AAPG2021	SELEXINI - Semantic Lexicon Induction for Interpretability and diversity	
Coordinated by :	Marie CANDITO and Carlos RAMISCH	48 months

SELEXINI: SEmantic LEXicon INduction for Interpretability and diversity in text processing

Abstract

Despite great enthusiasm for deep learning in NLP, concern is rising about its limitations. First, neural models are often blackboxes, and their behavior is hard to interpret. Second, benchmark-based evaluation overlooks biases, questioning the robustness and coverage of the resulting generalisations, yielding a landscape of overall diversity. The goal of the SELEXINI project is to address these issues by developing ***weakly supervised methods to induce semantic lexicons*** *from raw corpora, which will then be* ***seamlessly integrated with semantic text processing models****. Lexical units are seen as useful abstractions that allow representing complex phenomena (e.g. polysemy, similarity, multiword units) associated with interpretable labels, avoiding the overhead and opaqueness of contextualized embeddings (one vector per occurrence). Moreover, our lexicon will combine continuous data (embeddings, clusters) and symbolic data (labels). We will model single and multiword units, their senses, and their semantic frames (arguments, roles). Hence, we propose a new "by-construction" view on interpretability, which can be seen as an alternative to methods trying to dissect complex neural models. For extrinsic evaluation of interpretability and diversity, the induced lexicon will be integrated into standard deep learning models in downstream tasks requiring semantic information: machine reading comprehension and multiword expressions identification. We will develop an experimental protocol to assess the lexicon-corpus complementarity on diverse linguistic phenomena, and to assess the lexicon's usefulness for non-expert end users requiring interpretable results. We expect that this original approach will increase both the interpretability of models and the coverage of diverse phenomena (e.g. rare/unseen items in training data).*

I.Proposal's context, positioning and objective(s)

a. Objectives and research hypothesis

Contextualization: We are living in a time of **great enthusiasm for AI**, in particular for deep machine learning which has revolutionized several domains with impressive performances. In the last 5 years, the field of **natural language processing (NLP)** has been propelled from the position of a multidisciplinary research field to become a **major showcase for deep learning technology**. Hence, NLP is becoming **increasingly influential in our society**, thanks to versatile and novel applications.

Mainstream NLP systems are based on **neural models** such as LSTMs and transformers, performing end-to-end predictions which often bypass the need for linguistic expertise. They have allowed **major breakthroughs** and **fast progress** in tasks requiring high-level **semantic interpretation** such as natural language understanding. One of their strengths is the use of real-numbered vectors or **embeddings** to represent inputs and outputs in **continuous** rather than symbolic spaces. Embeddings can be pre-trained on large raw corpora and fine-tuned for specific tasks in **semi-supervised settings,** addressing challenging issues such as out-of-vocabulary units and limited annotations. Moreover, contextualized and sublexical embeddings are very popular, representing tokens' contexts beyond sentences (Devlin et al. 2019).

Scientific barriers to be lifted: As enthusiasm for deep learning grows, fostering fast development of more complex powerful architectures, the prevalence of neural (semi-)supervised learning raises concern about interpretability and robustness. Some of the most important challenges faced by NLP today are:

1. **Model opacity:** neural nets are composed of millions of real values corresponding to parameters of complex functions transforming input vectors into probability distributions over outputs. Thus, these models have become extremely **hard to inspect and interpret** without the use of sophisticated techniques (Rudin 2019). Despite the recent Bertology research (Rogers et al. 2020), it remains extremely **hard to explain these models' predictions,** why they fail, and how to improve them.

