RESEARCH CENTRE
Inria Nancy - Grand Est Center

IN PARTNERSHIP WITH:
CNRS, Université de Lorraine

2022
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

Inria

Contents

Project-Team SEMAGRAMME

Creation of the 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

B2. – Health

B9.6.8. – Linguistics

B9.9. – Ethics

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 [UL, Associate Professor, HDR]

- Karën Fort [Sorbonne Université, Associate Professor, HDR]

- Jacques Jayez [ENS de Lyon, Emeritus]

- Michel Musiol [UL, Professor, until Aug 2022, HDR]

- Guy Perrier [UL, Emeritus, HDR]

Post-Doctoral Fellow

- Marc Anderson [UL]

PhD Students

- William Babonnaud [Heinrich Heine Universität Düsseldorf, until Nov 2022]

- Samuel Buchel [INRIA]

- Hee-Soo Choi [UL]

- Marie Cousin [UL, from Oct 2022]

- Amandine Decker [UL, from Oct 2022]

- Maxime Guillaume [Yseop, CIFRE]

- Laurine Jeannot [CS GROUP, CIFRE, until Oct 2022]

- Amandine Lecomte [INRIA]

- Chuyuan Li [UL]

- Pierre Ludmann [UL]

- Siyana Pavlova [UL]

- Valentin Richard [UL]

- Priyansh Trivedi [INRIA]

Technical Staff

- Pierre Lefebvre [UL, Engineer, from May 2022 until Oct 2022]

- Vincent Tourneur [INRIA, Engineer, from Aug 2022]

- Vincent Tourneur [CNRS, Engineer, until Jul 2022]

Interns and Apprentices

- Marie Cousin [UL, from Mar 2022 until Aug 2022]

- Amandine Decker [UL, Intern, from Mar 2022 until Sep 2022]

- Fanny Ducel [Sorbonne Université, from Jul 2022 until Aug 2022]

- Violette Pelgrims [UL, from Nov 2022]

- Wenjun Sun [UL, from Mar 2022 until Aug 2022]

Administrative Assistants

- Isabelle Herlich [INRIA]

- Delphine Hubert [UL]

External Collaborators

- Mathieu Constant [UL, HDR]

- Michel Musiol [UL, from Sep 2022, HDR]

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.

Moreover, a transversal and transdisciplinary theme has been developed in the team in the past years: ethics in NLP and more generally in AI.

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 [51].

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 [52], [6].

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 [49]. 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 [53] and graph rewriting [2, 3] as models of natural language syntax. This includes the development of grammars for French [56], together with morphosyntactic 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 (also known as 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,

- (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 Highlights of the year

5.1 Science promotion

The members of the Sémagramme team are strongly devoted to the academic values of scientific freedom and peer review. They have continued and will continue to promote these values in spite of the current management style of the institute that seems to be more and more autocratic and even possibly toxic.

5.2 Awards

Marc Anderson's paper "Some Ethical Reflections on the EU AI Act" won the best short paper award at the 1st International Workshop on Imagining the AI Landscape after the AI Act (IAIL 2022), Vrije Universiteit Amsterdam (13th June 2022) [14] .

6 New software and platforms

6.1 New software

6.1.1 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 a piece of software for developing and using Abstract Categorial Grammars (ACG).

Release Contributions: This version fixes some bugs and adds some commands to the scripting language. Some internal modifications also prepare ACGtk to add extensions of ACGs such as weighting.

News of the Year: In addition to modifications to maintain the code (bug fixes, improvement of error messages, etc.), new functionalities have been added. The grammar of data files has been extended (in particular with UTF-8 symbols) and the command language to use ACGs has been redesigned. Documentation has been improved and a continuous integration workflow has been set up. Several optimizations (in particular regarding magic set rewriting) have been implemented.

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

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

Contact: Sylvain Pogodalla

Participants: Philippe De Groote, Pierre Ludmann, Jiri Marsik, Sylvain Pogodalla, Vincent Tourneur

6.1.2   Grew

Name: Graph Rewriting

Keywords: Semantics, Syntactic analysis, NLP, 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 2022, a few Grew software versions were released (1.10 is the latest one).

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 universal dependencies corpora (243 in 138 different languages in v2.11) are available and data from several other projects can also be queried. In 2022, 145,000 requests were received on the Grew-match server.

A new Python binding for the grew library was started and is in active development (`https://github.com/grew-nlp/grewpy`)

An article dedicated to the usage of Grew on semantic representation was published (`https://hal.inria.fr/hal-03724068`). Grew was a central tool in other publications (`https://hal.inria.fr/hal-03724129`, `https://hal.inria.fr/hal-03846825`)

Grew is also used in the Arborator-Grew software (`https://arborator.github.io/`), the annotation platform of the Autogramm ANR project.

URL: `https://grew.fr/`

Publications: hal-03724068, hal-03177701, hal-01930591, hal-01814386, hal-03021720, hal-03724068, hal-03724129, hal-03846825

Contact: Bruno Guillaume

Participants: Bruno Guillaume, Guy Perrier, Guillaume Bonfante

6.1.3   SLODiM

Name: SLODiM

Keywords: NLP, 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: The latest version integrates new treatments in particular at the level of the identification of the backchannels.

A version without the graphical representations is available without an account. Its purpose is to make visible the treatments that are produced in the system.

News of the Year: In 2022, Slodim was redesigned and splat in two parts in order to allow for thorough analyses, on the one hand, and a lean access that presents the computed data, without analyses, on the other hand. A transactional automatic analysis of dialogues was also implemented.

URL: `https://academics.slodim.loria.fr/`

Contact: Maxime Amblard

Partners: Loria, Université de Lorraine, CNRS

7 New results

7.1 Syntax-Semantics Interface

Participants: Maxime Amblard, William Babonnaud, Marie Cousin, Philippe de Groote, Bruno Guillaume, Maxime Guillaume, Pierre Ludmann, Sylvain Pogodalla, Siyana Pavlova, Valentin Richard, Priyansh Trivedi.

7.1.1 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 a generic feature structure framework. 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. 6).

We also investigated a naive way of simulating feature logic with atomic types on simplified frameworks [33].

Multityped ACG (mACG) Symbolic parsing with large coverage grammars usually leads to combinatorial explosion of syntactic ambiguities (a single expression has many syntactic analyses). A widespread method to tackle this issue is to use statistics and probabilities, leading for instance to probabilistic Context Free Grammars (pCFGs) and probabilistic Tree Adjoining Grammars (pTAGs). An important goal is then to also extend ACGs with probabilities or weights.

Yet, ACGs come with features that make this extension non-trivial. In particular, ACGs can be composed by making the parse structures of a grammar the surface structures of another ACG. The resulting composition is a full-flavored ACG. Because adding weights to ACGs naturally leads to refining admissible abstract structures (associated with a weight), ACG composition does not anymore correspond to functional composition. We introduced multityped ACGs (mACGs) [29]. Multityped ACGs are the underlying discrete mathematical structures that will support weighting extension. We showed that a suitable notion of composition can be defined for multityped ACGs as well.

Encoding of Meaning-Text Theory into ACGs Meaning-Text Theory (MTT) is a linguistic theory geared towards generating natural language expressions from semantic representations [54]. It relies on seven representation levels (e.g., semantics, deep syntax, surface syntax, etc.). Structures at each level are related to structures at the adjacent levels by rewriting devices. ACGs come with several composition modes, one of which in particular corresponds to transduction of (tree or graph) structures. We used this ability to study the extent to which MTT architecture can be encoded into ACGs. We showed that, while some of the MTT mechanisms can be faithfully accounted for within ACGs, some other ones require additional mechanisms [47].

7.1.2 Lexical Semantics and Linguistic Knowledge

Lexical Coercion and Types 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 has studied possible solutions to the acknowledged incompatibility between subtyping and Montagovian-style λ-calculus. He has shown that choosing a topos as categorical semantics for such λ-calculi enables one to obtain the same properties as Modern Type Theories. Moreover, he has developed a predicate calculus specifically designed for Montague semantics, which features covariant subtyping in a type-safe way and relies on the power of toposes to ensure all the necessary properties.

His doctoral dissertation [44] further develops this calculus by proposing a stronger and more general design and by proving several of its non-trivial properties. It uses a generalized notion of coercion as a mechanism of type change, which covers a wide variety of cases including subtyping, polysemy resolution, creative uses of language, and structural type shifts as introduced by Barbara Partee [55] ; it also comes with rules for coercion inference which enable the implementation of the calculus as a compositional system at the syntax-semantics interface. The dissertation also includes an extensive philosophical discussion on the necessary properties that types should have for proper natural language modeling and how the calculus implements them, as well as arguments in favor of an empirical investigation on the concrete set of types which would be needed for that purpose. A first experiment on how to extract types from corpus data has also been conducted in this setting, leading to promising results and more ideas and guidelines to improve them.

