RESEARCH CENTRE

Nancy - Grand Est

IN PARTNERSHIP WITH:

CNRS, Université de Lorraine

2020

ACTIVITY REPORT

Project-Team

SEMAGRAMME

Semantic Analysis of Natural Language

IN COLLABORATION WITH: Laboratoire lorrain de recherche en informatique et ses applications (LORIA)

DOMAIN

Perception, Cognition and Interaction

THEME

Language, Speech and Audio

Contents

Project-Team SEMAGRAMME

Creation of the Team: 2011 January 01, updated into Project-Team: 2013 July 01

Keywords

Computer sciences and digital sciences

A5.8. – Natural language processing

A7.2. – Logic in Computer Science

A9.4. – Natural language processing

Other research topics and application domains

B9.6.8. – Linguistics

1 Team members, visitors, external collaborators

Research Scientists

- Philippe de Groote [Team leader, Inria, Senior Researcher]
- Bruno Guillaume [Inria, Researcher]
- Sylvain Pogodalla [Inria, Researcher]

Faculty Members

- Maxime Amblard [Univ de Lorraine, Associate Professor, HDR]
- Michel Musiol [Inria secondment, Univ de Lorraine, Professor, HDR]
- Guy Perrier [Univ de Lorraine, Emeritus, HDR]

Post-Doctoral Fellow

- Marc Anderson [Univ de Lorraine, from Nov 2020]

PhD Students

- William Babonnaud [Univ de Lorraine]
- Clément Beysson [Univ de Lorraine, until Aug 2020]
- Maria Boritchev [Inria, until Aug 2020, Univ de Lorraine, from Sep 2020]
- Samuel Buchel [Inria]
- Amandine Lecomte [Inria, from Oct 2020]
- Chuyuan Li [Univ de Lorraine]
- Pierre Ludmann [Univ de Lorraine]
- Siyana Pavlova [Univ de Lorraine, from Oct 2020]
- Priyansh Trivedi [Inria, from Nov 2020]

Technical Staff

- Amandine Lecomte [Inria, Engineer, until May 2020]
- Pierre Lefebvre [Inria, Engineer]

Interns and Apprentices

- Hee-Soo Choi [Univ de Lorraine, from Jun 2020 until Aug 2020]
- Lucille Dumont [Univ de Lorraine, from May 2020 until Jul 2020]
- Louis Gleyo [Univ de Lorraine, from Jun 2020 until Jul 2020]
- Maxime Guillaume [Yseop, from Mar 2020 until Aug 2020]
- Morgane Pailler [Univ de Lorraine, from May 2020 until Jul 2020]
- Angeline Pintore [Univ de Lorraine, from May 2020 until Jul 2020]
- Clara Serruau [Univ de Lorraine, from Sep 2020]
- Vincent Tourneur [Univ de Lorraine, from Mar 2020 until Jul 2020]

Administrative Assistants

- Isabelle Herlich [Inria]
- Delphine Hubert [Univ de Lorraine]

External Collaborator

- Karën Fort [Sorbonne Université]

2 Overall objectives

2.1 Scientific Context

Computational linguistics is a discipline at the intersection of computer science and linguistics. On the theoretical side, it aims to provide computational models of the human language faculty. On the applied side, it is concerned with natural language processing and its practical applications.

From a structural point of view, linguistics is traditionally organized into the following sub-fields:

- Phonology, the study of language abstract sound systems.
- Morphology, the study of word structure.
- Syntax, the study of language structure, i.e., the way words combine into grammatical phrases and sentences.
- Semantics, the study of meaning at the levels of words, phrases, and sentences.
- Pragmatics, the study of the ways in which the meaning of an utterance is affected by its context.

Computational linguistics is concerned by all these fields. Consequently, various computational models, whose application domains range from phonology to pragmatics, have been developed. Among these, logic-based models play an important part, especially at the "highest" levels.

At the level of syntax, generative grammars may be seen as basic inference systems, while categorial grammars are based on substructural logics specified by Gentzen sequent calculi. Finally, model-theoretic grammars amount to sets of logical constraints to be satisfied.

At the level of semantics, the most common approaches derive from Montague grammars, which are based on the simply typed λ-calculus and Church's simple theory of types. In addition, various logics (modal, hybrid, intensional, higher-order...) are used to express logical semantic representations.

At the level of pragmatics, the situation is less clear. The word *pragmatics* has been introduced by Morris to designate the branch of philosophy of language that studies, besides linguistic signs, their relation to their users and the possible contexts of use. The definition of pragmatics was not quite precise, and, for a long time, several authors have considered (and some authors are still considering) pragmatics as the wastebasket of syntax and semantics. Nevertheless, as far as discourse processing is concerned (which includes pragmatic problems such as pronominal anaphora resolution), logic-based approaches have also been successful. In particular, Kamp's Discourse Representation Theory gave rise to sophisticated 'dynamic' logics. The situation, however, is less satisfactory than it is at the semantic level. On the one hand, we are facing a kind of logical "tower of Babel". The various pragmatic logic-based models that have been developed, while sharing underlying mathematical concepts, differ in several respects and are too often based on *ad hoc* features. As a consequence, they are difficult to compare and appear more as competitors than as collaborative theories that could be integrated. On the other hand, several phenomena related to discourse dynamics (e.g., context updating, presupposition projection and accommodation, contextual reference resolution...) are still lacking deep logical explanations. We strongly believe, however, that this situation can be improved by applying to pragmatics the same approach Montague applied to semantics, using the standard tools of mathematical logic.

Accordingly:

The overall objective of the Sémagramme project is to design and develop new unifying logic-based models, methods, and tools for the semantic analysis of natural language utterances and discourses. This includes the logical modeling of pragmatic phenomena related to discourse dynamics. Typically, these models and methods will be based on standard logical concepts (stemming from formal language theory, mathematical logic, and type theory), which should make them easy to integrate.

The project is organized along three research directions (i.e., *syntax-semantics interface*, *discourse dynamics*, and *common basic resources*), which interact as explained below.

2.2 Syntax-Semantics Interface

The Sémagramme project intends to focus on the semantics of natural languages (in a wider sense than usual, including some pragmatics). Nevertheless, the semantic construction process is syntactically guided, that is, the constructions of logical representations of meaning are based on the analysis of the syntactic structures. We do not want, however, to commit ourselves to such or such specific theory of syntax. Consequently, our approach should be based on an abstract generic model of the syntax-semantic interface.

Here, an important idea of Montague comes into play, namely, the "homomorphism requirement": semantics must appear as a homomorphic image of syntax. While this idea is almost a truism in the context of mathematical logic, it remains challenged in the context of natural languages. Nevertheless, Montague's idea has been quite fruitful, especially in the field of categorial grammars, where van Benthem showed how syntax and semantics could be connected using the Curry-Howard isomorphism. This correspondence is the keystone of the syntax-semantics interface of modern type-logical grammars. It also motivated the definition of our own Abstract Categorial Grammars [33].

Technically, an Abstract Categorial Grammar simply consists of a (linear) homomorphism between two higher-order signatures. Extensive studies have shown that this simple model allows several grammatical formalisms to be expressed, providing them with a syntax-semantics interface for free [34], [4].

We intend to carry on with the development of the Abstract Categorial Grammar framework. At the foundational level, we will define and study possible type theoretic extensions of the formalism, in order to increase its expressive power and its flexibility. At the implementation level, we will continue the development of an Abstract Categorial Grammar support system.

As said above, considering the syntax-semantics interface as the starting point of our investigations allows us not to be committed to some specific syntactic theory. The Montagovian syntax-semantics interface, however, cannot be considered to be universal. In particular, it does not seem to be well adapted to dependency and model-theoretic grammars. Consequently, in order to be as generic as possible, we intend to explore alternative models of the syntax-semantics interface. In particular, we will explore relational models where several distinct semantic representations can correspond to the same syntactic structure.