2. **Lack of diversity**: Most linguistic phenomena follow Zipf's law, i.e. few items are frequent and there is a long tail of rare ones. These few frequent items tend to be less diverse than the numerous items in the "Zipfian tail". Current models often favour the former and underperform in the latter, as they heavily rely on annotated data and are tuned for optimal global performances on **biased benchmarks**. Hence, quality is overestimated while **generalisation and robustness are rarely assessed** (Wisniewski & Yvon 2019).[1] Although awareness about diversity is rising (Narayan & Cohen 2015; Yang et al. 2018; Palumbo et al. 2020), **diversity is still largely neglected** to build and evaluate NLP systems.
3. **Simplifications of word-based models:** words are the basic units in many embedding models, although **multiword expressions** (MWEs), i.e. idiosyncratic word combinations, abound in languages (Constant et al. 2017). As most MWEs cannot be dealt with compositionally (Cordeiro et al. 2019), MWEs are often at the root of errors in word-based models. Moreover, assuming words as minimal semantic units also yields **meaning conflation**, mixing different meanings into one sub-specified vector (e.g. *crane* - animal or tool?).[2] Finally, word-based embeddings **lack (explicit) structure**, e.g. verbs' argumental structure.

Research hypothesis: For us, the notion of **lexicon** is central to attain interpretable and robust semantic processing (Savary et al. 2019). In a **semantic** lexicon, linguistic objects such as senses and frames act as trade-off representations between static word embeddings, which tend to conflate the different meanings of a word, and the opposite extreme of contextual embeddings obtained via pre-trained language models, in which each occurrence has a distinct representation. We hypothesise that the scientific barriers listed above can be addressed by developing methods to **induce semantic lexicons** from distributional patterns in raw corpora. These lexicons will combine continuous representations (embeddings) and interpretable descriptions (labels, frames, definitions, links to external lexicons...). Thus, we hypothesise that they will be usable both within robust NLP models for downstream tasks *and* in tasks requiring human interpretation.

Objectives: our goal is **to develop weakly supervised methods to induce semantic lexicons** which will then be **seamlessly integrated with neural text processing models**. Induction is understood here as automatic lexicon construction by learning from distributional and structural regularities in non-annotated large corpora. The induced lexicon will contain explicit labels, making it **more interpretable** than (contextualized) embeddings alone. Its entries will be generic enough to allow its reuse in several tasks. It will cover single and MWE entries, encoding their syntactic and semantic idiosyncrasies, building on the consortium's expertise in MWE processing. It will cover **diverse phenomena**, being complementary to annotated corpora, while providing sense frequency information. We will design evaluation protocols to drive our approach towards **cutting-edge robustness**, as compared to supervised methods alone. The lexicon will be evaluated extrinsically in downstream tasks, focusing on **interpretability and diversity**. Although we focus on French, the proposed methods should be applicable to any language (see Sec. III).

b. The SELEXINI semantic lexicon model

We start with the **SELEXINI terminological and conceptual framework**, which is the base of our methodology (Sec. I.c). Although it does not constitute a final contribution *per se* (WP1 addresses the lexicon's format), this minimal design constitutes the starting point for work packages 2-5, with frequent updates as the project goes along. We detail it here to facilitate the reading of the next sections.

Our framework is illustrated in Figure 1: two abstract events (frames), *stealing* and *flying*, are realised differently in sentences describing each situation and its participants. When analysing the model bottom-up, we start with a set of corpus sentences, containing **occurrences**, i.e. inflected forms of single words or morphosyntactic variants of MWEs, shown in bold. Yellow ellipses on the left illustrate the notion of **lemmas**, i.e. normalized canonical/base forms corresponding to how words appear in a dictionary.

[1] References having at least one author from the SELEXINI proposal consortium are shown in green.

[2] Examples in English for better readability, although SELEXINI prioritises language-independent methods.

Figure 1: SELEXINI's framework illustrating how lemmas and induced frames relate to occurrences.