7.1.3 Entity-Mention Information Extraction

Priyansh Trivedi investigated whether multi-task learning for entity-mention level NLP tasks can be generalized across multiple domains. Entity-mentions are referring expressions used to refer to entities in text, such as pronouns, nouns, and noun phrases. The tasks of detecting these mentions (named entity recognition), detecting mentions of the same entity (coreference resolution), and detecting semantic relations between these entities (relation extraction) are some of the well-researched problems in the field of natural language processing, or more specifically, information extraction. Interestingly, recent research works have demonstrated that solving these tasks together in a multi-task setting is mutually beneficial [58, 57]. However these advances assume the availability of aligned annotations, i.e., the same dataset to be annotated with labels from multiple tasks. In this work, we tried to find whether existing models can still be trained in a mutually beneficial manner when different datasets (domain adaptation) are annotated with different tasks (multi-task learning).

The preliminary findings indicate that existing models suffer from domain shift when trained in a multi-domain multi-task setting and generalize poorly when compared to models trained only in a multi-task setting. This negative result suggests that a future investigation on better domain adaptation techniques might be helpful, enabling these models to capitalize on a larger availability of samples from varied domains. Further analysis also provides further evidence to a previously known drawback of multi-task approaches to solving entity-mention level tasks, i.e., typically the annotation guidelines for selecting span boundaries differ across tasks. This problem is further compounded upon in a multi-domain setting. Ways to mitigate this challenge are currently explored, taking inspiration from word-level coreference models [50].

7.1.4 Graph-Based Semantics

We have been studying and comparing different existing semantic graph-based annotation frameworks (AMR, UCCA and DRS). The goal is to determine whether these frameworks are compatible and if they encode the same level of semantic information.

In [32], two of the currently popular semantic frameworks are considered: Abstract Meaning Representation (AMR)—a more abstract framework—and Universal Conceptual Cognitive Annotation (UCCA)—an anchored framework. A corpus-based approach was used to build two graph rewriting systems, a deterministic and a non-deterministic one, from the former to the latter framework. The graph rewriting systems were evaluated and ambiguities discovered when build the rules are reported.

In [13], we show how the online tool Grew-match can be used to make queries and visualize data from existing semantically annotated corpora. A dedicated syntax is available to construct simple to complex queries and execute them against a corpus. Such queries give transverse views of the annotated data, these views can help for checking the consistency of annotations in one corpus or across several corpora.

Grew-match can then be seen as an error mining tool: when inconsistencies are detected, it helps find the sentences which should be fixed. Finally, Grew-match can also be used as a side tool to assist annotation tasks helping to find annotation examples in existing corpora to be compared to the data to be annotated.

7.1.5 Semantics of quantification

In a joint work with Harry Bunt, we keep on working on a project that aims at establishing an interoperable annotation scheme for quantification phenomena as part of the ISO suite of standards for semantic annotation, known as the Semantic Annotation Framework. After a break, caused by the Covid-19 pandemic, the project was relaunched in early 2022 with a second working draft, which deals with certain issues in the annotation of quantification in a more satisfactory way than the original working draft [18].

7.1.6 Semantics of questions

Natural language statements are composed not only of declarative sentences but also of interrogative ones. Moreover, sentences cannot be categorized into purely declarative or purely interrogative sentences. Typically, a declarative statement may contain an indirect interrogative clause:

 (a) *I don't know where Mary is.*

Conversely, a direct interrogative clause may contain a declarative subordinate:

 (b) *Do you know that Mary is here ?*

This interaction between declarative and interrogative clauses is particularly present in dialogues, where the logical notion of *answerhood* is as significant as the one of *inference*.

In order to tackle this issue from a formal standpoint, we investigated the properties and possible uses of inquisitive semantics, which is a formal semantic theory based on a logic that provides a uniform treatment of both declarative and interrogative expressions. Valentin Richard is currently working on a semantic model for embedded interrogatives, like (a). This kind of sentence exhibits *de re*/*de dicto* ambiguity due to the attitude matrix verb (here *know*). Some further investigation is required to understand the interactions between this ambiguity, the inquisitive content and the dynamic referential effect of the interrogative pronoun.

7.2 Discourse Dynamics

> **Participants:**   Maxime Amblard, Maria Boritchev, Philippe de Groote, Jacques Jayez,
> Michel Musiol.

7.2.1 Dialogue Modeling

In [17], Maxime Amblard and Maria Boritchev present Dialogues in Games (DinG), a corpus of manual transcriptions of real-life, oral, spontaneous multi-party dialogues between French-speaking players of the board game Catan. The objective is to make available a quality resource for French, composed of long dialogues, to facilitate their studies in the style of [48]. In a general dialogue setting, participants share personal information, which makes it impossible to disseminate the resource freely and openly. In DinG, the attention of the participants is focused on the game, which prevents them from talking about themselves. In addition, they are conducting a study on the nature of the questions in dialogue, through annotation, in order to develop more natural automatic dialogue systems.

Together with Chloé Braud (IRIT), Maxime Amblard and Chuyuan Li design tools to automatically retrieve characteristic features of dialogues. The results are presented in [28], with a special focus on depression. This serious mental illness impacts the way people communicate, especially through their emotions, and, allegedly, the way they interact with others. This work examines depression signals in dialogues, a less studied setting that suffers from data sparsity. It is hypothesized that depression and emotion can inform each other, and the influence of dialogue structure through topic and dialogue act prediction is explored. Using a Multi-Task Learning (MTL) approach, the tasks mentioned above

are learned jointly with dialogue-tailored hierarchical modeling. Experiments are run on the DAIC and DailyDialog corpora (both contain dialogues in English) and show important improvements over state-of-the-art for depression detection (at best 70.6% F1). This demonstrates the correlation of depression with emotion and dialogue organization and the power of MTL to leverage information from different sources. A further extension of this work, currently under review, was worked on by Chuyuan Li during her sabbatical leave at the University of British Columbia in a collaboration with Giuseppe Carenini's group.

Maxime Amblard supervised the Master Thesis of Amandine Decker on topic shifts. Topics play an important role in coherence in dialogue, as what is currently being discussed constrains the possible contributions of the participants, and initiating a topic while the previous one is still under discussion may be confusing without appropriate signals. However, how to actually define the notion of topic is debated in linguistics and not sufficiently discussed in dialogue modeling. A precise description of topics and topic shifts in conversation would contribute to understanding what it is we perceive when we judge a sequence of utterances to be coherent. In order to understand the different mechanisms that license topic shifts in dialogue and the way participants acknowledge them, several pieces of conversation containing topic shifts were analyzed, with a focus on topic shifts based on extra-linguistic content and more unconventional topic shifts where speakers take advantage of word similarities to introduce a new topic with no semantic relation with the previous one.

Amandine Decker has been studying neural network learning on analogies for morphological issues. Analogical proportions are statements of the form "A is to B as C is to D". They support analogical inference and provide a logical framework to address learning, transfer, and explainability concerns. This logical framework finds useful applications in AI and natural language processing. It has been applied to solve morphological analogies using a retrieval approach named ANNr [30]. This deep learning framework encodes structural properties of analogical proportions and relies on a specifically designed embedding model capturing morphological characteristics of words. It is demonstrated that ANNr outperforms the state of the art on 11 languages. ANNr results for Navajo and Georgian, languages on which the model performs worse and better, are analyzed to explore potential correlations between the errors of ANNr and linguistic properties.

7.2.2 Pathological Discourse Modeling

Michel Musiol has obtained was in full-time delegation in the Sémagramme team until September 2022. This proximity made 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.

In the context of the Inria Exploratory Action ODiM, Maxime Amblard and Michel Musiol have been working on this issue with the constitution of a corpus of pathological discourses and the extension of the tool SLoDIM. The theoretical work focused on the formal definition of transactions in dialogue. With Samuel Buchel and Amandine Lecomte, they introduced a dynamic definition of backchannel words which are used to classify the dialogue units. In [27] they focused on the organizations and functions of backchannel verbal behavior in clinical interaction with a patient suffering from schizophrenia. In this corpus study, they addressed the issue of supportive dialogic behavior in clinical interviews. They started from the analysis of language interactions, assuming that the specific framework of the verbal interaction of clinical type is the privileged place of expressing "supportive dialogic behavior". On the basis of an empirical exploration, they proposed the elaboration of a dynamic interlocutory model of what could be interpreted as a form of attentive, supportive listening, based in particular on specific lexical indicators that are the backchannels. They extracted the properties of discourse configurations in which the supportive behavior of one of the interlocutors (e.g., the psychologist) has the effect of leading the speaker (e.g., the patient) to modify their argumentative strategy. The empirical analyses cover the entire recording time of 10 interviews with patients suffering from schizophrenia (collected in psychiatric hospitals) and 10 interviews with control subjects.