2.3 Discourse Dynamics

It is well known that the interpretation of a discourse is a dynamic process. Take a sentence occurring in a discourse. On the one hand, it must be interpreted according to its context. On the other hand, its interpretation affects this context, and must therefore result in an updating of the current context. For this reason, discourse interpretation is traditionally considered to belong to pragmatics. The cut between pragmatics and semantics, however, is not that clear.

As we mentioned above, we intend to apply to some aspects of pragmatics (mainly, discourse dynamics) the same methodological tools Montague applied to semantics. The challenge here is to obtain a completely compositional theory of discourse interpretation, by respecting Montague's homomorphism requirement. We think that this is possible by using techniques coming from programming language theory, in particular, continuation semantics, and the related theories of functional control operators.

We have indeed successfully applied such techniques in order to model the way quantifiers in natural languages may dynamically extend their scope [32]. We intend to tackle, in a similar way, other dynamic phenomena (typically, anaphora and referential expressions, presupposition, modal subordination...).

What characterizes these different dynamic phenomena is that their interpretations need information to be retrieved from a current context. This raises the question of the modeling of the context itself. At a foundational level, we have to answer questions such as the following. What is the nature of the information to be stored in the context? What are the processes that allow implicit information to be inferred from the context? What are the primitives that allow a context to be updated? How does the structure of the discourse and the discourse relations affect the structure of the context? These questions also raise implementation issues. What are the appropriate datatypes? How can we keep the complexity of the inference algorithms sufficiently low?

2.4 Common Basic Resources

Even if our research primarily focuses on semantics and pragmatics, we nevertheless need syntax. More precisely, we need syntactic trees to start with. We consequently need grammars, lexicons, and parsing algorithms to produce such trees. During the last years, we have developed the notion of interaction grammar [39] and graph rewriting [1, 2] as models of natural language syntax. This includes the development of grammars for French [41], together with morpho-syntactic lexicons. We intend to continue this line of research and development. In particular, we want to increase the coverage of our grammars for French, and provide our parsers with more robust algorithms.

Further primary resources are needed in order to put at work a computational semantic analysis of utterances and discourses. As we want our approach to be as compositional as possible, we must develop lexicons annotated with semantic information. This opens the quite wide research area of lexical semantics.

Finally, when dealing with logical representations of utterance interpretations, the need for inference facilities is ubiquitous. Inference is needed in the course of the interpretation process, but also to exploit the result of the interpretation. Indeed, an advantage of using formal logic for semantic representations is the possibility of using logical inference to derive new information. From a computational point of view, however, logical inference may be highly complex. Consequently, we need to investigate which logical fragments can be used efficiently for natural language oriented inference.

3 Research program

3.1 Overview

The research program of Sémagramme aims to develop models based on well-established mathematics. We seek two main advantages from this approach. On the one hand, by relying on mature theories, we have at our disposal sets of mathematical tools that we can use to study our models. On the other hand, developing various models on a common mathematical background will make them easier to integrate, and will ease the search for unifying principles.

The main mathematical domains on which we rely are formal language theory, symbolic logic, and type theory.

3.2 Formal Language Theory

Formal language theory studies the purely syntactic and combinatorial aspects of languages, seen as sets of strings (or possibly trees or graphs). Formal language theory has been especially fruitful for the development of parsing algorithms for context-free languages. We use it, in a similar way, to develop parsing algorithms for formalisms that go beyond context-freeness. Language theory also appears to be very useful in formally studying the expressive power and the complexity of the models we develop.

3.3 Symbolic Logic

Symbolic logic (and, more particularly, proof-theory) is concerned with the study of the expressive and deductive power of formal systems. In a rule-based approach to computational linguistics, the use of symbolic logic is ubiquitous. As we previously said, at the level of syntax, several kinds of grammars (generative, categorial...) may be seen as basic deductive systems. At the level of semantics, the meaning

of an utterance is captured by computing (intermediate) semantic representations that are expressed as logical forms. Finally, using symbolic logics allows one to formalize notions of inference and entailment that are needed at the level of pragmatics.

3.4 Type Theory and Typed lambda-Calculus

Among the various possible logics that may be used, Church's simply typed λ-calculus and simple theory of types (a.k.a. higher-order logic) play a central part. On the one hand, Montague semantics is based on the simply typed λ-calculus, and so is our syntax-semantics interface model. On the other hand, as shown by Gallin, the target logic used by Montague for expressing meanings (i.e. his intensional logic) is essentially a variant of higher-order logic featuring three atomic types (the third atomic type standing for the set of possible worlds).

4 Application domains

4.1 Deep Semantic Analysis

Our applicative domains concern natural language processing applications that rely on a deep semantic analysis. For instance, one may cite the following ones:

- textual entailment and inference,
- dialogue systems,
- semantic-oriented query systems,
- content analysis of unstructured documents,
- text transformation and automatic summarization,
- (semi) automatic knowledge acquisition.

4.2 Text Transformation

Text transformation is an application domain featuring two important sub-fields of computational linguistics:

- parsing, from surface form to abstract representation,
- generation, from abstract representation to surface form.

Text simplification or automatic summarization belong to that domain.

We aim at using the framework of Abstract Categorial Grammars we develop to this end. It is indeed a reversible framework that allows both parsing and generation. Its underlying mathematical structure of λ-calculus makes it fit with our type-theoretic approach to discourse dynamics modeling.

5 New software and platforms

5.1 New software

5.1.1 ACGtk

Name: Abstract Categorial Grammar Development Toolkit

Keywords: Natural language processing, NLP, Syntactic analysis, Semantics

Scientific Description: Abstract Categorial Grammars (ACG) are a grammatical formalism in which grammars are based on typed lambda-calculus. A grammar generates two languages: the abstract language (the language of parse structures), and the object language (the language of the surface forms, e.g., strings, or higher-order logical formulas), which is the realization of the abstract language.

ACGtk provides two software tools to develop and to use ACGs: acgc, which is a grammar compiler, and acg, which is an interpreter of a command language that allows one, in particular, to parse and realize terms.

Functional Description: ACGtk provides softwares for developing and using Abstract Categorial Grammars (ACG).

Release Contributions: This version fixes some bugs, including for the Opam package distribution. It also prepare supporting probabilistic ACG and Datalog Magic Set rewriting to optimize parsing.

News of the Year: The new version removes dependencies to obsolete libraries. It improves the command line interface and prepares the integration of new functionalities and optimizations.

URL: http://acg.loria.fr/

Publications: hal-01242154, hal-01328702, tel-01412765, inria-00112956, inria-00100529

Contacts: Philippe de Groote, Sylvain Pogodalla

Participants: Philippe de Groote, Jiri Marsik, Sylvain Pogodalla, Sylvain Salvati

5.1.2 Grew

Name: Graph Rewriting

Keywords: Semantics, Syntactic analysis, Natural language processing, Graph rewriting

Functional Description: Grew is a Graph Rewriting tool dedicated to applications in NLP. Grew takes into account confluent and non-confluent graph rewriting and it includes several mechanisms that help to use graph rewriting in the context of NLP applications (built-in notion of feature structures, parametrization of rules with lexical information).

News of the Year: In 2020, the Grew software version 1.4 was released. In this version, the syntax of pattern were enriched and the loading mechanism of CoNLL data was re-implemented (CoNLL is a format used in many syntactic annotation project but it is not officially defined, A more robust way of dealing with the format was implemented to be able to deal with a large set of usage of extensions of CoNLL).