Occurrences may have semantic arguments, in brackets and colored background in Figure 1. Although this is a polemic notion, we adopt the PARSEME definition of **semantic arguments** as participants of a situation which are both mandatory to fully specify it and specific to that situation (Savary et al., 2017). In example 1, *un téléphone (a telephone)* is mandatory to fully specify what is robbed (Y), whereas *cette nuit (this night)* is not specific to stealing events (example 2). Notice that these examples do not overtly specify the entity being robbed (Z), also mandatory for the robbery to exist. Units having at least one semantic argument are called **predicative**. Figure 1 omits non-predicative units such as *locomotive* and *palm tree*.

Different occurrences of the same lemmas can be grouped into lexical **senses** representing specific meanings, indicated by dashed boxes. In particular, occurrences of the polysemous lemma *voler (steal/fly)* belong to 2 senses. Lexical senses can have associated information, e.g. textual definitions, supersense labels, links to external resources (omitted). Senses can be interlinked to capture semantic similarity (synonymy or coarser semantic relatedness). We call these groups **semantic frames**, extending the traditional view of frames to include non-predicative units, similar to Wordnet's synsets. Frames may comprise **semantic slots** (similar to FrameNet's core frame elements), i.e. pieces of information that must be instantiated (by semantic arguments) for a lemma to evoke the frame. Part of these slots are filled by linguistic expressions in the corpus, and part have to be inferred (from linguistic or pragmatic context). The larger gray boxes show two frames and their slots, with indices and colors pointing to semantic arguments.

Figure 1 covers the most interpretable part, i.e. the structure and labels in the lexicon. Since the lexicon is induced from corpora, by construction, each of the elements (occurrences, lemmas, senses, etc.) is linked to dense representations, easy to integrate in NLP models. For instance, contextual embeddings of lemmas clustered into the same sense can be given as features for a natural language understanding system (WP4).

c. *Position of the project as it relates to the state of the art*

We briefly summarize three views on semantics for NLP: handcrafted lexicons, supervised learning for semantic text analysis, and weakly supervised lexicon induction, before we position our project.

Format, coverage, and discreteness issues in handcrafted semantic lexicons

The most popular semantic lexicon in NLP is the English Wordnet, which groups senses of nouns, verbs, adjectives and adverbs into *synsets* (sets of quasi-synonyms), linked by semantic relations (e.g. hypernymy, meronymy). MWEs are included in Wordnet, but represented as flat strings with no internal structure.

The relation between lexical units and semantic arguments is addressed by structured lexicons such as FrameNet, VerbNet, and PropBank, including efforts to connect them as SemLink (Bonial et al. 2013). They

differ mainly in the granularity of entries and slot labels (e.g. in a *Theft* event, FrameNet's frame-specific labels are *Perpetrator* and *Victim*, VerbNet's coarser generic *thematic roles* are *Agent* and *Source*, and PropBank's labels are *arg0, arg1*). FrameNet and PropBank are associated with annotated corpora. Also, abstract meaning representation graphs are built using PropBank senses (Banarescu et al. 2013).

Numerous projects aimed at (semi-)automatically creating similar resources for French. We can cite the WOLF Wordnet (Sagot & Fišer, 2008), constructed using parallel corpora, and a French version of VerbNet, obtained via semi-automatic alignment with the English version (Danlos et al. 2016). Manually annotated FrameNet corpora for French include ASFALDA (Djemaa et al. 2016), and CALOR (Marzinotto et al. 2018), covering 105 and 50 frames (i.e. about 1/10 of the English FrameNet). Finally, there is the *Réseau Lexical du Français* (RL-fr), based on meaning-text theory, which notably models MWEs (Polguère 2014).

Initially built by mapping English Wordnet to Wikipedia, Babelnet (Navigli & Ponzetto, 2012) became a highly multilingual lexical network, with version 5.0 now covering 500 languages (including French). Wiktionary is a collaborative multilingual lexicon built by and for humans, but also useful for NLP (Sérasset 2012). It covers 182 languages and has relatively high coverage and good quality for single words. Each lemma has a list of lexical senses and their definitions, often followed by example sentences.