Moreover, in [27] they have designed the beginnings of a methodology for the analysis of discourse

disorders that will have the particularity of helping to select the discontinuous sequences that are most likely to be carriers of thought disorders. They anticipated the development of a modeling system based on principles of pragmatic linguistics and formal semantics, which, applied to carefully selected discontinuous discourse sequences, will have a good chance of revealing the nature of the underlying thought disorders. They compared conjectures with the results of a previous study on the discovery of four "proven" types of discontinuous sequences, and showed which of these sequences can thus be considered as carrying thought disorders. They also analyzed these sequences by testing some principles of semantic modeling in order to identify the nature of the disorders and thought operations underlying the relevant discontinuous sequences. They show that discourse thought disorders should not be considered simply as an expression of a dysexecutive syndrome but also as a device likely to affect more complex thought operations such as the inferences involved in the representation system of the conversational context, in the meaning calculus of the utterances and in the speaker's meaning calculus.

In the MePheSTO project (DFKI-Inria AI project), a multimodal perspective on these issues was proposed. A protocol of a clinical multicenter prospective study has been proposed [10].

7.2.3   Cognitive traces of side issues

It is by now widely believed that natural language communication operates at several *levels*. This means that information is distributed across several partially independent dimensions. For instance, a sentence like *My stupid neighbor made noise again*, simultaneously conveys that my neighbor made noise (the truth-conditional content), that the speaker considers he is stupid (an expressive, side issue 1) and that he had made noise before (the presupposition of *again,* side issue 2). While these phenomena have been extensively described from an empirical perspective, there is at present no unified framework for representing their differences and possible interactions under a formal, computational or cognitive point of view.

We examined the *motor* effects of presuppositions, using the convenient lexical material of *factive* verbs, that is, verbs which presuppose the truth of the complement clause. For example, *Mary knows that Paul cheated on the exam* presupposes that Paul cheated on the exam (side issue) and asserts (truth-conditional content) that she believes that. It has been shown that the oral presentation of movement-related verbs like *jump* or *push* elicits some activation in the motor cortex and finally results into an involuntary contraction of the thumb-index arc, which can be recorded by a special electromagnetic cell, called a *grip force sensor*.

We adapted this technique to the case of factive verbs on a series of sentences of the form *Mary knows that Paul throws the ball*, compared with high base-level sentences like *Paul throws the ball* and low base-level sentences like *Paul does not throw the ball*. Summarizing, our results indicate that the sentences with the factive verbs elicit a very similar response to that of high base-level ones, and, as expected, a different response from that of low base-level one. This suggests that, at least for factive verbs, the presupposed status leaves no trace of a special cognitive treatment, which would lead for instance, to a delayed or weaker motor response.

However, when applied to more complex negative sentences like *Mary does not know that Paul throws the ball*, there is no evidence of a motor trace. This is in agreement with observations in the literature suggesting that, under negation, presuppositions are more 'difficult' to process than in simple assertive sentences. More precisely, in the case of motor response, negation interacts with the presupposition, which suggests that truth-conditional content and side issue cognitive processing cannot be totally separated [7].

7.3   Common Basic Resources

Participants: Maxime Amblard, Hee-Soo Choi, Philippe de Groote, Bruno Guillaume,
Guy Perrier, Sylvain Pogodalla, Karën Fort, Valentin Richard.

7.3.1 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 in the development and the improvement of the French data in this international initiative. In 2022, two new versions (2.10 and 2.11) of the UD data were released.

During 2022, they continued 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. In [22], they bring to the fore some advantages to first develop a new treebank in Surface-Syntactic Universal Dependencies (SUD) annotation scheme, even if the goal is to obtain a UD treebank. Theoretical benefits of SUD are presented, as well as UD-compatible SUD innovations. The two-way UD, SUD conversion is explained, as well as the possibility to customize the conversion for a given language. The paper concludes by a practical guide for the development of a SUD treebank.

A website was built to present the framework (guidelines, data). The Sémagramme team 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 to produce the SUD version of all UD available data.

Hee-Soo Choi, Bruno Guillaume and Karën Fort continued working on linguistic typology based on UD annotated data. They used GREW to enrich the UD annotations and studied the respective word order of verbs with their subjects and objects on 74 languages and compared with other linguistic works.[1] They focused on four Greenberg's universal, providing results which constitute new typological information based on large amounts of data that can fill in gaps in the existing databases. The corpus-based approach allowed them to evaluate the consistency between corpora of the same language and showed a great variation according to the corpus types: oral language, written language in newspapers, tweets, poetry, novels, grammars, etc. This work was published at the Syntax Fest in March 2022 [19].

7.3.2 Induction of Descriptive Grammars

The ANR project Autogramm (Induction of descriptive grammar from annorated corpora) started in January 2022. The goal of this project is to automate, as far as possible, the extraction of descriptive grammars and grammatical descriptions from annotated corpora for linguistic and typological studies. The project also promotes the development of treebanks for low-resourced languages, in order to extract quantitative descriptive grammars for these languages.

A new corpus was added to the SUD project in Beja, a Cushitic language spoken in Sudan. This is the first treebank for Beja in the UD/SUD project. It has been built from the conversion and enhancement of an Interlinear Glossed Text (IGT). The paper [26] presents this corpus and describes the choice to use a morph-based annotation and its consequences; the processing chain from an IGT to a morph-based dependency treebank and a word-based treebank; and several interesting constructions in Beja.

With Santiago Herrera and Sylvain Kahane, in [37], the first results regarding the extraction of rules from annotated corpora were presented.

7.3.3 Mapping Lexical Resources

Hee-Soo Choi started her PhD in October 2021 on mapping French heterogeneous lexical resources. Lexical resources are essential for the development of tools and methods for the various tasks of NLP. These resources are heterogeneous in their size, their construction and their level of linguistic description. This variety opens the way to group or link/map these resources. In [34], she presents a state of the art on French lexical resources and discusses the different features of a lexical resource, the resources based on mapping and the approaches used for this purpose.

---

[1] Code web page

7.3.4 FENEC

Karën Fort worked with Alice Millour (Université Paris 8), Yoann Dupont (Université Paris 3) and Alexane Jouglar (former M1 student) on the creation of a balanced sample corpus for French named entity recognition. The created corpus, FENEC, is freely available and presented in a paper which was published at TALN 2022 [39].

7.3.5 Sentence Semantic Similarity Corpus

In the context of the CODEINE ANR project and more specifically of Nicolas Hiebel's PhD thesis, Karën Fort worked with Aurélie Névéol (LISN-CNRS) and Olivier Ferret (CEA) on the creation of a sentence semantic similarity corpus for French on the clinical domain. A paper on the subject was presented at LREC 2022 [25]. An adaptation in French was subsequently presented at TALN 2022 [38].

7.4 Logical Resources

When designing new logical frameworks, there are desiderata to achieve. It may require some prior intuition in order to create the right structure which can fulfill these desiderata. This intuition is especially hard to get when working with formal inference rules. Valentin Richard introduced a graphical language helping to visualize the connections between rules in a sequent calculus [43], called proof tree graphs. In this directed hypergraph, vertices are set of sequents and edges are rules.

7.5 Ethics and biases

Participants: Karën Fort, Maxime Amblard, Marc Anderson.

7.5.1 Ethics@Loria

Karën Fort originated a working group at LORIA for AI ethics (ethics@loria), involving researchers from various teams, including Maxime Amblard, Marc Anderson, Armelle Brun (BIRD), Mathieu d'Aquin (Orpailleur), Christophe Cerisara (Synalp), Anne Bonneau (Multispeech), Slim Ouni (Multispeech) and Abdessamad Imine (Pesto). Aurore Coince helped manage the group. Ethics@loria proposed the Doctoral training on Ethics "Write your dystopia".

7.5.2 Evaluating stereotypes in masked language models in many languages

Karën Fort led the creation of a French corpus of stereotyped/anti-stereotyped pairs of sentences (women cannot drive / men cannot drive) with Aurélie Névéol (LISN-CNRS), Yoann Dupont (Sorbonne Université) and Julien Bezançon (M2 student at Sorbonne). The corpus was created by translating and adapting an American corpus (CrowsPairs), then adding more specific French stereotypes (for example against Gypsies) obtained via a citizen science platform, LanguageArc. It was then used to evaluate the main masked models for French (including CamemBERT and FlauBERT). A paper on the subject was accepted at the A* conference in NLP ACL 2022 [5]. An adaptation in French was subsequently presented at TALN 2022 [40]. The usage of the citizen science platform LanguageArc was detailed in a paper presented at the LREC 2022 workshop on Novel Incentives in Data Collection from People [21].