The Grew-match tool (http://match.grew.fr) is an online service available where a user can query different corpora with graph matching requests. All UD corpora (183 in 104 different languages in v2.7) are available and data from several other projects can also be queried. In 2020, 114,000 requests were received on the Grew-match server.

Grew is used in a new software Arborator-Grew (https://arborator.github.io/). See https://hal.inria.fr/hal-03021720v1

URL: http://grew.fr/

Publications: hal-01930591, hal-01814386, hal-03021720

Contacts: Bruno Guillaume, Guy Perrier

Participants: Bruno Guillaume, Guy Perrier, Guillaume Bonfante

5.1.3 SLODiM

Name: SLODiM

Keywords: Natural language processing, Discourse, Dialogue, French

Functional Description: SLODiM is a software package for the analysis of oral French. It is more particularly developed to allow the analysis of interviews with clinicians in order to identify language behaviours characteristic of mental pathologies.

Release Contributions: first complete version

URL: https://team.inria.fr/semagramme/odim/

Contacts: Maxime Amblard, Pierre Lefebvre

Partners: Loria, Université de Lorraine, CNRS

6 New results

6.1 Syntax-Semantics Interface

Participants William Babonnaud, Philippe de Groote, Maxime Guillaume, Pierre Ludmann, Sylvain Pogodalla, Maxime Amblard, Bruno Guillaume, Siyana Pavlova.

6.1.1 Abstract Categorial Grammars

Feature Structure ACG has proven to be a powerful framework with well-defined theoretical properties. It was however lacking a facility which is useful and widely used for grammar engineering: feature structures. The latter are often used to express in a concise way some combinatorial properties related to morphosyntactic properties of expressions, for instance subject-verb agreement.

We worked on extending the ACG type system to provide such feature structures. This extension relies on a restricted addition of product (records) and dependent types. We also considered the reduction of grammars using this extension to Datalog programs (which is used to implement ACG parsing in ACGtk, see Sec. 5).

Probabilistic ACG (pACG) Symbolic parsing with large coverage grammars usually leads to combinatorial explosion of syntactic ambiguities (a single expression has many syntactic analysis). Whereas people easily disambiguate such expressions, often without even noticing, automatic systems need to use additional information. The latter is usually provided in terms of probabilities or weights associated to parse structures. We worked on endowing ACG with such a mechanism using probabilistic tree automata[36]. This allowed us to characterize minimal reduced pACG as simple probabilistic context-free formalisms, and to encode pCFG[43, 26] and pTAG[44, 42, 28] into pACG.

6.1.2 Lexical Semantics

The lexicon model underlying Montague semantics is an enumerative model that would assign a meaning to each atomic expression. This model does not exhibit any interesting structure. In particular, polysemy problems are considered as homonymy phenomena: a word has as many lexical entries as it has senses, and the semantic relations that might exist between the different meanings of a same word are ignored. To overcome these problems, models of generative lexicons have been proposed in the literature. Implementing these generative models in the realm of the typed λ-calculus necessitates a calculus with notions of subtyping and type coercion. In this context, William Babonnaud and Philippe de Groote have developed a simply-typed λ-calculus dedicated to the treatment of the lexical phenomena of restrictive selection and type coercion. This calculus features records and record types, subtyping

through explicit coercion, and bounded polymorphism. They have shown that coercion inference is decidable and discussed the canonicity of the inferred solutions [8].

6.1.3 Graph-based Semantics

Siyana Pavlova started her PhD in November 2020. She began to study and compare different existing semantic graph-based annotation frameworks (AMR, UCCA and DRS). The goal is to determine how these frameworks are compatible and if they encode the same level of semantic information. Clara Serruau is working on the same topic with a focus of DRS annotation available in the Parallel Meaning Bank (https://pmb.let.rug.nl/).

6.2 Discourse Dynamics

Participants Maxime Amblard, Maria Boritchev, Philippe de Groote, Bruno Guillaume, Pierre Ludmann, Michel Musiol.

6.2.1 Dialogue Modeling

Maxime Amblard and Maria Boritchev pursue the development of a dynamic model of dialogue for questions and answers. Formal studies of discourse raise numerous interrogations on the nature and the definition of the way consecutive sentences combine with one another. The shift from discourse to dialogue brings forward even more specific issues. Dialogue acts are more intrinsically connected because of the dynamicity of the interaction. In [9] they introduce a proof of concept of a formal compositional treatment of the relationship between consecutive utterances. Starting from neo-Davidsonian event semantics, we propose to use the relative response set as an intermediate set tool that allows us to define notions of question-answer correspondence, model the effect of clarification requests on previous utterances and compute semantic representations of dialogue interactions. In this perspective, they finished the development of the DinG corpus.

Maxime Amblard continue a common work with Chloé Braud on Formal and Statistical Modelling of dialogues. In the PhD thesis of Chuyuan Li, they design tools to automatically retrieve characteristic features of dialogues. They present preliminary results in [16]. Dealing with human-human dialogues makes for a realistic situation, but it calls for strategies to represent context and to face data sparsity. They highlight the biases in the model and argue for future developments delixicalised.

6.2.2 Dialogue Dynamics

Maria Boritchev and Philippe de Groote have developped a dynamic model of dialogue [10]. This model is based on insights and ideas developed by Jonathan Ginzburg [38]. It takes advantage of inquisitive semantics [29], which allows to model both declarative and interrogative sentences in a uniform way. It appeals to ideas derived from classical epistemic logic in order to model the knowledge states of the dialogue participants, and includes a context-updating mechanism based on the type-theoretic dynamic logic developed in [40].

6.2.3 Pathological Discourse Modelling

Michel Musiol has obtained a full-time delegation in the Sémagramme team. This proximity makes it possible to set up a more active collaboration on the issue of pathological discourse modeling. He has worked on the development of the possibility of testing his conjectures on the cognitive and psychopathological profile of the interlocutors, in addition to information provided by the model of ruptures and incongruities in pathological discourse. This methodological system makes it possible to discuss, or even evaluate, the heuristic potential of the computational models developed on the basis of empirical facts.

Maxime Amblard and Michel Musiol were awarded by an Inria Exploratory Action on this issues ODiM. This year we recruited the project's collaborators. In addition, they started the constitution of a new resource and a new tool SLoDIM. The theoretical work focused on the formal definition of transactions in

dialogue. To do so, with Samuel Buchel and Amandine Lecomte, they introduce a dynamic definition of back channel words which are used to classify the dialogue units. With Manuel Rebuschi they proposed a survey of linguistic modeling of dialogue including patients with schizophrenia [19].

6.3 Common Basic Resources

Participants Maxime Amblard, Clément Beysson, Philippe de Groote, Bruno Guillaume, Guy Perrier, Sylvain Pogodalla, Karën Fort.

6.3.1 FR-FraCas

Maxime Amblard, Clement Beysson, Philippe de Groote, Bruno Guillaume and Sylvain Pogodalla carried on the development of FR-FraCas, a French version of the FraCas test suite [31] which is an inference test suite, in English, for evaluating the inferential competence of different NLP systems and semantic theories. There currently exists a multilingual version of the resource for Farsi, German, Greek, and Mandarin. Sémagramme completed the first translation into French of the test suite. The latter has been publicly released [1].

In [7], the French version of the FraCaS test suite is presented. The paper describes linguistic choices that had to be made when translating the FraCaS test suite in French, and discusses some of the issues that were raised by the translation. It also reports an experiment ran with 18 French native speakers in order to test both the translation and the logical semantics underlying the problems of the test suite. Such an experiment provides a way of checking the hypotheses made by formal semanticists against the actual semantic capacity of speakers (in the present case, French speakers), and allows us to compare the obtained results with the ones of similar experiments that have been conducted for other languages [30, 27].