Making MWE-aware lexicons is hard, as MWEs are partly regular and partly idiosyncratic. This calls for representing them jointly with single words, e.g. as predicates with some lexicalized arguments, not mapped to semantic roles (McShane et al., 2015). However, MWEs exhibit wide morphosyntactic variability and semantic non-compositionality, calling for complex yet large-scale descriptions.

Pros and cons: Handcrafted lexical resources abound for English and exist for some other languages. They allow fine-grained encoding of complex phenomena (e.g. RL-fr, FrameNet), favouring linguistic precision. Still, their granularity is fixed and often considered too fine for semantic NLP tasks (Lacerra et al. 2020). Moreover, the main limitation of handcrafted lexicons is the **huge effort required to reach decent coverage**. FrameNet has been an ongoing project for more than 20 years, with still insufficient coverage for English, and even more so for French. The RL-fr has described about 28K lexical units in more than 10 years.

For MWEs, coverage is even weaker. We found that 55.2% of the MWEs annotated in the PARSEME-FR corpus (Candito et al., 2020) are absent from Wiktionary (this ratio is only 7.6% for content words). Furthermore, the syntactic structure and variability of MWEs are mostly neglected or modelled informally in lexicons with decent coverage (Lichte et al., 2018). Another limitation of hand-crafted lexicons stands in **discreteness**, whereas major recent NLP breakthroughs lie in using continuous representations. The fact that most languages **lack corpora** annotated with these lexicons makes it much harder to build embeddings for lexical entries. Lack of annotation is also a well known bottleneck for large-coverage WSD (see below).

Lack of gold annotated data for WSD and semantic parsing

Lexicons are particularly useful not only per se, but as nodes of semantic representations, i.e. for semantic parsing. This supposes to map in-context word forms to lexical entries: a task known as **word sense disambiguation (WSD)**. Current WSD approaches are either (semi-)supervised or knowledge-based, the former seemingly outperforming the latter (Raganato et al. 2017). For instance, Vial et al. (2019) reach high performance for English by injecting knowledge-based information into a supervised neural model.

Pros and cons: English has rich resources including both a lexical resource (Wordnet) and a Wordnet-annotated corpus (SemCor), but gold **sense-annotated data is rare or nonexistent for most languages**. For French, we are aware of a small noun sense disambiguation dataset (Navigli et al., 2013), the French SemEval dataset (Segonne et al. 2019) with annotations for 50 verbs, and the FrSemCor corpus tagged for supersenses of nouns and nominal MWEs (Barque et al. 2020). Knowledge-based methods (i.e. using handcrafted dictionary definitions and examples) are an alternative, but dictionary examples remain insufficient for supervised methods. For instance, Segonne et al. (2019) obtained an F-score of 0.43 for WSD for French verbs when training on Wiktionary examples. Also, dictionary examples do not model sense frequency, crucial for a task in which the "most frequent sense" baseline is strong.

Self-training is another method to augment data. The EuroSense corpus (Delli Bovi et al. 2017) is a multilingual corpus which was automatically annotated for BabelNet senses. Although it led to small WSD improvements when used to augment training data, its quality is too low to serve as training data per se (Segonne et al. 2019), and the BabelNet inventory proved highly redundant. Concerning data for semantic

analysis beyond WSD, it is well known that e.g. frame-annotated data, or more sophisticated semantic representations such as AMR (Banarescu et al. 2013) are even more difficult to obtain at a large scale.

Moreover, because high-quality sense-annotated data is so rare, annotation of MWEs within sense-annotated data is often minimal and/or inconsistent. MWEs can be (partly) non-compositional, so some or all the components do not contribute their usual senses to the sense of the whole,e.g. *pay* in *to pay a visit*. Being frequent, MWEs should constitute a major challenge in WSD. Still, work specifically addressing the impact of MWEs on WSD is relatively scarce (Del Corro et al. 2014; Schneider et al. 2016).