Karën Fort contacted researchers interested in creating a CrowsPairs corpus for their language, in order to test the language models. The group got bigger and bigger and now includes: Margot Mieskes and Jonathan Baum for German, Claudia Borg for Maltese, Luciana Benotti and Laura Alonso Alemany for Argentinian Spanish, Wolfgang Sebastian Schmeisser Nieto for Spain Spanish, Sergio Zanotto and Matteo Radaelli for Italian, Francielle Vargas for Brazilian Portuguese, Yongjian Chen, with Karën Fort's M1 students Yuyan Qian and Shailynn Xie, for Chinese and the M1 students Sarah Saidi and Thiziri Saci for standard Arabic. The French CrowsPairs team joined the effort (Aurélie Névéol, Yoann Dupont and Julien Bezançon). The group has been working together since July 2022, with different levels of progress. The corpora will be made freely available once they are ready, along with the code to test the language

models and the guidelines we followed for the adaptation. It's important to notice that this work has been performed without any funding.

7.5.3 NLP for NLP and Ethics

Karën Fort worked with Yves Lepage (Waseda University, Japan), Gaël Lejeune (Sorbonne Université) and Fanny Ducel (M1 student at Sorbonne Université) on the evaluation of the application of the Bender's rule in NLP research papers. The Bender's rule states that NLP researchers should name the language they work on, but it is often not applied, especially when the language dealt with is English, giving the false impression that the research can apply to any language or that English is universal. A paper on a comparison between LREC and ACL proceedings was presented at LREC 2022 [20]. Another paper analyzing the situation concerning TALN and ACL was presented at TALN 2022 [35].

Karën Fort originated a research group on the impact of the BigTech companies on NLP (industry presence, potential thematic shifts, participation in paper authorship). This group is composed of Aurélie Névéol (LISN), Saif Mohammad (National Research Council Canada), Mohamed Abdalla (University of Toronto), Terry Lima Ruas and Jan Philip Wahle (Wuppertal University) and Fanny Ducel (M2 student at Sorbonne University). The group regularly gathered from September to December 2022 to perform both an automatic study on all the ACL Anthology papers and a manual one from the ACL 2022 papers. The results are being submitted at ACL 2023.

7.5.4 Ethics in AI Integration into Industry

The ongoing collaboration between Karën Fort and Marc Anderson resulted in the publication of two journal papers in 2022. One is about ethics by design for real in the industrial context of the AI-Proficient project [9]. The other criticizes the "human in" (the loop/command, etc.) terminology and proposes a new grid of analysis of the interaction between the human and the system [1]. They also presented their approach and insights relative to a specific logistics centered Use Case in the project [15] at SOHOMA'22. Marc Anderson presented his view on whether the future of AI ethics is interdisciplinary at the conference Where AI Ethics Should Go [41].

In the context of ethical issues in the AI-Proficient project, Marc Anderson authored a conference paper for IAIL 2022 on ethical issues around the EU AI Act [14]. Marc Anderson also coauthored (with Fernandez et al.) AI-Proficient's Deliverable 4.1: Human-machine interaction and feedback mechanisms (Design and specification) which was publicly released in June 2022, and has coauthored three other project deliverables scheduled for release in spring 2023.

8 Bilateral contracts and grants with industry

8.1 Bilateral Grants with Industry

8.1.1 C&S Group

Participants: Maxime Amblard, Philippe de Groote, Laurine Jeannot.

The Sémagramme team has set up a Cifre contract with CS Group. The subject of the thesis concerns the automatic analysis of software specifications (expressed in natural language) with the aim of extracting a logical representation and detecting possible inconsistencies and ambiguities of the requirements. The thesis started in 2021, and prematurely ended for health reasons in 2022.

8.1.2 Yseop

Participants: Philippe de Groote, Maxime Guillaume, Sylvain Pogodalla.

The Sémagramme team has set up a Cifre thesis contract with Yseop on ACG extensions and use in an industrial environment.

9 Partnerships and cooperations

9.1 International initiatives

9.1.1 Participation in other International Programs
IMPRESS

Participants: Philippe de Groote, Sylvain Pogodalla, Priyansh Trivedi.

Title: Improving Embeddings with Semantic Knowledge

Partner Institutions:

- DFKI, Saarbrücken, Germany
- MAGNET, Inria Lille-Europe
- Multispeech, Inria Nancy-Grand Est
- Sémagramme, Inria Nancy-Grand Est

Date/Duration: October 2020–September 2023

Coordination: Pascal Denis (MAGNET) and Ivana Kruijff-Korbayova (DFKI)

Additionnal info/keywords: The IMPRESS project aims at investigating 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.

MePheSTO

Participants: Maxime Amblard, Amandine Lecomte, Michel Musiol.

Title: Digital Phenotyping 4 Psychiatric Disorders from Social Interaction

Partner Institutions:

- DFKI, Saarbrücken, Germany
- Sémagramme, Inria Nancy-Grand Est
- StARS, Inria Sophia Antipolis-Méditerranée

Date/Duration: 2020–2023

Coordination: Maxime Amblard (Sémagramme) and François Bremond (StARS)

Additionnal info/keywords: MePheSTO 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.

External Partners:

- Nice University Clinic (Prof. Dr. Philippe Robert, University Côte d'Azur)
- University Clinic of Saarland (Prof. Dr. Matthias Riemenschneider)
- Centre Psychothe'rapique de Nancy (Prof. Dr Vincent Laprevote)
- Dept. Of Psychiatry, University of Oldenburg, Karl Jaspers Klinik (Prof. Dr. Dr. Rene' Hurlemann)

9.1.2 International research visitors

Amandine Decker

Visited institution: Gothenburg University

Country: Suède

Dates: 03/01/2021 - 08/31/2022

Context of the visit: Amandine Decker realised her Master thesis under the join supervision of Maxime Amblard and Ellen Breitholtz. She visited Gothenburg University for a few months.

Mobility program/type of mobility: Master program in NLP and Erasmus exchange

Chuyuan Li

Visited institution: British Columbia University

Country: Canada

Dates: 09/01/2021 - 02/28/2022

Context of the visit: The Université de Lorraine offers an exchange program for PhD students. With Chloé Braud and Maxime Amblard, Chuyuan Li has proposed a research program to work on uptodate Dep Learning approaches for Dialogue parsing.

Mobility program/type of mobility: Long stays program: DReaM (Université de Lorraine)

William Babonnaud

Visited institution: Heinrich-Heine-Universität Düsseldorf, Institut für Linguistik

Country: Germany

Dates: Regular stays between September 2021 and November 2022

Context of the visit: William Babonnaud's fourth year of PhD has been funded by the Deutsche Forschungsgemeinschaft (German Research Foundation) as part of the project "Coercion and Copredication as Flexible Frame Composition", setting up a collaboration with the project team.

Mobility program/type of mobility: Research stays

9.2   European initiatives

9.2.1   H2020 projects

AI-Proficient

Participants: Marc Anderson, Karën Fort.

Title: Artificial intelligence for improved production efficiency, quality and maintenance

Duration: November 2020–October 2023

Coordinator: Benoit Iung (CRAN, Université de Lorraine)

Partners: Université de Lorraine (coordination), Continental (industrial partner), Ineos (industrial partner), Institute Mihailo Pupin (Serbia), Tekniker (Spain), Ibermatica (Spain), TenForce (Belgium), VTT (Finland), Inos Hellas (Greece), ATC Athens Technology Center (Greece)

Participants: Marc Anderson, Karën Fort

Abstract: AI-Proficient carries out research on integrating AI services into the industrial contexts of three factories located in France, Belgium, and Germany. Two Sémagramme members make up the ethics team for the project since its beginning in 2020: Karën Fort (project ethics officer) and Marc Anderson (postdoctoral fellow).

9.3   National initiatives

9.3.1   INRIA Exploratory Action: ODiM

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

Title: Outils informatisés d'aide au Diagnostic des Maladies mentales

Duration: 2019–2022

Coordinator: Maxime Amblard

Partners: Inria, Université de Lorraine, Psychiatric hospital Montperrin (Aix-en-Provence)

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

Abstract: ODiM is an interdisciplinary project, which mixed psychiatry-psychopathology, linguistics, formal semantics and digital sciences. The goal was to theoretically replace the paradigm of language and thought disorders (LTD) used in the Mental Health sector with a semantic-formal and cognitive model of Discourse disorders (SD). These disorders are translated into pathognomonic signs, making them complementary tools for diagnosing and screening vulnerable individuals before the onset of psychosis.

