools and infrastructure) and TOTh 2020 (Terminology & Ontology: Theories and applications).
- Benoît Sagot: Reviewer for ACL 2020 and EACL 2021.
- Djamé Seddah: Reviewer for ACL 2020, EMNLP 2020, COLING 2020, EACL 2021 and NLP+CSS (computational social sceince).

10.1.2 Journal

Member of the Editorial Boards

- Éric de La Clergerie: Member of the editorial board for *ActuIA* (Large Audience Newspaper).
- Rachel Bawden: Member of the editorial board for *Northern European Journal of Language Technology*.

Reviewer - Reviewing Activities

- Éric de La Clergerie: Reviewer for *Natural Language Engineering* and *JDMDH (Journal of Data Mining & Digital Humanities)*.

10.1.3 Invited Talks

- Alix Chagué: ALFA Seminar on “Artificial intelligence for the analysis of the alfonsine corpus” at l’Observatoire de Paris, Paris, France (15 Dec 2020): “EScriptorium: an application for handwritten text recognition”.
- Éric de La Clergerie: GDR Lift workshop on “Linguistique informatique, formelle et de terrain” (11 Dec 2020): “SMG Simple MetaGrammars”.
- Pedro Javier Ortiz Suárez: Séminaires du Lattice, École normale supérieure / Laboratoire Lattice (22 Sep 2020): “Des Méthodes de TAL modernes pour l’Enrichissement de Documents”.
- Djamé Seddah:
  - Seminar at Axa REV (26 Apr 2020): “Deep Learning and the rise of CamemBERT”.
  - Invited talk at the Paris NLP Meetup (22 Jan 2020): “Sesame street-based naming schemes must fade out, long live CamemBERT et le French fromage!”.

10.1.4 Research Administration

- Pedro Javier Ortiz Suárez: Member of the user committee for the Jean Zay supercomputer (Invited member).
- Laurent Romary:
  - Member of the scientific board for the ELEXIS Interoperability and Sustainability Committee (ISC) (ELEXIS is the European Lexicographic Infrastructure.).

  - Member of the scientific board for the Schloss Dagstuhl Scientific Advisory Board (Website).
  - President of the scientific board for ABES (Website).
  - Member of the international advisory board for the Research Infrastructure project LINDAT/CLARIAH-CZ.

- Djamé Seddah: Member of the scientific board and Vice President of the French NLP society (ATALA).

10.2 Teaching - Supervision - Juries

10.2.1 Teaching

- Alix Chagué: Master's course (M1) as part of the Masters "Archives", "Histoire" et "Lettres et langues". Introduction à LaTeX (5 hours). Institut d'études culturelles et internationales (IECI), Université Versailles-Saint-Quentin-en-Yvelines, France.

- Alix Chagué: Master's course (M1) as part of the Masters "Archives", "Histoire" et "Lettres et langues". Introduction à Linux (5 hours). Institut d'études culturelles et internationales (IECI), Université Versailles-Saint-Quentin-en-Yvelines, France.

- Alix Chagué: Master's course (M2) as part of the Master "Technologies Numériques Appliquées à l'Histoire". Introduction à la programmation avec Python (10 hours). École nationale des chartes, France.

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

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

- Louis Martin: Master's course (M2) as part of the Master "Mathématiques, Vision Apprentissage". Speech and Language Processing (3 hours), coorganised with Vincent Lepetit. ENS Paris-Saclay, France.

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

- Clémentine Fourrier: Eq. Master's course (M1,M2) Module TDLog - Techniques de Developpement Logiciel (computer science course in Python) (18 hours). École des Ponts ParisTech, France.

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

- Loïc Grobol: Master's course (M1) as part of the pluriTAL. Introduction à la fouille de textes (39 hours). Université Sorbonne Nouvelle, France.

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

- Mathilde Regnault: Bachelor's course (L2) as part of the Licence Sciences du langage. Informatique et industries de la langue (20 hours). Université Sorbonne Nouvelle, Paris, France.

- Pedro Javier Ortiz Suárez: Bachelor's course (L2) as part of the Licence de Sciences et Technologies. Eléments de programmation 2 (None hours). Sorbonne Université, Paris, France.