Word sense and semantic frame induction: high coverage, but low interpretability

Since full-coverage WSD remains unsolved for most languages, an alternative is **word sense induction** (**WSI**), in which the different senses of words are automatically induced from large unlabeled corpora. The two main classes of WSI techniques consist in clustering (partitioning) occurrences of a given lemma, versus partitioning "word ego-networks", i.e. graphs of lemmas semantically related to the target ambiguous lemma. In the former, we mention latent variable models (e.g. Amplayo et al. 2018), who use a language model to generate candidate in-context substitutes for word instances, and then cluster the vectors of substitutes to both induce senses and associate instances to senses. Amrani et al. (2019) show that, using substitutes from BERT yields substantial performance improvements. An alternative class of WSI methods uses graph-clustering algorithms over word ego-networks: a lemma is linked to semantic neighbours taken from a dictionary, as in Ustalov et al. (2019), from distributional models, or from in-context neighbors.

Closely related to word sense induction, the task of **semantic frame induction (SFI)** consists in identifying more structured units than word senses, namely **semantic frames**. These are viewed here as typed feature structures consisting of an event or state type (e.g. commercial transaction), its labeled semantic argument slots (e.g. *Buyer, Seller, Goods*), and the lexical units that can evoke the frame (e.g. *sell.v, purchase.n)*. Compared to word senses, frames cope with **(i) lexical variation,** as senses with similar meaning can be grouped in the same frame, and **(ii) semantic valency properties**, as semantic frames list slots for the participants to the event or state, possibly along with their syntactic properties. These slots are labeled with semantic roles of various granularity (cf. above, hand-craft lexicons). Semantic frame parsing, that is, identifying the frames evoked in a text, and the elements filling their semantic slots, provides a schematic representation of the eventualities evoked in a text.

Previous work on semantic frame induction, especially SemEval 2019 shared task on lexical frame induction (QasemiZadeh et al. 2019), considers a pipeline of two tasks: (i) predicates are grouped into semantic classes and then (ii) their arguments are grouped into semantic roles. Other methods directly induce full frames. Central to these methods is the use of syntactic trees, from which predicate instances (in general verbal predicates) are extracted along with their dependents. For instance, Kallmeyer et al. (2018) treat semantic frame labels and semantic role labels as latent variables in a generative probabilistic model. Ustalov et al. (2018) approach the task through clustering of subject-verb-object triples.

Pros and cons: WSI and SFI tackle the coverage issues of pre-defined sense and frame inventories, in which senses/frames may be missing, may not be relevant to a domain, or may not have the appropriate granularity. The coverage advantage of induction methods is crucial in the case of semantic frames, as decades of FrameNet-based projects have produced insufficient data to cope with applied semantic tasks (e.g. machine reading comprehension, natural language inference). Another advantage of WSI over WSD is that **WSI methods generally provide a large sense-tagged corpus as a by-product** of sense induction. Moreover, with respect to handcrafted lexicons, the induced senses/frames do embed frequency information, and for most methods, **they are associated with distributed representations** (embeddings) by construction. Finally, compared to supervised WSD, the sense annotations obtained through WSI have necessarily better coverage than those present in manually annotated data (although limited to the corpus used for sense induction), and sense and frame induction constitute automatic and reusable procedures.

However, these advantages come at the cost of limitations. As opposed to handcrafted dictionaries, the induced senses are **much noisier** than handcraft lexicons, and **generally lack interpretability**, though some methods can provide interpretable induced senses representations in the form of lists of representative semantic neighbors and/or sparse but interpretable features capturing the typical context of the senses (Panchenko et al., 2018). Another limitation is that these methods generally **do not take a full account of**

MWEs, either working at the level of single-word lemmas, or focusing on the induction task, assuming gold MWEs as in the frame induction SemEval 2019 shared task (QasemiZadeh et al. 2019).