9.3.2   ANR Project: CoDeinE

Participants: Karën Fort, Bruno Guillaume.

Title: artificial text COrpus DEsIgNed Ethically automatic synthesis of clinical documents

Duration: March 2021–February 2025

Coordinator: Aurélie Névéol (Limsi)

Partners: CRC, CEA List, LISN, LORIA

Participants: Bruno Guillaume, Karën Fort (local coordinator)

Abstract: Machine learning methods have become prevalent in language technologies. They rely on annotated corpora to train models and evaluate algorithms. The CoDeinE project proposes to address the lack of shareable corpora in sensitive domains such as health or banking. The key idea of the project is to use confidential corpora to automatically generate synthetic texts that mimic the linguistic properties of real documents while preserving confidentiality. We will use clinical documents in electronic patient records as a case study. Furthermore, the project will rely on Games With A Purpose and crowd sourcing to validate and annotate the synthesized texts.

9.3.3   ANR Project: Autogramm

Participants: Bruno Guillaume, Karën Fort, Guy Perrier.

Title: Induction of descriptive grammar from annotated corpora

Duration: January 2022–December 2025

Coordinator: Sylvain Kahane (Université Paris Nanterre)

Partners: MoDyCo, LACITO, LISN, Inria Nancy - Grand Est

Participants: Bruno Guillaume (local coordinator), Karën Fort, Guy Perrier

Abstract: The goal of this project is to automate, as far as possible, the extraction of descriptive grammars and grammatical descriptions from annotated corpora for linguistic and typological studies. The project also promotes the development of treebanks for under-endowed languages, in order to extract quantitative descriptive grammars for these languages. The project uses the annotation scheme SUD (Surface-syntactic Universal Dependencies), the query tool Grew-match and the annotation tool ArboratorGrew.

10   Dissemination

Participants: Maxime Amblard, Marc Anderson, Hee-Soo Choi, Marie Cousin, Amandine Decker, Philippe de Groote, Karën Fort, Bruno Guillaume, Jacques Jayez, Chuyuan Li, Pierre Ludmann, Michel Musiol, Guy Perrier, Sylvain Pogodalla, Valentin Richard, Vincent Tourneur.

10.1   Promoting scientific activities

10.1.1   Scientific events: organisation

General chair, scientific chair

- Michel Musiol: Chair and moderator of the session 'Pour constituer leur objet, les neurosciences peuvent-elles se passer d'un point de vue psychanalytique ?', 1st International Colloquium of the Réseau Francophone Psychanalyse Et Neurosciences, 'Dialogues between Psychoanalysis and Neurosciences: a new approach to psychic life', University of Lorraine, Nancy, 18th November 2022 and 19th November 2022.

Member of the organizing committees

- Marc Anderson: member of the organizing committee of the AI for Manufacturing (AI4M) Workshop - Hosted by DAAM International Symposium (October 2022).

- Karën Fort: co-organizer, with Sylvain Loiseau (LACITO) and Berthold Crysman (LLF) of the GDR LIFT summer school on Annotation. From 30th May 2022 to 3rd June 2022. .

10.1.2 Scientific events: selection

Area Chair of conference

- Karën Fort: Area Chair for the Ethics in NLP track for COLING 2022.

Member of the conference program committees

- Marc Anderson: member of the program committee of the AI for Manufacturing (AI4M) Workshop - Hosted by DAAM International Symposium (October 2022).

Reviewer

- Maxime Amblard: ARR 2022, TALN 2022, Coling 22, EMNLP 22.

- Marc Anderson: APMS Smart Manufacturing and Logistics Systems 2022, SOHOMA 2022.

- Karën Fort: AACL 2022, LREC 2022, TALN 2022, CMLF 2022, Games and NLP 2022, Journées conjointe des GDR LIFT et TAL 2022, journée ATALA RobusTAL.

- Sylvain Pogodalla: ACL Rolling Review, Conférence Nationale en Intelligence Artificielle collection (CNIA 2022), Traitement Automatique des Langues Naturelles 2022 (TALN 2022).

- Siyana Pavlova: Coling 22.

10.1.3 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.

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

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

Reviewer - reviewing activities

- Maxime Amblard: Traitement Automatique des Langues, Journal of Logic, Language and Information, Computer Speech & Language.

- Marc Anderson: Transactions of the Charles S. Peirce Society, Dewey Studies.

- Karën Fort: PLOS ONE.

- Sylvain Pogodalla: Traitement Automatique des Langues, Logical Methods in Computer Science.

10.1.4   Invited talks

- Karën Fort: invited to present a state of the art on ethics in NLP at the joint GDR LIFT and TAL conference in Marseille, on 14th November 2022: "Éthique et TAL : ce dont on parle, ce dont on ne parle plus, ce dont on ne parle pas (un état de l'art)" [36].

- Jacques Jayez: invited talk at the *Implicit Manipulation in Politics - Quantitatively Assessing the Tendentiousness of Speeches* (IMPAQTS) conference , 27th April 2022 and 28th April 2022: "(Innocent?) Bias inside language".

10.1.5   Invited seminars

- Marc Anderson: 15th December 2022, Department ISET - CRAN, UL "Éthique dès la conception dans un projet d'informatique industrielle : l'exemple du projet AI-Proficient".

- Hee-Soo Choi: 20th May 2022, GRLMC, Universitat Rovira i Virgili, Spain "Investigating Dominant Word Order on Universal Dependencies with Graph Rewriting".

- Karën Fort: OFAI seminar (Austria) on 12th October 2022: "Ethics and NLP: What we Talk About, What we Don't Talk About Anymore, What we Never Talked About".

- Karën Fort: IXXI seminar on 10th June 2022: "Éthique et TAL : ce dont on parle, ce dont on ne parle plus, ce dont on ne parle pas (un état de l'art)".

- Karën Fort: panel on teaching ethics at the ERCIM project meeting.

10.1.6   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.

- Marc Anderson: Josiah Royce Society Board Member - Representative at Large (from 2020).

- Karën Fort: Co-chair of the ACL Ethics committee (2021–2026).

- Karën Fort: member of the ethics committee of the LIAvignon Chaire.

- Karën Fort: principal investigator of LIFT 2, a follow-up to the LIFT GDR (Linguistique Informatique, Formelle et de Terrain) to start on 2024.

10.1.7   Scientific expertise

- Maxime Amblard: evaluation for the Université Grenoble-Alpes (internal call for project proposals, allowing for PhD position funding).

- Karën Fort: evaluation for the Deutsche Forschungsgemeinschaft (German Research Foundation), evaluation for the Marie Sklodowska-Curie Actions COFUND / C2W (Université de Mons, Belgique).

10.1.8   Research administration

- Maxime Amblard:

  – Member of conseil d'administration of Université de Lorraine.

  – 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 grade promotion committee 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).

- Karën Fort

  - Member of CNU 27 (Computer Science): participation to qualifications, promotions and suivi de carrière.

- 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.

10.2 Teaching - Supervision - Juries

10.2.1 Teaching

- Licence:

  - Maxime Amblard, AI Introduction, 15h, L1, Université de Lorraine, France.
  - Maxime Amblard and Chuyuan Li NLP for beginners, 20h, L2, Université de Lorraine, France.
  - Maxime Amblard and Chuyuan Li Linguistic engineering, 20h, L3, Université de Lorraine, France.
  - Hee-Soo Choi, Corpus processing (introduction to Python), 24h, Université de Lorraine, France.
  - Karën Fort, Relational databases, 55h, L3, Sorbonne Université, France.
  - Pierre Ludmann, Informatics 2, 27h, Mines Nancy, Université de Lorraine, France.
  - Pierre Ludmann, Databases, 114h, Polytech Nancy, Université de Lorraine, France.
  - Pierre Ludmann, Java project, 16h, Polytech Nancy, Université de Lorraine, France.
  - Pierre Ludmann, Year-long project, 20h, Polytech Nancy, Université de Lorraine, France.
  - Pierre Ludmann, Mentoring, 40h, Polytech Nancy, Université de Lorraine, France.
  - Valentin Richard, Langages de Scripts, 50h, L3, Université de Lorraine, France.