During her internship, Morgane Pailler builds the syntactic annotation of the full French version of the test suite and propose a semantic interpretation of a subset of the test suite. Her work is will be the starting point of the next work planned on the test suite.

6.3.2 Universal Dependencies and Surface Syntactic Universal Dependencies

The Universal Dependencies project (UD) aims at building a syntactic dependency scheme which allows for similar analyses for several different languages. Bruno Guillaume and Guy Perrier are active in the UD community, and participate to the development and the improvement of the French data in this international initiative.

During 2020, they continue working, in collaboration with Sylvain Kahane, Kim Gerdes and their teams on the promotion of the Surface Syntactic Universal Dependencies (SUD) framework. SUD is an annotation scheme for syntactic dependency treebanks, that is almost isomorphic to UD (Universal Dependencies). Contrary to UD, it is based on syntactic criteria (favoring functional heads) and the relations are defined on distributional and functional bases [37]

A website was built to present the framework (guidelines, data)[2]. The Sémagramme teams is notably in charge of the GREW-based tools for conversion with the UD framework. These conversion tools are used both to produce the UD data for a few SUD native treebanks and of to produce the SUD version of all UD available data.

During her internship in summer 2020, Hee-Soo Choi worked on linguistic typology based on UD annotated data. She has used Grew to enrich the UD annotations and studied the respective word order of verbs with thier subjects and objects on 74 languages and compare with other linguistic works. The work will be submitted to a conference in February 2021.

[1] https://gitlab.inria.fr/semagramme-public-projects/resources/french-fracas
[2] https://surfacesyntacticud.github.io/

6.3.3 Rigor Mortis

In [11], Karën Fort, Bruno Guillaume, Mathieu Constant, Nicolas Lefèbvre and Yann-Alan Pilatte present Rigor Mortis, a gamified crowdsourcing platform[3] designed to evaluate the intuition of the speakers, then train them to annotate multi-word expressions (MWEs) in French corpora. They previously showed that the speakers' intuition is reasonably good (65% in recall on non-fixed MWE) [35]. After a training phase using some of the tests developed in the PARSEME-FR project, they obtain 0.685 in F-measure at an experimentally determined 25% threshold (number of players who annotated the same segment).

6.3.4 PARSEME

Mathieu Constant and Bruno Guillaume participates to the PARSEME-FR projects. With other researchers implied in the project, in [6], they present the enrichment of a French treebank of various genres with a new annotation layer for multiword expressions (MWEs) and named entities (NEs). The contribution with respect to previous work on NE and MWE annotation is the particular care taken to use formal criteria, organized into decision flowcharts, shedding some light on the interactions between NEs and MWEs. Moreover, in order to cope with the well-known difficulty to draw a clear-cut frontier between compositional expressions and MWEs, sufficient criteria only were chosen. As a result, annotated MWEs satisfy a varying number of sufficient criteria, accounting for the scalar nature of the MWE status. In addition to the span of the elements, annotation includes the subcategory of NEs (e.g., person, location) and one matching sufficient criterion for non-verbal MWEs (e.g., lexical substitution). The 3,099 sentences of the treebank were double-annotated and adjudicated, with attention to cross-type consistency and compatibility with the syntactic layer. Overall inter-annotator agreement on non-verbal MWEs and NEs reached 71.1%. The annotated corpus is released on http://hdl.handle.net/11234/1-3429.

Bruno Guillaume was a one of the organisers of the edition 1.2 of the PARSEME Shared Task[4] on Semi-supervised Identification of Verbal Multiword Expressions [15] presents. Lessons learned from previous editions indicate that VMWEs have low ambiguity, and that the major challenge lies in identifying test instances never seen in the training data. Therefore, this edition focuses on unseen VMWEs. The organisers have split annotated corpora so that the test corpora contain around 300 unseen VMWEs, and they provide non-annotated raw corpora to be used by complementary discovery methods. Annotated (http://hdl.handle.net/11234/1-3367) and raw (http://hdl.handle.net/11234/1-3416) corpora were released in 14 languages. The semi-supervised challenge attracted 7 teams who submitted 9 system results. The paper describes the effort of corpus creation, the task design, and the results obtained by the participating systems, especially their performance on unseen expressions.

6.3.5 Less-resourced languages

Karën Fort continued working with her PhD student, Alice Millour, on crowdsourcing for less-resourced languages, especially with no standard orthography. This allowed for the publication of two papers, one is a state of the art of language resources for non-standardized languages [13], the other one is a replication experiment [17] concerning the development of deep-learning-based tagger for Alsatian.

6.3.6 enetCollect

The enetCollect COST action produced a reflexion on crowdsourcing for language learning which was published at LREC 2020 [14].

Another result, concerning a European survey on the usage of crowdsourcing by teachers, was published in a journal [5].

6.3.7 DinG

Maria Boritchev and Maxime Amblard finished the development of DinG. Ding is a transcription corpus of oral French, based on multilogues between 3 to 4 people playing the board game *Catan*. It was created to

[3] http://rigor-mortis.org/
[4] http://multiword.sourceforge.net/sharedtask2020/

study human dialogue based on attested, spontaneous and unconstrained, without personal information, oral data in French. It allows the study of long interactions, going beyond informative exchanges.

The games have been recorded in social event at the university. The setting is designed with minimally intrusive device to be quickly forgotten. The participants could thus concentrate on their interactions.

Conversation is one of the strengths of *Catan*, which integrates the principle of negotiation. There are few long silences during a game. It is a game of resources that the different players obtain according to their developments on the board and the result obtained with the dice. Depending on the situation, they can negotiate resources with the other players.

The recordings were then processed to produce a transcribed version of the games. For this a guide was developed and transcribers were recruited. Each recording was treated individually: segmentation into turns, transcription according to the guide, verification by a super annotator. Thus, the resource produced is of very good quality. It contains 14 hours of recording for 22k speaking turns and 115k words.

6.3.8 ArboratorGrew

Bruno Guillaume was implied in the development of ArboratorGrew[5]. ArboratorGrew [12] is a collaborative annotation tool for treebank development. ArboratorGrew combines the features of two preexisting tools: Arborator and Grew. Arborator is a widely used collaborative graphical online dependency treebank annotation tool. Grew is a tool for graph querying and rewriting specialized in structures needed in NLP, i.e. syntactic and semantic dependency trees and graphs. Arborator-Grew is a complete redevelopment and modernization of Arborator, replacing its own internal database storage by a new Grew API, which adds a powerful query tool to Arborator's existing treebank creation and correction features. This includes complex access control for parallel expert and crowd-sourced annotation, tree comparison visualization, and various exercise modes for teaching and training of annotators. Arborator-Grew opens up new paths of collectively creating, updating, maintaining, and curating syntactic treebanks and semantic graph banks.

7 Partnerships and cooperations

7.1 International initiatives

7.1.1 Inria international partners

Declared Inria international partners Sémagramme is part of the Inria-DFKI project IMPRESS. Its goals are are to investigate the integration of semantic knowledge into embeddings and its impact on selected downstream tasks, to extend this approach to multimodal and mildly multilingual settings, and to develop open source software and lexical resources, focusing on video activity recognition as a practical testbed. The project is lead by Pascal Denis (MAGNET, Inria Lille-Europe), and Multispeech (Inria Nancy-Grand Est) member of this project.