- Pedro Javier Ortiz Suárez: Bachelor's course (L2) as part of the Licence de Sciences et Technologies. Mathématiques discrètes (40 hours). Sorbonne Université, Paris, France.

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

- Alix Chagué: Animated eScriptorium Tutorial, Online (1 Oct 2020): Prise en main d'eScriptorium (2.5 hours)

- Clémentine Fourrier: Animated Coding dojo, ENPC Paris (1 Sep 2020): TDLOG coding dojo, coorganised with Theodo (CS startup) (6 hours)

10.2.2 Supervision

- PhD defended: Mohamed Khemakhem, "Automatic extraction of lexical content and structure from digitized legacy dictionaries" (1 Sep 2016–31 Aug 2020). Supervised by Laurent Romary. PhD defended on 1 Oct 2020.

- PhD defended: Jack Bowers, "Corpus and lexicon for Mixtec, TEI" (1 Oct 2016–8 Oct 2020). Supervised by Laurent Romary. PhD defended on 8 Oct 2020.

- PhD defended: Loïc Grobol, "Coreference resolution" (1 Oct 2016–15 Jul 2020). Supervised by Frédéric Landragin, Marco Dinarelli and Éric de La Clergerie. PhD defended on 15 Jul 2020.

- PhD in progress: Axel Herold, "Extraction of etymological information from digital dictionaries" (1 Oct 2016–present). Supervised by Laurent Romary.

- PhD in progress: Mathilde Regnault, "Medieval French processing" (1 Oct 2017–15 Sep 2020). Supervised by Sophie Prévost and Éric de La Clergerie.

- PhD in progress: Louis Martin, "Automatic text simplification" (1 Jun 2018–present). Supervised by Benoît Sagot and Éric de La Clergerie.

- PhD in progress: José Carlos Rosales Núñez, "Machine translation for user-generated content" (1 Jun 2018–present). Supervised by Guillaume Wisniewksi and Djamé Seddah.

- PhD in progress: Pedro Javier Ortiz Suárez, "NLP and IE from 17th century encyclopedia" (1 Oct 2018–present). Supervised by Laurent Romary and Benoît Sagot.

- PhD in progress: Benjamin Muller, "NLP for social media texts" (1 Oct 2018–present). Supervised by Benoît Sagot and Djamé Seddah.

- PhD in progress: Clémentine Fourrier, "Neural architecture development, computational historical linguistics" (1 Oct 2019–present). Supervised by Laurent Romary, Benoît Sagot and Rachel Bawden.

- PhD in progress: Robin Algayres, "Unsupervised Automatic Speech Recognition in low resource conditions" (1 Oct 2019–present). Supervised by Emmanuel Dupoux and Benoît Sagot.

- PhD in progress: Lionel Tadjou Tadonfouet, "Conversations Disentanglement" (1 Mar 2020–present). Supervised by Laurent Romary and Éric de La Clergerie.

10.2.3 Juries

- Kim Gerdes: Reviewer of the PhD committee for Mohammed Galal at the University of Ain Shams, Egypt on 8 Jan 2020. Title: *Les constructions exceptives du français et de l'arabe : syntaxe et interface sémantique-syntaxe*

- Kim Gerdes: President of the PhD committee for Marie-Amélie Botalla at the Université Sorbonne Nouvelle, Paris, France on 14 May 2020. Title: *Modélisation de la production des énoncés averbaux : le cas des compléments différés*