- Master:

  - Maxime Amblard and Siyana Pavlova, Python Programming, 30h, M1 NLP (IDMC), Université de Lorraine, France.
  - Maxime Amblard and Siyana Pavlova, Methods for NLP, 36h, M1 NLP (IDMC), Université de Lorraine, France.
  - Maxime Amblard, NLP project, 20h, M1 NLP (IDMC), Université de Lorraine, France.
  - Maxime Amblard, Marie Cousin, and Amandine Decker, Formalisms and Syntax, 24h, M2 NLP (IDMC), Université de Lorraine, France.
  - Maxime Amblard and Valentin Richard, Discourse and Dialogue, 18h, M2 NLP (IDMC), Université de Lorraine, France.
  - Maxime Amblard, software project, 10h, M2 NLP (IDMC), Université de Lorraine, France.

- – Marie Cousin, Agile Method and Scrum, 4h, M2 NLP (IDMC), Université de Lorraine, France.

- – Amandine Decker, LaTeX/Shell, 6h, M1 NLP (IDMC), Université de Lorraine, France.

- – Philippe de Groote, Formal Logic, 22h, M1 NLP (IDMC), Université de Lorraine, France.

- – Philippe de Groote, Formal languages, 22h, M1 NLP (IDMC), Université de Lorraine, France.

- – Philippe de Groote, Computational Semantics, 18h, M2 NLP (IDMC), Université de Lorraine, France.

- – Philippe de Groote, Computational structures and logics for natural language modeling, 18h, M2 NLP (IDMC), Université Paris Diderot – Paris 7, France.

- – Karën Fort, Data Ethics, 18h, M1 ISF (Informatique et Statistique financières), *Université Panthéon Assas*, France.

- – Karën Fort, Ethics and NLP (English), 17h30, M2 NLP (IDMC), Université de Lorraine, France.

- – Karën Fort, Introduction to ethics, 5h, M2 CogSci (IDMC), Université de Lorraine, France.

- – Karën Fort, Formal Grammar, 39h, M1, *Sorbonne Université*, France.

- – Karën Fort, Corpora, resources and tools for linguistics, 39h, M1, *Sorbonne Université*, France.

- – Karën Fort, Ethics and NLP, 15h, M2, *Sorbonne Université*, France.

- – Karën Fort, Ethics and NLP (English), 12h, M1/M2, *University of Malta*, Malta.

- – Karën Fort, Collaborative annotation for NLP, 30h, M2, *Sorbonne Université*, France.

- – Karën Fort, Crowdsourcing for NLP, 3h, M2, *Nanterre Université*, France.

- – Bruno Guillaume, Written Corpora TAL (English), 45h, M1 NLP (IDMC), Université de Lorraine, France.

- – Chuyuan Li, Introduction to G5K (English), 4h, M1 NLP (IDMC), Université de Lorraine, France.

- – Pierre Ludmann, Software Engineering 2, 10h, Mines Nancy, Université de Lorraine, France.

- – Pierre Ludmann, Compilation, 27.5h, Mines Nancy, Université de Lorraine, France.

- – Pierre Ludmann, Department project, 5h, Mines Nancy, Université de Lorraine, France.

- – Pierre Ludmann, Software Engineering, 42h, Polytech Nancy, Université de Lorraine, France.

- – Pierre Ludmann, Web project, 6h, Polytech Nancy, Université de Lorraine, France.

- – Pierre Ludmann, End-of-study interships, 20h Polytech Nancy, Université de Lorraine, France.

- – Pierre Ludmann, End-of-study project, 20h Polytech Nancy, Université de Lorraine, France.

- – Pierre Ludmann, Mentoring, 48h, Polytech Nancy, Université de Lorraine, France.

- – Vincent Tourneur, UML beginners (English), 10h, M1 NLP (IDMC), Université de Lorraine, France.

- – Siyana Pavlova, Data Science, 9h, M1 NLP (IDMC), Université de Lorraine, France.

- Doctorate:

  - – Karën Fort, Scientific integrity, 6h, *École doctorale 5, Faculté des lettres, Sorbonne Université*.

  - – Karën Fort, Annotation Quality, 4h30, *École d'été* of the GDR LIFT, May 2022, Banyuls-sur-Mer, France.

  - – Karën Fort, Maxime Amblard, Marc Anderson, Ethics training: write your dystopia, 2 days, 3rd November 2022 and 4th November 2022, *École doctorale IAEM*, Nancy, France.

- International Summer School:

  - – Philippe de Groote with Yoad Winter, Introduction to natural language formal semantics, at the 32nd European Summer School in Logic, Language and Information (ESSLLI 2021).

10.2.2 Supervision

HDR • Karën Fort, *Myriadisation et éthique pour le traitement automatique des langues.* Defended on 23rd November 2022 [45].

PhD • William Babonnaud, *Sémantique lexicale, compositionnalité et coercions. Fondements théoriques des types sémantiques.* Defended on 22nd November 2022 [44].

PhD in progress

- Samuel Buchel, *Linguistic, Semantic and Cognitive Modelling of Dialogical Incongruities and Discontinuities in The Interaction with The Schizophrenic Patients,* since December 2019. Supervision: Maxime Amblard and Michel Musiol.
- Hee-Soo Choi, *Lier des ressources lexicales du français en vue d'une interopérabilité entre niveaux linguistiques,* since October 2021. Supervision: Karën Fort and Mathieu Constant.
- Marie Cousin, *Modélisation de paraphrase dans les grammaires catégorielles abstraites,* since October 2022. Supervision: Philippe de Groote and Sylvain Pogodalla.
- Amandine Decker, *Modelling Topic-level Interaction in Pathological Conversations,* since October 2022. Supervision: Maxime Amblard and Ellen Breitholtz (University of Gothenburg, Sweden).
- Maxime Guillaume, *Structures de traits pour les Grammaires Catégorielles Abstraites,* since July 2021. Supervision: Philippe de Groote and Raphaël Salmon (Yseop).
- Laurine Jeannot, *Robust Semantic and Discourse Analysis of Natural Language System Specifications,* from January 2021 to October 2022. Supervision: Maxime Amblard and Philippe de Groote.
- Amandine Lecomte, *Analyse longitudinale de prise en charge psychothérapeutique de patients psychiatriques et de patients atteints de maladies neurodégénératives : informatisation et modélisation dialogique des indices comportementaux associés à l'efficacité (vs échec) des stratégies de prise en charge tentées par les thérapeutes,* since October 2019. Supervision: Michel Musiol and Alexandra König.
- Chuyuan Li, *Formal and Statistical Modeling of Dialogue,* since October 2019. Supervision: Maxime Amblard and Chloé Braud (IRIT).
- Pierre Ludmann, *Dynamic Construction of Discursive Structures,* since September 2017. Supervision: Philippe de Groote and Sylvain Pogodalla.
- Siyana Pavlova, *Tools and Methods for Semantic Annotation,* since November 2020. Supervision: Maxime Amblard and Bruno Guillaume.
- Valentin Richard, *Aspects compositionnels et dynamiques de la sémantique inquisitrice,* since September 2021. Supervision: Philippe de Groote and Reinhart Muskens (Universiteit van Amsterdam, ILLC).
- Priyansh Trivedi, *Injecting Lexical and Semantic Knowledge into Word, Phrasal and Sentence Embeddings,* since November 2021. Supervision: Philippe de Groote and Pascal Denis.

10.2.3 Juries

- Karën Fort: member of the PhD jury of Anaëlle Baledent (examiner), "De la complexité de l'annotation manuelle : méthodologie, biais et recommandations", *Université de Caen Normandie,* 1st December 2022.
- Guy Perrier: reviewer of Rita Hijazi's PhD thesis, "Simplification syntaxique de textes à base de représentations sémantiques exprimées avec DMRS", *Université d'Aix-Marseille,* 14th December 2022.

10.3 Popularization

- Maxime Amblard is a member of the scientific committee of)i(interstices.

10.3.1 Interventions

- Marie Cousin and Amandine Decker: 17th November 2022, presentation to high school students in the context of the "Chiche !" initiative (lycée Saint-Exupéry, Fameck, France).

- Karën Fort: 30th September 2022, Radio Campus France, Têtes Chercheuses podcast: interview on crowdsourcing.

- Karën Fort: 30th September 2022, Nuit européenne des Chercheurs in Nancy: impro lab with the Crache-Texte company "Le miroir virtuel : La société face aux algorithmes".

- Valentin Richard gave a talk [46] at LORIA's PhD Seminar on 5th October 2022.

11 Scientific production

11.1 Major publications

[1] M. Anderson and K. Fort. 'Human Where? A New Scale Defining Human Involvement in Technology Communities from an Ethical Standpoint'. In: *International Review of Information Ethics* (25th Aug. 2022). URL: https://hal.inria.fr/hal-03762035.

[2] 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.

[3] 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.

[4] 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.