Sémagramme is part of the Inria-DFKI project MePheSTO. It is an interdisciplinary research project that envisions a scientifically sound methodology based on artificial intelligence methods for the identification and classification of objective, and thus measurable, digital phenotypes of psychiatric disorders. MePheSTO has a solid foundation of clinically motivated scenarios and use-cases synthesized jointly with clinical partners. Important to MePheSTO is the creation of a multimodal corpus including speech, video, and biosensors of social patient-clinician interactions, which serves as the basis for deriving methods, models and knowledge. Important project outcomes include technical tools and organizational methods for the management of medical data that implement both ELSI and GDPR requirements, demonstration scenarios covering patients' journeys including early detection, diagnosis support, relapse prediction, therapy support. The project is co-lead by François Bremond (Star, Inria Sofia Antipolis).

[5] https://arborator.github.io/

7.2 European initiatives

7.2.1 FP7 & H2020 Projects

AI Proficient Sémagramme is part of the AI Proficient ICT-38-2020 - Artificial intelligence for manufacturing project (see https://cordis.europa.eu/project/id/957391), coordinated by the CRAN laboratory of Université de Lorraine.

By combining human knowledge with AI capabilities, the EU-funded AI-PROFICIENT project will develop proactive control strategies to improve manufacturing processes in terms of production efficiency, quality and maintenance. The overall goal is to increase the positive impact of AI technology on the manufacturing process as a whole, while keeping the human in a central position, assuming supervisory (human-on-the-loop) and executive (human-in-command) roles. By identifying the effective means for human-machine interaction, the project will assist Europe's manufacturing and process industry to improve production planning and execution.

Karën Fort is the Project Ethics Officer and as such is responsible for the ethical dimensions of the project. Marc Anderson was hired as a post-doc researcher on the project is carrying out research on AI Ethics by Design in the Sémagramme team.

7.2.2 Collaborations in European programs, except FP7 and H2020

enetCollect COST action Sémagramme is part of the European Network for Combining Language Learning with Crowdsourcing Techniques (enetCollect) COST action (see https://enetcollect.eurac.edu/), which has been prolonged until September 2021. The action aims at unlocking a crowdsourcing potential available for all languages and at triggering an innovation breakthrough for the production of language learning material, such as lesson or exercise content, and language-related datasets such as, among others, NLP language resources.

Karën Fort is Management Committee member for France and was leading the Working Group 5 of the action (Application-oriented specifications for an ethical, legal and profitable solution) but she resigned in 2020 due to a potential conflict of interest with the AI Proficient external ethical advisor, Katerina Zdravkova (University of Skopje, Macedonia), who was vice leader of WG5.

LITHME COST action Sémagramme is part of the Language In The Human-Machine Era (LITHME) COST action (see http://lithme.eu/). LITHME aims at shining a light on the ethical implications of emerging language technologies. Karën Fort is Management Committee member for France.

7.2.3 Collaborations with major European organizations

7.3 National initiatives

7.3.1 ODiM

Outils informatisés d'aide au Diagnostic des Maladies mentales

2019 - 2022

Coordinator: Maxime Amblard

Participants: Maxime Amblard, Vincent-Thomas Barrouillet, Samuel Buchel, Amandine Lecomte, Chuyuan Li, Michel Musiol

Abstract:

ODiM is an interdisciplinary project, at the interface of psychiatry-psychopathology, linguistics, formal semantics and digital sciences. It aims to replace the paradigm of Language and Thought Disorders (LTD) as used in the Mental Health sector with a semantic-formal and cognitive model of Discourse Disorders (DD). These disorders are translated into pathognomonic signs, making them complementary diagnostic tools as well as screening for vulnerable people before the psychosis's trigger. The project has three main components.

The work is based on real data from interviews with patients with schizophrenia. A data collection phase in partner hospitals and with a control group, consisting of interviews and neuro-cognitive tests, is therefore necessary.

The data collection will allow the development of the theoretical model, both in psycholinguistic and semantic formalization for the identification of diagnostic signs. The success of such a project requires the extension of the analysis methodology in order to increase the model's ability to identify sequences with symptomatic discontinuities.

If the general objective of the project is to propose a methodological framework for defining and understanding diagnostic clues associated with psychosis, we also wish to equip these approaches by developing software to automatically identify these clues, both in terms of discourse and language behaviour.

7.3.2 ANR CoDeinE

The ANR project CoDeinE (artificial text COrpus DEsIgNed Ethically automatic synthesis of clinical documents) is coordinated by Aurélie Névéol (Limsi). Sémagramme is one the partner: Karën Fort (local coordinator) and Bruno Guillaume are involved in the project.

7.3.3 GDR LIFT

Sémagramme participates in GDR LIFT (Linguistique Informatique, Formelle et de Terrain). Karën Fort is co-chair (with G. Wisniewski) of the axis 2: Linguistique et évaluation des systèmes de traitement automatique des langues.

8 Dissemination

8.1 Promoting scientific activities

8.1.1 Scientific events: organization

General chair, scientific chair

- Maxime Amblard: Co-chair of ETeRNAL2: atelier Ethique et TRaitemeNt Automatique des Langues de la conférence JEP-TALN-RECITAL 2020 [18].
- Karën Fort: Co-chair of ETeRNAL2: atelier Ethique et TRaitemeNt Automatique des Langues de la conférence JEP-TALN-RECITAL 2020 [18].
- Sylvain Pogodalla: Co-chair of JEP-TALN-RECITAL 2020: the 6th joined conference JEP (Journées d'Études sur la Parole, 33rd edition), TALN (Conférence sur le Traitement Automatique des Langues Naturelles, 27th edition) and RÉCITAL (Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues, 22nd edition) [21, 20, 22, 23].

EThics Committee Chair

- Karën Fort: Co-chair (with Dirk Hovy) of the ethics committee of EMNLP2020 (see `https://2020.emnlp.org/organizers/ethics-committee`.)

8.1.2 Scientific events: selection

Chair of conference program committees

- Sylvain Pogodalla: Co-chair of the 6th joined conference JEP (Journées d'Études sur la Parole, 33rd edition), TALN (Conférence sur le Traitement Automatique des Langues Naturelles, 27th edition) and RÉCITAL (Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues, 22nd edition) [21, 20, 22, 23].
- Maxime Amblard and Karën Fort: Co-chair of ETeRNAL2: atelier Ethique et TRaitemeNt Automatique des Langues de la conférence JEP-TALN-RECITAL 2020 [18].

Member of the conference program committees

- Maxime Amblard: 6th joined conference JEP (Journées d'Études sur la Parole, 33rd edition), TALN (Conférence sur le Traitement Automatique des Langues Naturelles, 27th edition) and RÉCITAL (Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues, 22nd edition)

- Philippe de Groote: SEM 2020 (ninth joint conference on lexical and computational semantics), SCiL 2021 (fourth meeting of the Society for Computation in Linguistics), PaM 2020 (Conference on probability and Meaning).

Reviewer

- Maxime Amblard: 28th International Conference on Computational Linguistics (COLING'2020) The 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP2020) the 29th International Joint Conference on Artificial Intelligence and the 17th Pacific Rim International Conference on Artificial Intelligence (IJCAI-PRICAI 2020), 6th joined conference JEP (Journées d'Études sur la Parole, 33rd edition), TALN (Conférence sur le Traitement Automatique des Langues Naturelles, 27th edition) and RÉCITAL (Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues, 22nd edition), NLPinAI.

- Karën Fort: 28th International Conference on Computational Linguistics (COLING'2020), the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing (AACL-IJCNLP'2020), *SEM 2020 workshop, Citizen Linguistics in Language Resource Development workshop 2020, REPROLANG 2020: Shared Task on the Reproduction of Research Results in Science and Technology of Language, LREC 2020, TALN 2020, RÉCITAL 2020, Atelier EGC Humains et IA, 2020, Colloque La fabrique de la participation culturelle. Plateformes numériques et enjeux démocratiques, 2020.