- Benoît Sagot: Reviewer of the PhD committee for Alice Millour at the Sorbonne Université, Paris, France on 14 Dec 2020. Title: *Myriadisation de ressources linguistiques pour le traitement automatique de langues non-standardisées*
- Benoît Sagot: Examiner of the Master committee for Antoine Yang at the École Polytechnique & ENS Paris-Saclay, France on 12 Sep 2020. Title: *Leveraging large scale video data for Video Question Answering*
- Benoît Sagot: Examiner of the PhD committee for Jack Bowers at the Université Paris Sciences et Lettres on 8 Oct 2020. Title: *Language Documentation and Standards in Digital Humanities: TEI and the documentation of Mixtepec-Mixtec*
- Benoît Sagot: Reviewer of the PhD committee for Jacobo Levy Abitbol at the ENS de Lyon on 9 Jan 2020. Title: *Computational detection of socioeconomic inequalities*
- Benoît Sagot: Examiner of the Master committee for Lucas Elbert at the ENS Paris-Saclay on 10 Sep 2020. Title: *Explorations with a voice-type classifier for child-centered audio recordings*
- Alix Chagué: Examiner of the Master committee for Lucas Terriel at the Ecole nationale des chartes on 19 Oct 2020. Title: *Représenter et évaluer les données issues du traitement automatique d'un corpus de documents historiques. L'exemple de la reconnaissance des écritures manuscrites dans les répertoires de notaires du projet LectAuRep.*
- Alix Chagué: Examiner of the Master committee for Jean-Damien Genero at the Ecole nationales des chartes on 19 Oct 2020. Title: *Valoriser le traitement automatique des données : Le cas des Ouvriers des deux mondes*
- Éric de La Clergerie: Examiner of the PhD committee for Loïc Grobol at the Université Sorbonne Nouvelle, Paris, France on 15 Jul 2020. Title: *Coreference resolution for spoken French*

10.3 Popularization

10.3.1 Articles and contents

Authored article

- Floriane Chiffoleau, with the Digital Intellectuals Blog, authored an article for "Encoding an XML Tree model for my corpus". Online, 25 Mar 2020.
- Floriane Chiffoleau, with the Digital Intellectuals Blog, authored an article for "Starting a new project – Discovering its source material". Online, 31 Mar 2020.
- Floriane Chiffoleau, with the Digital Intellectuals Blog, authored an article for "Working through minor issues". Online, 7 Apr 2020.
- Floriane Chiffoleau, with the Digital Intellectuals Blog, authored an article for "How to produce a model for the transcription". Online, 7 May 2020.
- Alix Chagué, with Time Us Blog, authored an article for "Constitution d'un corpus textuel sur les monographies de le play". Online, 11 Jun 2020.
- Floriane Chiffoleau, with the Digital Intellectuals Blog, authored an article for "Difficulties in creating the transcription model". Online, 17 Jun 2020.
- Jean-Damien Genero, with Time Us Blog, authored an article for "Les ouvriers des deux mondes : des images aux urls". Online, 19 Jun 2020.
- Floriane Chiffoleau, with the Digital Intellectuals Blog, authored an article for "How to produce a model for the segmentation". Online, 17 Jul 2020.
- Floriane Chiffoleau, with the Digital Intellectuals Blog, authored an article for "Transcribing the corpus". Online, 30 Jul 2020.

- Floriane Chiffoleau, with the Digital Intellectuals Blog, authored an article for “Encoding the corpus”. Online, 10 Aug 2020.
- Lucas Terriel, with Lectaurep Blog, authored an article for “Le saviez-vous ? Les répertoires de notaires ne sont pas seulement des images numérisées !”. Online, 6 Oct 2020.
- Éric de La Clergerie authored an article for “Traitement Automatisé du Langage avec : les LSTM, BERT, GPT-3”. Journal ActuIA (numero 2), 20 Oct 2020.
- Alix Chagué, with Lectaurep Blog, authored an article for “Switching from Transkribus to eScriptorium with Aspyre”. Online, 1 Nov 2020.
- Floriane Chiffoleau, with the Digital Intellectuals Blog, authored an article for “Publication of my digital edition – Working with TEI Publisher”. Online, 4 Dec 2020.

Media interview

- Éric de La Clergerie was interviewed as part of “Chaîne TV Big Data et IA Paris (diffusion 18/12/2020)”. Paris, 14 Sep 2020.

10.3.2 Education

- Clémentine Fourrier participated in “Filles et Maths, une équation lumineuse”. Sorbonne, 1 May 2020 (2 hours).
- Clémentine Fourrier participated in “Filles et Maths, une équation lumineuse”. Online, 1 Aug 2020 (2 hours).
- Clémentine Fourrier, with Eloise Berthier, was the main organiser of “Rendez-Vous des Jeunes Mathématiciennes et Informaticiennes Inria 2020”. Inria Paris and online, 1 Oct 2020 (14 hours).
- Rachel Bawden participated in “Rendez-Vous des Jeunes Mathématiciennes et Informaticiennes Inria 2020”. Inria Paris and online, 1 Oct 2020 (1 hour).
- Pedro Javier Ortiz Suárez participated in “Rendez-Vous des Jeunes Mathématiciennes et Informaticiennes Inria 2020”. Inria Paris and online, 1 Oct 2020 (6 hours).
- Clémentine Fourrier participated in “Rendez-Vous des Jeunes Mathématiciennes et Informaticiennes ENS 2020”. Online for the ENS, 1 Dec 2020 (2 hours).