[5] A. Névéol, Y. Dupont, J. Bezançon and K. Fort. 'French CrowS-Pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English'. In: ACL 2022 - 60th Annual Meeting of the Association for Computational Linguistics. Dublin, Ireland, 22nd May 2022. URL: https://hal.inria.fr/hal-03629677.

[6] 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.

[7] R. Reinecke, T. A. Nazir, S. Carvallo and J. Jayez. 'Factives at hand: When presupposition mode affects motor response'. In: *Journal of Experimental Psychology* (2022). DOI: 10.1037/xge0001167. URL: https://hal.archives-ouvertes.fr/hal-03538732.

11.2 Publications of the year

International journals

[8] M. Anderson and K. Fort. 'Human Where? A New Scale Defining Human Involvement in Technology Communities from an Ethical Standpoint'. In: *International Review of Information Ethics* (25th Aug. 2022). URL: https://hal.inria.fr/hal-03762035.

[9] M. M. Anderson and K. Fort. 'From the Ground Up: Developing a Practical Ethical Methodology for Integrating AI into Industry'. In: *AI & Society: Knowledge, Culture and Communication* (11th July 2022). DOI: 10.1007/s00146-022-. URL: https://hal.inria.fr/hal-03754737.

[10] A. König, P. Müller, J. Tröger, H. Lindsay, J. Alexandersson, J. Hinze, M. Riemenschneider, D. Postin, E. Ettore, A. Lecomte, M. Musiol, M. Amblard, R. Hurlemann, F. F. Bremond and M. Balazia. 'Multimodal phenotyping of psychiatric disorders from social interaction: Protocol of a clinical multicenter prospective study'. In: *Personalized Medicine in Psychiatry* 33-34 (July 2022), p. 100094. DOI: 10.1016/j.pmip.2022.100094. URL: https://hal.inria.fr/hal-03724844.

[11]   M. Musiol, M. Rebuschi, S. Buchel, A. Lecomte, P. de Groote and M. Amblard. 'The Issue of the Analysis of Thought Disorders in the Discourse of Schizophrenic Subjects: A Methodological Proposition'. In: *L'Évolution Psychiatrique* 87.2 (13th Apr. 2022), pp. 347–369. DOI: `10.1016/j.evo psy.2022.04.004`. URL: `https://hal.inria.fr/hal-03724030`.

[12]   R. Reinecke, T. A. Nazir, S. Carvallo and J. Jayez. 'Factives at hand: When presupposition mode affects motor response'. In: *Journal of Experimental Psychology* 151.9 (Sept. 2022), pp. 2173–2194. DOI: `10.1037/xge0001167`. URL: `https://hal.science/hal-03538732`.

International peer-reviewed conferences

[13]   M. Amblard, B. Guillaume, S. Pavlova and G. Perrier. 'Graph Querying for Semantic Annotations'. In: ISA 2022 : 18th Joint ACL - ISO Workshop on Interoperable Semantic Annotation. Marseille, France, 20th June 2022. URL: `https://hal.inria.fr/hal-03724068`.

[14]   *Best Paper*
       M. M. Anderson. 'Some Ethical Reflections on the EU AI Act'. In: *CEUR Workshop Proceedings*. IAIL 2022: 1st International Workshop on Imagining the AI Landscape After the AI Act. Vol. 3221. Amsterdam, Netherlands, Sept. 2022. URL: `https://hal.inria.fr/hal-03921857`.

[15]   M. M. Anderson and K. Fort. 'Ethical Internal Logistics 4.0: Observations and Suggestions from a Working Internal Logistics Case'. In: 12th International Workshop on Service Oriented, Holonic and Multi-Agent Manufacturing Systems for Industry of the Future. Bucarest, Romania, 22nd Sept. 2022. URL: `https://hal.inria.fr/hal-03768317`.

[16]   T. Anikina, N. Skachkova, J. Renner and P. Trivedi. 'Anaphora Resolution in Dialogue: System Description (CODI-CRAC 2022 Shared Task)'. In: CODI-CRAC 2022. Gyeongju, South Korea, Oct. 2022. URL: `https://hal.inria.fr/hal-03925147`.

[17]   M. Boritchev and M. Amblard. 'A Multi-Party Dialogue Ressource in French'. In: LREC 2022 - 13th Edition of Language Resources and Evaluation Conference. Marseille, France, 20th June 2022. URL: `https://hal.inria.fr/hal-03724046`.

[18]   H. Bunt, M. Amblard, J. Bos, K. Fort, P. de Groote, B. Guillaume, C. Li, P. Ludmann, M. Musiol, S. Pavlova, G. Perrier and S. Pogodalla. 'Quantification Annotation in ISO 24617-12, Second Draft'. In: Proceedings of the 13th Language Resources and Evaluation Conference (LREC 2022). Marseille, France: European Language Resources Association, 20th June 2022, pp. 3407–3416. URL: `https://hal.inria.fr/hal-03724057`.

[19]   H.-S. Choi, B. Guillaume and K. Fort. 'Corpus-based Language Universals Analysis using Universal Dependencies'. In: SyntaxFest Quasy 2021 - Quantitative Syntax. Proceedings of the Second Workshop on Quantitative Syntax (Quasy, SyntaxFest 2021). Sofia, Bulgaria, 21st Mar. 2022. URL: `https://hal.inria.fr/hal-03501774`.

[20]   F. Ducel, K. Fort, G. Lejeune and Y. Lepage. 'Do we Name the Languages we Study? The #BenderRule in LREC and ACL articles'. In: LREC 2022 - International Conference on Language Resources and Evaluation (LREC). Marseille, France, 20th June 2022. URL: `https://hal.inria.fr/hal-036805 61`.

[21]   K. Fort, A. Névéol, Y. Dupont and J. Bezançon. 'Use of a Citizen Science Platform for the Creation of a Language Resource to Study Bias in Language Models for French: a case study'. In: 2nd LREC Workshop on Novel Incentives in Data Collection from People. Marseille, France, 25th June 2022. URL: `https://hal.inria.fr/hal-03693686`.

[22]   K. Gerdes, B. Guillaume, S. Kahane and G. Perrier. 'Starting a new treebank? Go SUD! Theoretical and practical benefits of the Surface-Syntactic distributional approach'. In: SyntaxFest Depling 2021 - 6th International Conference on Dependency Linguistics. Sofia, Bulgaria, 21st Mar. 2022. URL: `https://hal.inria.fr/hal-03509136`.

[23]   G. Greco, M. Moortgat, V. D. Richard and A. Tzimoulis. 'Lambek-Grishin Calculus: focusing, display and full polarization'. In: Topology, Algebra, and Categories in Logic (TACL 2022). Coimbra, Portugal, 20th June 2022, pp. 79–81. URL: `https://hal.science/hal-03727761`.

[24]   P. de Groote. 'Deriving formal semantic representations from dependency structures'. In: *Proceedings of the 19th International Workshop of Logic and Engineering of Natural Language Semantics 19 (LENLS19*. LENLS19 - Logic and Engineering of Natural Language Semantics 19. Tokyo, Japan, Nov. 2022, pp. 87–91. URL: https://hal.inria.fr/hal-03940888.

[25]   N. Hiebel, O. Ferret, K. Fort and A. Névéol. 'CLISTER: A corpus for semantic textual similarity in French clinical narratives'. In: LREC 2022 - 13th Language Resources and Evaluation Conference. Vol. 2022. LREC 2022 - Proceedings of the 13th Conference on Language Resources and Evaluation. Marseille, France: European Language Resources Association, 2022, pp. 4306–4315. URL: https://hal-cea.archives-ouvertes.fr/cea-03740484.

[26]   S. Kahane, M. Vanhove, R. Ziane and B. Guillaume. 'A morph-based and a word-based treebank for Beja'. In: *SyntaxFest*. TLT 2021 - 20th International Workshop on Treebanks and Linguistic Theories. Sofia, Bulgaria, 2021. URL: https://hal.science/hal-03494462.

[27]   A. Lecomte, S. Buchel, N. Franck, C. Demily, M. Amblard and M. Musiol. 'Organisations et fonctions du comportement verbal de type "backchannels" dans l'interaction clinique avec la personne souffrant de schizophrénie'. In: 8eme Congrès mondial de linguistique française. Orléans, France, 4th July 2022. URL: https://hal.inria.fr/hal-03724063.

[28]   C. Li, C. Braud and M. Amblard. 'Multi-Task Learning for Depression Detection in Dialogs'. In: *à paraître*. 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2022). Edinbourg, United Kingdom, 7th Sept. 2022, pp. 1–8. DOI: 10.48550/arXiv.2208.10250. URL: https://hal.inria.fr/hal-03724305.