- Bruno Guillaume: 28th International Conference on Computational Linguistics (COLING'2020)

- Pierre Ludmann: 28th International Conference on Computational Linguistics (COLING'2020), RÉCITAL 2020 (Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues, 22nd edition).

- Guy Perrier: Universal Dependencies Workshop 2020 (UDW 2020).

- Sylvain Pogodalla: 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020).

8.1.3 Journal

Member of the editorial boards

- Maxime Amblard: Member of the editorial board of the journal Traitement Automatique des Langues, in charge of the *pdf pipeline*.

- Philippe de Groote: area editor of the *FoLLI-LNCS series*.

- Sylvain Pogodalla: Member of the editorial board of the journal Traitement Automatique des Langues, in charge of the *Résumés de thèses* section.

- Michel Musiol: Psychological and educational sciences (Université d'ElOued Ed)

Reviewer - reviewing activities

- Maxime Amblard:
  - Journal of Logic, Language and Information
  - Mathematics
  - PlosOne
- Philippe de Groote: Studia Logica
- Karën Fort: Frontiers 2020
- Michel Musiol:
  - Journal of french linguistic studies
  - Education et Société Inclusives

8.1.4 Invited talks

- Bruno Guillaume was invited to give a Lattice seminar in Paris on January 14th.
- Guy Perrier was invited to give a talk at the seminar of "Master Industries de la Langue de Grenoble" on November 6th.
- Michel Musiol: Approches interactionnelles et tentatives de modélisation du trouble schizophrénique. Centre Hospitalier Universitaire de Nice, Département de Psychiatrie, on October 21th.

8.1.5 Leadership within the scientific community

- Maxime Amblard: Management Committee of the OLKI project (Lorraine Université d'Excellence project - PIA), co-leader of the workpackage 2 on NLP activities.

8.1.6 Scientific expertise

- Maxime Amblard is member of the scientific board of the INJS - Institut National des Jeunes Sourds.
- Marc Anderson served as project review panel member for AI Ethics related projects in the New Frontiers in Research Fund 2020 Exploration Competition (Government of Canada).
- Michel Musiol: expertise CIFRE ANRT

8.1.7 Research administration

- Maxime Amblard
  - Member of conseil scientifique of Université de Lorraine
  - Standing invitee at the pôle scientifique AM2I of Université de Lorraine
  - Member of the Sénat Académique of Université de Lorraine
  - Member of the progress commission of Université de Lorraine
  - Member of the administration council of the Institut des sciences du digital, management et cognition
  - Head of the master in Natural Language Processing (master 1 and 2)
- Philippe de Groote:
  - Member of the *bureau du comité des projets d'Inria Nancy – Grand Est.*
  - Member of the scientific council of the LIRMM, *Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier*

- Bruno Guillaume:
  - Head of the Natural Language Processing and Knowledge Discovery department of the LORIA laboratory
  - Manager (with Alain Polguère) of the CPER (Contrat de Plan État-Région) "Langues, Connaissances et Humanités Numériques".
- Sylvain Pogodalla:
  - Elected member of the comité de centre Inria Nancy – Grand Est,
  - In charge of the local *commission IES (information et édition scientifique)* of the Inria Nancy – Grand Est and LORIA.
  - Member of the national *commission IES* of Inria.
- Michel Musiol:
  - member of the Professor selection committee, neuropsychology, section 16 (Université de Lorraine)
  - member of the Professor selection committee, clinical psychology, section 16 (Université de Lorraine)
  - member of the MCF selection committee, psychology and psychiatry, section 16 (Université de Paris)
  - Assistant Director of UMR 7118 Atilf CNRS, until September
  - member of the CLCS scientifical pole

8.2 Teaching - Supervision - Juries

8.2.1 Teaching

- Licence:
  - Maxime Amblard, AI Introduction, 15h, L1, Université de Lorraine, France.
  - Maxime Amblard, Maria Boritchev and Chuyuan Li NLP for beginners, 10h, L2, Université de Lorraine, France.
  - Maxime Amblard, Maria Boritchev and Chuyuan Li Linguistic engineering, 10h, L3, Université de Lorraine, France.
  - Maria Boritchev, Formalisms and reasoning representations , 20h, L3, Université de Lorraine, France.
  - Maria Boritchev, Algorithmic 1, 22h, L1, Université de Lorraine, France.
  - Pierre Ludmann, Informatics 2, 20h, Mines de Nancy, France.
- Master:
  - Maxime Amblard, Chuyuan Li and Siyana Pavlova, Python Programming, 30h, M1 NLP, Université de Lorraine, France.
  - Maxime Amblard, Methods for NLP, 36h, M1 NLP, Université de Lorraine, France.
  - Maxime Amblard, Formalisms and Syntax, 24h, M2 NLP, Université de Lorraine, France.
  - Maxime Amblard, Discourse and Dialogue, 18h, M2 NLP, Université de Lorraine, France.
  - Philippe de Groote, Formal Logic, 22h, M1 NLP, Université de Lorraine, France.
  - Philippe de Groote, Formal languages, 22h, M1 NLP, Université de Lorraine, France.
  - Philippe de Groote, Computational Semantics, 18h, M2 NLP, Université de Lorraine, France.
  - Philippe de Groote, Computational structures and logics for natural language modeling, 18h, M2 NLP, Université Paris Diderot – Paris 7, France.
  - Karën Fort, Data ethics (English), 3h, M2 NLP and cog. Sces (IDMC), Université de Lorraine, France.
  - Bruno Guillaume, Written Corpora TAL (english), 44h, M1 NLP, Université de Lorraine, France.

8.2.2 Tutorials

- Karën Fort co-organized the tutorial on Reviewing Natural Language Processing Research (Introductory) at ACL 2020, with Kevin Cohen, Margot Mieskes and Aurélie Névéol.[24]

8.2.3 Supervision

- PhD in progress:
  - William Babonnaud, *Lexical semantics, compositionality and type coercion*, since September 2018, Philippe de Groote.
  - Maria Boritchev, *Dialogue Dynamics Modeling in the Simple Theory of Types*, since September 2017, Maxime Amblard and Philippe de Groote.
  - Pierre Ludmann, *Dynamic construction of discursive structures*, since September 2017, Philippe de Groote and Sylvain Pogodalla.
  - Chuyuan Li, *Formal and statistical modeling of dialogue*, since October 2019, Maxime Amblard and Chloé Braud.
  - Samuel Buchel, *Linguistic, semantic and cognitive modelling of dialogical incongruities and discontinuities in the interaction with the schizophrenic patients*, since December 2019, Maxime Amblard and Michel Musiol.
  - Siyana Pavlova, *Tools and methods for semantic annotation*, since November 2020, Maxime Amblard and Bruno Guillaume.
  - Priyansh Trivedi, *injecting lexical and semantic knowledge into word, phrasal and sentence embeddings*

8.3 Popularization

8.3.1 Internal or external Inria responsibilities

- Maxime Amblard is the vice head of editorial board of Interstices.info

8.3.2 Articles and contents

- Maxime Amblard publish an interstices.info article [25]
- Maxime Amblard has written the scenaris of two cartoon movies about AI and NLP for the OLKi project.
- Talk about Artificial Intelligence for a general audience for Institut des Sciences du Digital, Management et Cognition.
- Presentation of the ODiM project on national broadcast, La méthode Scientifique - France Culture

8.3.3 Education

- Marc Anderson gave weekly open/public lectures on ethics and value for business/industry in the context of the logic of Hyperthematics (`https://www.ideatrek.io`).
- Maxime Amblard Unplugged Computer Science on grammars, rabbits and carrots - afternoon with undergraduate students.