10.3.3 Interventions

- Éric de La Clergerie gave a talk at “AI Paris 2020”. Paris Porte de Versailles, 14 Sep 2020 (1 hour).
- Éric de La Clergerie gave a talk at the “Seminaire Club Speech Bertin IT”. Paris, 13 Oct 2020 (1 hour).
- Benoît Sagot gave a talk at “France is AI”. Online, 16 Nov 2020.
- Éric de La Clergerie participated in “Battle Lab Rens”. Online, 25 Nov 2020 (20 hours).
- Alix Chagué, with ADEMEC (Association des Diplômés et Etudiants des Masters de l’Ecole des chartes), was the main organiser of “ADEMEC Monthly Workshops”. Ecole nationale des chartes, 1 Dec 2020 (2 hours).

11 Scientific production

11.1 Major publications

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

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

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

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

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

[6] C. Ribeyre, É. Villemonte de La Clergerie and D. Seddah. 'Because Syntax does Matter: Improving Predicate-Argument Structures Parsing Using Syntactic Features'. In: *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Denver, USA, United States, June 2015. URL: https://hal.archives-ouvertes.fr/hal-01174533.

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

[8] B. Sagot. 'The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French'. In: *7th international conference on Language Resources and Evaluation (LREC 2010)*. Valletta, Malta, May 2010. URL: https://hal.inria.fr/inria-00521242.

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

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

[11] R. Tsarfaty, D. Seddah, Y. Goldberg, S. Kübler, Y. Versley, M. Candito, J. Foster, I. Rehbein and L. Tounsi. 'Statistical Parsing of Morphologically Rich Languages (SPMRL) What, How and Whither'. In: *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*. États-Unis Los Angeles: Association for Computational Linguistics, 2010, pp. 1–12.

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

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

11.2 Publications of the year

International journals

[14] A. Bertino, L. Foppiano, L. Romary and P. Mounier. 'Leveraging Concepts in Open Access Publications'. In: *Journal of Data Mining and Digital Humanities* 2019 (15th June 2020). URL: https://hal.inria.fr/hal-01981922.

[15] X. Chen and K. Gerdes. 'Dependency Distances and Their Frequencies in Indo-European Language'. In: *Journal of Quantitative Linguistics* (18th June 2020), pp. 1–20. DOI: 10.1080/09296174.2020.1771135. URL: https://hal.archives-ouvertes.fr/hal-03168332.

[16] L. Foppiano and L. Romary. 'Entity-fishing: a DARIAH entity recognition and disambiguation service'. In: *Journal of the Japanese Association for Digital Humanities* 5.1 (19th Nov. 2020), pp. 22–60. DOI: 10.17928/jjadh.5.1_22. URL: https://hal.inria.fr/hal-01812100.

International peer-reviewed conferences

[17] R. Algayres, M. S. Zaiem, B. Sagot and E. Dupoux. 'Evaluating the reliability of acoustic speech embeddings'. In: INTERSPEECH 2020 - Annual Conference of the International Speech Communication Association. Shanghai / Vitrtual, China, 25th Oct. 2020. URL: https://hal.inria.fr/hal-02977539.

[18] F. Alva-Manchego, L. Martin, A. Bordes, C. Scarton, B. Sagot and L. Specia. 'ASSET: A Dataset for Tuning and Evaluation of Sentence Simplification Models with Multiple Rewriting Transformations'. In: ACL 2020 - 58th Annual Meeting of the Association for Computational Linguistics. Seattle / Virtual, United States, 5th July 2020. URL: https://hal.inria.fr/hal-02889823.