[29]   P. Ludmann, S. Pogodalla and P. de Groote. 'Multityped Abstract Categorial Grammars and Their Composition'. In: WoLLIC 2022 - 28th International Workshop on Logic, Language, Information, and Computation. Vol. 13468. Lecture Notes in Computer Science. Iaşi, Romania: Springer International Publishing, 9th Sept. 2022, pp. 105–122. DOI: 10.1007/978-3-031-15298-6_7. URL: https://hal.inria.fr/hal-03781596.

[30]   E. Marquer, S. Alsaidi, A. Decker, P.-A. Murena and M. Couceiro. 'A Deep Learning Approach to Solving Morphological Analogies'. In: *Lecture Notes in Computer Science*. 30th International Conference on Case-Based Reasoning (ICCBR2022). Nancy, France: Springer, 12th Sept. 2022, pp. 159–174. DOI: 10.1007/978-3-031-14923-8_11. URL: https://hal.inria.fr/hal-0366 0625.

[31]   A. Névéol, Y. Dupont, J. Bezançon and K. Fort. 'French CrowS-Pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English'. In: ACL 2022 - 60th Annual Meeting of the Association for Computational Linguistics. Dublin, Ireland, 22nd May 2022. URL: https://hal.inria.fr/hal-03629677.

[32]   S. Pavlova, M. Amblard and B. Guillaume. 'How much of UCCA can be predicted from AMR?' In: ISA 2022 : 18th Joint ACL - ISO Workshop on Interoperable Semantic Annotation. Marseille, France, 20th June 2022. URL: https://hal.inria.fr/hal-03724129.

[33]   V. D. Richard. 'Simulating Feature Structures with Simple Types'. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop. Hybrid: Seattle, United States: Association for Computational Linguistics, July 2022, pp. 113–122. DOI: 10.18653/v1/2022.naacl-srw.15. URL: https://hal.science/hal-03943681.

National peer-reviewed Conferences

[34]   H.-S. Choi. 'State of the art : Linking French Lexical Resources'. In: TALN + RECITAL 2022 - 29e Conférence sur le Traitement Automatique des Langues Naturelles. 24e Rencontres Etudiants Chercheurs en Informatique pour le pour le Traitement Automatique des Langues (RECITAL). Avignon, France, 2022, pp. 55–68. URL: https://hal.science/hal-03705816.

[35]   F. Ducel, K. Fort, G. Lejeune and Y. Lepage. 'Contrastive and diachronic study of unmentioned (by default ?) languages in TALN and ACL'. In: Conférence sur le Traitement Automatique des Langues Naturelles (TALN). Avignon, France, 27th June 2022. URL: https://hal.inria.fr/hal-0368056 5.

[36]   K. Fort. 'Éthique et TAL : ce dont on parle, ce dont on ne parle plus, ce dont on ne parle pas (un état de l'art)'. In: LIFT TAL 2022 - Journées Jointes des Groupements de Recherche « Linguistique Informatique, Formelle et de Terrain » et « Traitement Automatique des Langues ». Marseille, France, 14th Nov. 2022. URL: https://hal.inria.fr/hal-03943792.

[37]   S. Herrera, S. Kahane and B. Guillaume. 'Grammar rules extraction from treebanks : a tool and some first results'. In: *Actes des journées jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain (LIFT) et Traitement Automatique des Langues (TAL)*. Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain (LIFT) et Traitement Automatique des Langues (TAL). Marseille, France: CNRS, 2022, pp. 93–98. URL: https://hal.sc ience/hal-03846825.

[38]   N. Hiebel, O. Ferret, K. Fort and A. Névéol. 'CLISTER : A Corpus for Semantic Textual Similarity in French Clinical Narratives'. In: Conférence sur le Traitement Automatique des Langues Naturelles (TALN). Avignon, France, 27th June 2022. URL: https://hal.inria.fr/hal-03680571.

[39]   A. Millour, Y. Dupont, A. Jouglar and K. Fort. 'FENEC : a balanced sample corpus for French named entity recognition'. In: Conférence sur le Traitement Automatique des Langues Naturelles (TALN). Avignon, France, 27th June 2022. URL: https://hal.inria.fr/hal-03680569.

[40]   A. Névéol, Y. Dupont, J. Bezançon and K. Fort. 'French CrowS-Pairs : Extending a challenge dataset for measuring social bias in masked language models to a language other than English'. In: Conférence sur le Traitement Automatique des Langues Naturelles (TALN). Avignon, France, 27th June 2022. URL: https://hal.inria.fr/hal-03680574.

Conferences without proceedings

[41]   M. M. Anderson. 'Is the Future of AI Ethics Interdisciplinary?' In: Where AI Ethics Should Go. Tubingen, Germany, 30th June 2022. URL: https://hal.inria.fr/hal-03922361.

[42]   P. de Groote. 'Ty2 revisited'. In: HyLo 2022 — Workshop on Hybrid Logic and Applications. Chania, Greece, 6th Apr. 2022. URL: https://hal.inria.fr/hal-03940832.

[43]   V. D. Richard. 'Introducing Proof Tree Automata and Proof Tree Graphs: Extended abstract'. In: SLSS 2022 - 11th Scandinavian Logic Symposium. Bergen, Norway, 17th June 2022, pp. 10–12. URL: https://hal.science/hal-03698581.

Doctoral dissertations and habilitation theses

[44]   W. Babonnaud. 'Lexical Semantics, compositionality and coercions. Theoretical foundations of semantic types'. Université de Lorraine, 22nd Nov. 2022. URL: https://hal.science/tel-0393 5669.

[45]   K. Fort. 'Crowdsourcing and ethics for Natural Language Processing'. ED n°77 : Informatique - Automatique - Électronique - Électrotechnique - Mathématiques de Lorraine (IAEM-Lorraine), 23rd Nov. 2022. URL: https://theses.hal.science/tel-03873000.

Reports & preprints

[46]   V. D. Richard. *Is Natural Language Processing (NLP) Going in the Right Direction?: Getting started with NLP ethics*. 5th Oct. 2022. URL: https://hal.science/hal-03813509.

Other scientific publications

[47]   M. Cousin. 'Text generation with abstract categorial grammars and the meaning-text theory'. Vandoeuvre-les -Nancy: Grenoble INP Ensimag, 5th Sept. 2022, p. 69. URL: https://hal.inria.f r/hal-03942766.

11.3   Cited publications

[48]   N. Asher, J. Hunter, M. Morey, B. Farah and S. Afantenos. 'Discourse Structure and Dialogue Acts in Multiparty Dialogue: the STAC Corpus'. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 2721–2727. URL: https://aclanthology.org/L16-1432.

[49]   P. de Groote. 'Towards a Montagovian account of dynamics'. In: *Proceedings of the 16th Semantics and Linguistic Theory Conference (SALT 16)*. Ed. by M. Gibson and J. Howell. 2006. DOI: 10.3765/salt.v16i0.2952.

[50]   V. Dobrovolskii. 'Word-Level Coreference Resolution'. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 7670–7675. DOI: 10.18653/v1/2021.emnlp-main.605. URL: https://aclanthology.org/2021.emnlp-main.605.

[51]   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, July 2001, pp. 148–155. URL: https://hal.inria.fr/inria-00100529.

[52]   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). http://www.springerlink.com/content/1572-9583/, pp. 421–438. DOI: 10.1007/s10849-004-2114-x. URL: https://hal.inria.fr/inria-00112956.

[53]   B. Guillaume and G. Perrier. 'Interaction Grammars'. In: *Research on Language and Computation* 7.2-4 (2009), pp. 171–208. DOI: 10.1007/s11168-010-9066-x. URL: https://hal.inria.fr/inria-00568888.

[54]   I. Mel'čuk. *Semantics: From Meaning to Text*. Vol. 1. Studies in Language Companion Series 129. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2012.

[55]   B. H. Partee. 'Noun Phrase Interpretation and Type-shifting Principles'. In: *Studies in discourse representation theory and the theory of generalized quantifiers*. Ed. by J. Groenendijk, D. de Jongh and M. Stokhof. De Gruyter, 1986, pp. 115–143.

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

[57]   Ö. Sevgili, A. Shelmanov, M. Arkhipov, A. Panchenko and C. Biemann. 'Neural entity linking: A survey of models based on deep learning'. In: *Semantic Web* 13.3 (2022), pp. 527–570. DOI: 10.3233/SW-222986.

[58]   K. Zaporojets, J. Deleu, C. Develder and T. Demeester. 'DWIE: An entity-centric dataset for multi-task document-level information extraction'. In: *Information Processing & Management* 58.4 (2021), p. 102563. DOI: 10.1016/j.ipm.2021.102563.