8.3.4 Interventions

- Maxime Amblard: Long talk (2 x 2 hours) about Artificial Intelligence and NLP for Moovie studies students at IECA - Université de Lorraine

9 Scientific production

9.1 Major publications

[1] G. Bonfante and B. Guillaume. 'Non-size increasing Graph Rewriting for Natural Language Processing'. In: *Mathematical Structures in Computer Science* 28.08 (2018), pp. 1451–1484. DOI: 10.1017/S0960129518000178. URL: https://hal.inria.fr/hal-00921038.

[2] G. Bonfante, B. Guillaume and G. Perrier. *Application of Graph Rewriting to Natural Language Processing*. Vol. 1. Logic, Linguistics and Computer Science Set. ISTE Wiley, 2018, p. 272. URL: https://hal.inria.fr/hal-01814386.

[3] P. de Groote and M. Kanazawa. 'A Note on Intensionalization'. In: *Journal of Logic, Language and Information* 22.2 (2013), pp. 173–194. DOI: 10.1007/s10849-013-9173-9. URL: https://hal.inria.fr/hal-00909207.

[4] S. Pogodalla. 'A syntax-semantics interface for Tree-Adjoining Grammars through Abstract Categorial Grammars'. In: *Journal of Language Modelling* 5.3 (2017), pp. 527–605. DOI: 10.15398/jlm.v5i3.193. URL: https://hal.inria.fr/hal-01242154.

9.2 Publications of the year

International journals

[5] Š. Arhar Holdt, R. Zviel-Girshin, E. Gajek, I. Durán-Muñoz, K. Fort, P. Bago, C. Hatipoglu, R. Kasperavičienė, S. Koeva, I. Lazić Konjik, L. Miloshevska, A. Ordulj, C. Rodosthenous, E. Volodina, T. Weber and L. Zanasi. 'Language Teachers and Crowdsourcing: Insights from a Cross-European Survey'. In: *Rasprave Instituta za hrvatski jezik i jezikoslovlje* 46.1 (2nd Sept. 2020), pp. 1–28. DOI: 10.31724/rihjj.46.1.1. URL: https://hal.inria.fr/hal-02974069.

[6] M. Candito, M. Constant, C. Ramisch, A. Savary, B. Guillaume, Y. Parmentier and S. R. Cordeiro. 'A French corpus annotated for multiword expressions and named entities'. In: *Journal of Language Modelling* (2021). URL: https://hal.archives-ouvertes.fr/hal-03016721.

International peer-reviewed conferences

[7] M. Amblard, C. Beysson, P. de Groote, B. Guillaume and S. Pogodalla. 'A French Version of the FraCaS Test Suite'. In: LREC 2020 - Language Resources and Evaluation Conference. Marseille, France, 11th May 2020, p. 9. URL: https://hal.inria.fr/hal-02619239.

[8] W. Babonnaud and P. de Groote. 'Lexical selection, coercion, and record types'. In: LENLS17 : Logic & Engineering of Natural Language Semantics. Online, Japan, 15th Nov. 2020. URL: https://hal.inria.fr/hal-03076311.

[9] M. Boritchev and M. Amblard. 'There is as yet Insufficient Data for a Meaningful Answer'. In: SemDial - WatchDial The 24th Workshop on the Semantics and Pragmatics of Dialogue. Brandeis, United States: https://www.brandeis.edu/nasslli2020/watchdial-main/, 19th July 2020. URL: https://hal.inria.fr/hal-02930715.

[10] M. Boritchev and P. de Groote. 'On dialogue modeling: a dynamic epistemic inquisitive approach'. In: LENLS17 : Logic & Engineering of Natural Language Semantics. Online, Japan, 15th Nov. 2020. URL: https://hal.inria.fr/hal-03065236.

[11] K. Fort, B. Guillaume, Y.-A. Pilatte, M. Constant and N. Lefèbvre. 'Rigor Mortis: Annotating MWEs with a Gamified Platform'. In: LREC 2020 - Language Resources and Evaluation Conference. Marseille, France, 11th May 2020. URL: https://hal.inria.fr/hal-02571466.

[12] 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.

[13] A. Millour and K. Fort. 'Text Corpora and the Challenge of Newly Written Languages'. In: 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020). Proceedings of the 1st Joint SLTU and CCURL Workshop. Marseille, France, 11th May 2020. URL: https://hal.archives-ouvertes.fr/hal-02611209.

[14] L. Nicolas, V. Lyding, C. Borg, C. Forascu, K. Fort, K. Zdravkova, I. Kosem, J. Cibej, Š. A. Holdt, A. Millour, A. König, C. Rodosthenous, F. Sangati, U. u. Hassan, A. Katinskaia, A. Barreiro, L. Aparaschivei and Y. Hacohen-Kerner. 'Creating Expert Knowledge by Relying on Language Learners: a Generic Approach for Mass-Producing Language Resources by Combining Implicit Crowdsourcing and Language Learning'. In: LREC 2020 - Language Resources and Evaluation Conference. Marseille, France, 11th May 2020. URL: https://hal.inria.fr/hal-02879883.

[15] C. Ramisch, A. Savary, B. Guillaume, J. Waszczuk, M. Candito, A. Vaidya, V. B. Mititelu, A. Bhatia, U. Iñurrieta, V. Giouli, T. Güngör, M. J. Polyu, T. Lichte, C. Liebeskind, J. Monti, R. Ramisch, S. Stymne, A. Walsh and H. Xu. 'Edition 1.2 of the PARSEME Shared Task on Semi-supervised Identification of Verbal Multiword Expressions'. In: Joint Workshop on Multiword Expressions and Electronic Lexicons (MWE-LEX 2020). Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons. Barcelona, Spain: https://www.aclweb.org/anthology/volumes/2020.mwe-1/, Dec. 2020. URL: https://hal.archives-ouvertes.fr/hal-03014927.

National peer-reviewed Conferences

[16] M. Amblard, C. Li, C. Braud, C. Demily, N. Franck and M. Musiol. 'Investigating Learning Methods Applied to Language Specificity of Persons with Schizophrenia'. 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*. 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. Nancy, France: https://jep-taln2020.loria.fr/, 2020, pp. 12–26. URL: https://hal.archives-ouvertes.fr/hal-02784752.

[17] A. Millour, K. Fort and P. Magistry. 'Replicating and extending for Alsatian : "POS tagging for low-resource languages by adaptingword embeddings"'. 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). 2e atelier Éthique et TRaitemeNt Automatique des Langues (ETeRNAL)*. 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). 2e atelier Éthique et TRaitemeNt Automatique des Langues (ETeRNAL). Nancy, France, 2020, pp. 29–37. URL: https://hal.archives-ouvertes.fr/hal-02750224.

Edition (books, proceedings, special issue of a journal)

[18] G. Adda, M. Amblard and K. Fort, eds. *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). 2e atelier Éthique et TRaitemeNt Automatique des Langues (ETeRNAL)*. Nancy, France: http://talnarchives.atala.org/ateliers/2020/ETeRNAL/, 2020. URL: https://hal.archives-ouvertes.fr/hal-02750218.

[19] M. Amblard, M. Musiol and M. Rebuschi. *(In)coherence of Discourse. Formal and conceptual issues of language*. Language, Cognition and Mind (Chungmin Lee "Editor Springer book series). 2020. URL: https://hal.archives-ouvertes.fr/hal-02501028.

[20] C. Benzitoun, C. Braud, L. Huber, D. Langlois, S. Ouni, S. Pogodalla and S. Schneider, eds. *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. Vol. 2. Volume 2 : Traitement Automatique des Langues Naturelles. Nancy, France: http://talnarchives.atala.org/TALN/TALN-2020/, 2020. URL: https://hal.archives-ouvertes.fr/hal-02784750.