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

[20] M. Fabre, P. J. Ortiz Suárez, B. Sagot and É. Villemonte de La Clergerie. 'French Contextualized Word-Embeddings with a sip of CaBeRnet: a New French Balanced Reference Corpus'. In: CMLC-8 - 8th Workshop on the Challenges in the Management of Large Corpora. Marseille, France: https://lrec2020.lrec-conf.org/media/proceedings/Workshops/Books/CMLC-8book.pdf, 16th May 2020. URL: https://hal.inria.fr/hal-02678358.

[21] C. Fourrier and B. Sagot. 'Comparing Statistical and Neural Models for Learning Sound Correspondences'. In: LT4HALA 2020 : First Workshop on Language Technologies for Historical and Ancient Languages. Marseille, France, 12th May 2020. URL: https://hal.inria.fr/hal-02529929.

[22] C. Fourrier and B. Sagot. 'Methodological Aspects of Developing and Managing an Etymological Lexical Resource: Introducing EtymDB 2.0'. In: LREC 2020 - 12th Language Resources and Evaluation Conference. Marseille, France, 11th May 2020. URL: https://hal.inria.fr/hal-02678100.

[23] S. Gabay, L. Rondeau Du Noyer and M. Khemakhem. 'Selling autograph manuscripts in 19th c. Paris: digitising the Revue des Autographes'. In: IX Convegno AIUCD. Milan, Italy: https://aiucd2020.unicatt.it, 15th Jan. 2020. URL: https://hal.archives-ouvertes.fr/hal-02388407.

[24] H. Gonen, G. Jawahar, D. Seddah and Y. Goldberg. 'Simple, Interpretable and Stable Method for Detecting Words with Usage Change across Corpora'. In: ACL 2020 - 58th Annual Meeting of the Association for Computational Linguistics. Seattle / Virtual, United States, 5th July 2020, pp. 538–555. DOI: 10.18653/v1/2020.acl-main.51. URL: https://hal.inria.fr/hal-03161637.

[25] G. Guibon, M. Courtin, K. Gerdes and B. Guillaume. 'When Collaborative Treebank Curation Meets Graph Grammars: Arborator With a Grew Back-End'. In: LREC 2020 - 12th Language Resources and Evaluation Conference. Marseille, France: http://www.lrec-conf.org/proceedings/lrec2020/index.html, 11th May 2020. URL: https://hal.inria.fr/hal-03021720.

[26] G. Guibon and B. Sagot. 'OFrLex: A Computational Morphological and Syntactic Lexicon for Old French'. In: LREC 2020 - 12th Language Resources and Evaluation Conference. Marseille, France, 11th May 2020, pp. 11–16. URL: https://hal.inria.fr/hal-02677957.

[27] F. Khan, L. Romary, A. Salgado, J. Bowers, M. Khemakhem and T. Tasovac. 'Modelling Etymology in LMF/TEI: The Grande Dicionário Houaiss da Língua Portuguesa Dictionary as a Use Case'. In: LREC 2020 - 12th Language Resources and Evaluation Conference. Marseille, France: http://www.lrec-conf.org, 11th May 2020. URL: https://hal.inria.fr/hal-02618067.

[28] T. Kristanti and L. Romary. 'DeLFT and entity-fishing : Tools for CLEF HIPE 2020 Shared Task'. In: CLEF 2020 - Conference and Labs of the Evaluation Forum. Vol. 2696. CLEF 2020 Working Notes. Thessaloniki / Virtual, Greece: http://ceur-ws.org/Vol-2696/, 22nd Sept. 2020. URL: https://hal.inria.fr/hal-02974946.

[29] 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, 5th July 2020. DOI: 10.18653/v1/2020.acl-main.645. URL: https://hal.inria.fr/hal-02889805.

[30] L. Martin, É. Villemonte de La Clergerie, B. Sagot and A. Bordes. 'Controllable Sentence Simplification'. In: LREC 2020 - 12th Language Resources and Evaluation Conference. Marseille, France: http://www.lrec-conf.org/proceedings/lrec2020/index.html, 11th May 2020. URL: https://hal.inria.fr/hal-02678214.

[31] P. J. Ortiz Suárez, Y. Dupont, G. Lejeune and T. Tian. 'SinNer@Clef-Hipe2020 : Sinful adaptation of SotA models for Named Entity Recognition in French and German'. In: CLEF 2020 Working Notes. Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum. Thessaloniki / Virtual, Greece: https://impresso.github.io/CLEF-HIPE-2020/, 21st Oct. 2020. URL: https://hal.inria.fr/hal-02984746.

[32] 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. Marseille, France: http://www.lrec-conf.org, 11th May 2020. URL: https://hal.inria.fr/hal-02617950.

[33] 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: https://acl2020.org, 5th July 2020. DOI: 10.18653/v1/2020.acl-main.156. URL: https://hal.inria.fr/hal-02863875.

[34] 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, 5th July 2020. DOI: 10.18653/v1/2020.acl-main.107. URL: https://hal.inria.fr/hal-02889804.

National peer-reviewed Conferences

[35] C. Fourrier. 'Sound change and neural networks: preliminary experiments'. In: *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 3 : Rencontre des Étudiants Chercheurs en Informatique pour le TAL*. JEP-TALN-RECITAL 2020 - 33ème Journées d'Études sur la Parole, 27ème Conférence sur le Traitement Automatique des Langues Naturelles, 22ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues. Nancy / Virtuel, France, 2020, pp. 110–122. URL: https://hal.archives-ouvertes.fr/hal-02786192.

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

[37] L. Martin, B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, E. Villemonte de la Clergerie, B. Sagot and D. Seddah. 'CAMEMBERT Contextual Language Models for French: Impact of Training Data Size andHeterogeneity'. In: *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2 : Traitement Automatique des Langues Naturelles*. JEP-TALN-RECITAL 2020 - 33ème Journées d'Études sur la Parole, 27ème Conférence sur le Traitement Automatique des Langues Naturelles, 22ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues. Nancy / Virtuel, France, 2020, pp. 54–65. URL: https://hal.archives-ouvertes.fr/hal-02784755.

Conferences without proceedings

[38] M. Khemakhem, S. Gabay, B. Joyeux-Prunel, L. Romary, L. Saint-Raymond and L. Rondeau Du Noyer. 'Information Extraction Workflow for Digitised Entry-based Documents'. In: DARIAH Annual event 2020. Zagreb / Virtual, Croatia: https://dariah-ae-2020.sciencesconf.org/, 26th May 2020. URL: https://hal.archives-ouvertes.fr/hal-02508549.

Scientific book chapters

[39] A. Baillot. 'Zahlenwahn oder Textliebe? Digitale Philologie als Disziplin und als Weltanschauung'. In: *Machines/Maschinen. Les machines dans l'espace germanique: de l'automate de Kempelen à Kraftwerk*. p. 379-388. Nantes, France, Sept. 2020. URL: https://halshs.archives-ouvertes.fr/halshs-01562486.

[40] J. Edmond, F. Fischer, L. Romary and T. Tasovac. '9. Springing the Floor for a Different Kind of Dance: Building DARIAH as a Twenty-First-Century Research Infrastructure for the Arts and Humanities'. In: *Digital Technology and the Practices of Humanities Research*. Feb. 2020, pp. 207–234. DOI: 10.11647/OBP.0192.09. URL: https://hal.inria.fr/hal-02464622.

[41] J. Edmond and L. Romary. '3. Academic Publishing'. In: *Digital Technology and the Practices of Humanities Research*. Feb. 2020, pp. 49–80. DOI: 10.11647/OBP.0192.03. URL: https://hal.inria.fr/hal-02464616.

Doctoral dissertations and habilitation theses

[42] J. Bowers. 'Language Documentation and Standards in Digital Humanities: TEI and the documentation of Mixtepec-Mixtec'. École Pratique des Hauts Études, 8th Oct. 2020. URL: https://tel.archives-ouvertes.fr/tel-03131936.

[43] L. Grobol. 'Coreference resolution for spoken French'. Université Sorbonne Nouvelle - Paris 3, 15th July 2020. URL: https://hal.archives-ouvertes.fr/tel-02928209.

[44] M. Khemakhem. 'Standard-based Lexical Models for Automatically Structured Dictionaries'. Université de Paris, 1st Oct. 2020. URL: https://tel.archives-ouvertes.fr/tel-03153438.