[21] C. Benzitoun, C. Braud, L. Huber, D. Langlois, S. Ouni, S. Pogodalla and S. Schneider, eds. *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 1 : Journées d'Études sur la Parole*. Nancy, France, 2020. URL: https://hal.archives-ouvertes.fr/hal-02798507.

[22] C. Benzitoun, C. Braud, L. Huber, D. Langlois, S. Ouni, S. Pogodalla and S. Schneider, eds. *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 : 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). Vol. 3. Volume 3 : Rencontre des Étudiants Chercheurs en Informatique pour le TAL. Nancy, France: http://talnarchives.atala.org/RECITAL/RECITAL-2020/, 2020. URL: https://hal.archives-ouvertes.fr/hal-02786181.

[23] C. Benzitoun, C. Braud, L. Huber, D. Langlois, S. Ouni, S. Pogodalla and S. Schneider, eds. *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 4 : Démonstrations et résumés d'articles internationaux*. JEP-TALN-RECITAL 2020. Vol. 4. Volume 4 : Démonstrations et résumés d'articles internationaux. Nancy, France: http://talnarchives.atala.org/TALN/TALN-2020/, 2020. URL: https://hal.archives-ouvertes.fr/hal-02768750.

Other scientific publications

[24] K. Bretonnel Cohen, K. Fort, M. Mieskes and A. Névéol. *Reviewing Natural Language Processing Research*. Seattle, United States, 5th July 2020. DOI: 10.18653/v1/2020.acl-tutorials.4. URL: https://hal.inria.fr/hal-02943568.

9.3 Other

Scientific popularization

[25] M. Amblard. 'Le problème des 8 reines'. In: *Interstices* (20th Nov. 2020). URL: https://hal.inria.fr/hal-03131246.

9.4 Cited publications

[26] T. L. Booth and R. A. Thompson. 'Applying Probability Measures to Abstract Languages'. In: *IEEE Transactions on Computers* 22.5 (May 1973), pp. 442–450. DOI: 10.1109/T-C.1973.223746. URL: https://doi.org/10.1109/T-C.1973.223746.

[27] S. Chatzikyriakidis, R. Cooper, S. Dobnik and S. Larsson. 'An overview of Natural Language Inference Data Collection: The way forward?' In: *Proceedings of the Computing Natural Language Inference Workshop*. 2017. URL: https://www.aclweb.org/anthology/W17-7203.

[28] D. Chiang. 'Statistical Parsing with an Automatically-Extracted Tree Adjoining Grammar'. In: *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*. Hong Kong: Association for Computational Linguistics, Oct. 2000, pp. 456–463. DOI: 10.3115/1075218.1075276. URL: https://www.aclweb.org/anthology/P00-1058.

[29] I. Ciardelli, J. Groenendijk and F. Roelofsen. *Inquisitive Semantics*. Oxford Surveys in Semantics and Pragmatics. Oxford University Press, 2018.

[30] R. Cooper, S. Chatzikyriakidis and S. Dobnik. *Testing the FraCaS test suite*. Presentation at the Unshared Task "Theory and System analysis with FraCaS, MultiFraCaS and JSeM Test Suites" of Logical Engineering of Natural Language Semantics 13 (LENLS 13). 2016.

[31] R. Cooper, D. Crouch, J. Van Eijck, C. Fox, J. Van Genabith, J. Jaspars, H. Kamp, D. Milward, M. Pinkal, M. Poesio, S. Pulman, T. Briscoe, H. Maier and K. Konrad. *Using the framework*. Tech. rep. Technical Report LRE 62-051 D-16. The FraCaS Consortium, 1996.

[32] P. de Groote. 'Towards a Montagovian Account of Dynamics'. In: *16th Semantics and Linguistic Theory conference - SALT2006*. Ed. by M. Gibson and J. Howell. Tokyo, Japan, 2006. URL: https://journals.linguisticsociety.org/proceedings/index.php/SALT/article/view/2952/0.

[33] P. de Groote. 'Towards abstract categorial grammars'. In: *Association for Computational Linguistics, 39th Annual Meeting and 10th Conference of the European Chapter*. Colloque avec actes et comité de lecture. internationale. Toulouse, France: Association for Computational Linguistics, July 2001, pp. 148–155. URL: http://hal.inria.fr/inria-00100529/en.

[34] P. de Groote and S. Pogodalla. 'On the expressive power of abstract categorial grammars: Representing context-free formalisms'. In: *Journal of Logic, Language and Information* 13.4 (2004), pp. 421–438.

[35] K. Fort, B. Guillaume, M. Constant, N. Lefèbvre and Y.-A. Pilatte. '"Fingers in the Nose": Evaluating Speakers' Identification of Multi-Word Expressions Using a Slightly Gamified Crowdsourcing Platform'. In: *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*. Santa Fe, United States, Aug. 2018, pp. 207–213. URL: https://hal.archives-ouvertes.fr/hal-01912706.

[36] Z. Fülöp and H. Vogler. 'Weighted Tree Automata and Tree Transducers'. In: *Handbook of Weighted Automata*. Ed. by M. Droste, W. Kuich and H. Vogler. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. Chap. 9, pp. 313–403. DOI: 10.1007/978-3-642-01492-5_9. URL: https://doi.org/10.1007/978-3-642-01492-5_9.

[37] K. Gerdes, B. Guillaume, S. Kahane and G. Perrier. 'Improving Surface-syntactic Universal Dependencies (SUD): surface-syntactic relations and deep syntactic features'. In: *TLT 2019 - 18th International Workshop on Treebanks and Linguistic Theories*. Paris, France, Aug. 2019. URL: https://hal.inria.fr/hal-02266003.

[38] J. Ginzburg. *The interactive stance*. Oxford University Press, 2012.

[39] B. Guillaume and G. Perrier. 'Interaction Grammars'. In: *Research on Language & Computation* 7 (2009), pp. 171–208.

[40] E. Lebedeva. 'Expressing discourse dynamics through continuations'. Thèse de Doctorat. Université de Lorraine, 2012.

[41] G. Perrier. 'A French Interaction Grammar'. In: *International Conference on Recent Advances in Natural Language Processing - RANLP 2007*. Ed. by G. Angelova, K. Bontcheva, R. Mitkov, N. Nicolov and K. Simov. IPP & BAS & ACL-Bulgaria. Borovets, Bulgarie: INCOMA Ltd, Shoumen, Bulgaria, 2007, pp. 463–467. URL: http://hal.inria.fr/inria-00184108/en/.

[42] P. Resnik. 'Probabilistic Tree-Adjoining Grammar as a Framework for Statistical Natural Language Processing'. In: *Proceedings of the 14th Conference on Computational Linguistics (COLING 1992). Volume 2*. Nantes, France: Association for Computational Linguistics, 1992, pp. 418–424. DOI: 10.3115/992133.992135. URL: https://www.aclweb.org/anthology/C92-2065.

[43] A. Salomaa. 'Probabilistic and Weighted Grammars'. In: *Information and Control* 15.6 (1969), pp. 529–544. DOI: 10.1016/S0019-9958(69)90554-3. URL: http://www.sciencedirect.com/science/article/pii/S0019995869905543.

[44] Y. Schabes. 'Stochastic Lexicalized Tree-Adjoining Grammars'. In: *Proceedings of the 14th Conference on Computational Linguistics (COLING 1992). Volume 2*. Nantes, France: Association for Computational Linguistics, 1992, pp. 425–432. DOI: 10.3115/992133.992136. URL: https://www.aclweb.org/anthology/C92-2066/.