Reports & preprints

[45] J. Bowers. *Pathways and patterns of metaphor and metonymy in Mixtepec-Mixtec body-part terms*. 7th June 2020. URL: https://hal.inria.fr/hal-02075731.

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

[47] F. Chiffoleau. *Rapport d'avancement sur le projet DAHN (avec le soutien du MESRI)*. Inria Paris, 25th May 2020. URL: https://hal.inria.fr/hal-02619488.

[48] L. Foppiano, S. Dieb, A. Suzuki, P. Baptista de Castro, S. Iwasaki, A. Uzuki, M. G. Esparza Echevarria, Y. Meng, K. Terashima, L. Romary, Y. Takano and M. Ishii. *SuperMat: Construction of a linked annotated dataset from superconductors-related publications*. 28th Jan. 2021. URL: https://hal.inria.fr/hal-03101177.

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

[50] 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*. Oct. 2020. URL: https://hal.inria.fr/hal-03109106.

[51] B. Muller, Y. Elazar, B. Sagot and D. Seddah. *First Align, then Predict: Understanding the Cross-Lingual Ability of Multilingual BERT*. 8th Mar. 2021. URL: https://hal.inria.fr/hal-03161685.

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

[53] A. Riabi, T. Scialom, R. Keraron, B. Sagot, D. Seddah and J. Staiano. *Synthetic Data Augmentation for Zero-Shot Cross-Lingual Question Answering*. 13th Jan. 2021. URL: https://hal.inria.fr/hal-03109187.

[54] E. Tóth-Czifra and L. Romary. *The Heritage Data Reuse Charter: from principles to research workflows*. 12th Feb. 2020. URL: https://halshs.archives-ouvertes.fr/halshs-02475692.

[55] N. Truan and L. Romary. *Building, Encoding, and Annotating a Corpus of Parliamentary Debates in XML-TEI: A Cross-Linguistic Account*. 18th June 2020. URL: https://halshs.archives-ouvertes.fr/halshs-03097333.

Other scientific publications

[56] J.-D. Généro. *Le corpus des Ouvriers des deux mondes : des images et des URLs*. 19th June 2020. URL: https://hal.archives-ouvertes.fr/hal-03118736.

[57] L. Romary. *An editorial and technical journey into Post Publication Peer Review (PPPR)*. Berlin / Virtual, Germany, 2nd Oct. 2020. URL: https://hal.inria.fr/hal-02960535.

[58] L. Romary. *Multilingual content management and standards with a view on AI developments*. Turin / Virtual, Italy, 6th Oct. 2020. URL: https://hal.inria.fr/hal-02961857.

[59] L. Romary. *TEI guidelines: born to be open*. Vienne, Austria, 10th June 2020. URL: https://hal.inria.fr/hal-02864525.

[60] L. Terriel. *Le saviez-vous ? Les répertoires de notaires ne sont pas seulement des images numérisées !* 6th Oct. 2020. URL: https://hal.archives-ouvertes.fr/hal-03138879.

11.3 Other

Scientific popularization

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

[62] A. Chagué, L. Terriel and L. Romary. *Des images au texte : LECTAUREP, un projet de reconnaissance automatique d'écriture*. Lille / Virtual, France, 18th Nov. 2020. URL: https://hal.archives-ouvertes.fr/hal-03008579.

[63] L. Romary. 'Découpler gestion des manuscrits de publication et évaluation par les pairs : la plateforme de gestion de revues Épisciences'. In: *I2D – Information, données & documents* 2 (2020). URL: https://hal.inria.fr/hal-03033488.

[64] L. Romary. *Les données de la recherche*. Nancy / Virtual, France, 12th Nov. 2020. URL: https://hal.inria.fr/hal-03006187.

11.4 Cited publications

[65] F. Alva-Manchego, L. Martin, C. Scarton and L. Specia. 'EASSE: Easier Automatic Sentence Simplification Evaluation'. In: *EMNLP-IJCNLP 2019 - Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (demo session)*. Hong Kong, China, Nov. 2019, pp. 49–54. URL: https://hal.inria.fr/hal-02272950.

[66] 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.

[67] 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.

[68] 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.

[69] 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.

[70] 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.

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

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

[73] 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.

[74] 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.

[75] 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.

[76] *Language resource management — Lexical markup framework (LMF) — Part 2: Machine-readable dictionary (MRD) model*. Standard. Geneva, CH: International Organization for Standardization, 2020.

[77] 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.

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

[79] 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.

[80] 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.

[81] A. Lavie and M. J. Denkowski. 'The Meteor metric for automatic evaluation of machine translation'. In: *Machine Translation* 23.2-3 (2009), pp. 105–115.

[82] 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.

[83] 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).

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

[85] 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.

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

[87] 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.

[88] B. Muller, Y. Elazar, B. Sagot and D. Seddah. *First Align, then Predict: Understanding the Cross-Lingual Ability of Multilingual BERT*. 2021. eprint: arXiv:2101.11109.

[89] B. Muller, B. Sagot and D. Seddah. *Can Multilingual Language Models Transfer to an Unseen Dialect? A Case Study on North African Arabizi*. 2020. eprint: arXiv:2005.00318.

[90] 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.

[91] K. Papineni, S. Roukos, T. Ward and W.-J. Zhu. 'Bleu: a Method for Automatic Evaluation of Machine Translation'. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA, July 2002, pp. 311–318. DOI: 10.3115/1073083.1073135. URL: https://www.aclweb.org/anthology/P02-1040.

[92] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee and L. Zettlemoyer. 'Deep contextualized word representations'. In: *Proc. of NAACL*. 2018.

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

[94] L. Rello, R. Baeza-Yates, S. Bott and H. Saggion. 'Simplify or help?: text simplification strategies for people with dyslexia'. In: *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*. ACM. 2013, p. 15.

[95] L. Rello, R. Baeza-Yates, L. Dempere-Marco and H. Saggion. 'Frequent words improve readability and short words improve understandability for people with dyslexia'. In: *IFIP Conference on Human-Computer Interaction*. Springer. 2013, pp. 203–219.

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

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

[98] A. M. Rush, R. Reichart, M. Collins and A. Globerson. 'Improved Parsing and POS Tagging Using Inter-sentence Consistency Constraints'. In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. EMNLP-CoNLL '12. Jeju Island, Korea, 2012, pp. 1434–1444.

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

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

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

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

[103] B. Sagot, D. Nouvel, V. Mouilleron and M. Baranes. 'Extension dynamique de lexiques morphologiques pour le français à partir d'un flux textuel'. In: *TALN - Traitement Automatique du Langage Naturel*. Les sables d'Olonne, France, June 2013, pp. 407–420. URL: https://hal.inria.fr/hal-00832078.

[104] C. Scarton, M. De Oliveira, A. Candido Jr, C. Gasperin and S. M. Aluísio. 'SIMPLIFICA: a tool for authoring simplified texts in Brazilian Portuguese guided by readability assessments'. In: *Proceedings of the NAACL HLT 2010 Demonstration Session*. Association for Computational Linguistics. 2010, pp. 41–44.

[105] Y. Scherrer and B. Sagot. 'A language-independent and fully unsupervised approach to lexicon induction and part-of-speech tagging for closely related languages'. In: *Language Resources and Evaluation Conference*. European Language Resources Association. Reykjavik, Iceland, May 2014. URL: https://hal.inria.fr/hal-01022298.

[106] S. Schuster, É. Villemonte de La Clergerie, M. Candito, B. Sagot, C. D. Manning and D. Seddah. 'Paris and Stanford at EPE 2017: Downstream Evaluation of Graph-based Dependency Representations'. In: *EPE 2017 - The First Shared Task on Extrinsic Parser Evaluation*. Proceedings of the 2017 Shared Task on Extrinsic Parser Evaluation. Pisa, Italy, Sept. 2017, pp. 47–59. URL: https://hal.inria.fr/hal-01592051.

[107] 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.

[108] 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.

[109] 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.

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

[111] 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.

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

[113] S. Štajner, M. Popović, H. Saggion, L. Specia and M. Fishel. 'Shared task on quality assessment for text simplification'. In: *qats2016: LREC 2016 Workshop & Shared Task on Quality Assessment for Text Simplification (QATS)*. Portorož, Slovenia, 2016, pp. 22–31. URL: https://madoc.bib.uni-mannheim.de/41134/.

[114] 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.

[115] 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.

[116] 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.