SELEXINI's position

Our proposal is positioned in the word sense and semantic frame **induction** paradigm. We will develop automatic and reusable methods to induce a semantic lexicon for French. We believe that this induced lexicon provides an original trade-off between standard word embeddings (all meanings of words are conflated) and contextualized BERT-like embeddings (all word occurrences have different representations).

Originality of the objectives: SELEXINI has original and ambitious goals with respect to the landscape above. First, we will design a **sound lexicon model**, inspired by the sophistication reached in handcrafted lexicons, **taking into account specificities of MWEs**. In contrast to standard embeddings, we will induce **structured semantic units** including syntactic and semantic valency (i.e. semantic frames and their slots, WP2). As a by-product, this procedure will generate an **automatically sense-annotated corpus**, which can bootstrap large-coverage WSD for French. One weakness of WSI, also touching contextualized word embeddings, is the low interpretability of results. SELEXINI includes the **generation of interpretable textual descriptions** for the induced units (WP3). The resulting hybrid lexicon will **link dense embeddings to symbolic descriptions**, thus proposing a trade-off between practical usefulness and explicit labels. Its evaluation will be based on applicability, putting special emphasis on its **integration within downstream applications**, the **interpretability of results** (WP4), and the **diversity of the phenomena** covered (WP5).

Originality of the methodology: The experience of the partners in PARSEME-FR leads us to **taking MWEs into account** from the very beginning, and not as extensions to word-based models (WP1). The application of **constrained clustering to WSI** is a promising original idea (WP2). We will use existing lexicons, namely Wiktionary, to learn **more structured** units than techniques such as retrofitting or AutoExtend (Rothe & Schütze, 2015), while retaining **control on the induced sense inventory structure and granularity**. We prefer Wiktionary over other resources because Wiktionary's senses proved more suitable than Babelnet for French verbs (Segonne et al. 2019). Moreover, Wiktionary is quite large for many languages (24 languages with more than 50,000 entries). Thus, although focusing on French, the project will yield methods applicable to other languages. Finally, we will design a new generic framework to evaluate interpretability (WP4) and diversity (WP5) in downstream tasks, thanks to the induced lexicon.

d. Methodology and risk management

SELEXINI has 5 scientific work packages (WPs) in addition to project coordination (WP0). **WP1** groups technical tasks: corpus preprocessing, embedding pre-training, pre-lexicon extraction. **WP2**, at the core of the project, covers the induction of senses and frames via clustering. **WP3** adds human-readable labels on top of the induced frames, creating new MWE entries when necessary. In WPs 4 and 5, we assume that the induced lexicon increases interpretability/diversity of downstream tasks. For **WP4**, we hope that the lexicon can help interpret the capabilities of a machine reading comprehension system. In **WP5**, we design a new evaluation framework to measure the diversity of a corpus and apply it to MWE identification.

WP0: Coordination and dissemination

Partner in charge: LLF (M. Candito) and LIS (C. Ramisch)
Objectives:

- Ensure WP synchronization, so that ideas, resources and methods are shared among partners
- Release resources under open licences, publish in top-tier open-access conferences and journals

➢ WP0.1: Communication and project management tools

Scientific coordinators will meet online every 2 months for WP progress updates, with minutes on the wiki of a project management tool also containing internal documentation (e.g. GitLab). For closer collaboration on tasks that are critical to the project's success (namely WP1, WP2), we will adopt Agile tools (e.g. Trello) and methodology. The goal is to frequently release intermediate versions of the induced lexicon, foster inter-WP collaboration and minimise risks. Yearly (online) workshops will allow members to interact, attend invited talks and hands-on sessions. Announcements will be posted on the project's mailing list.
Deliverables: D0.1.1 internal wiki; **D0.1.2** mailing lists; **D0.1.3** Agile management and version control tools

➢ WP0.2: Shared computational infrastructure

Development and experiments require access to computational resources such as graphical processing units (GPUs) and shared storage for corpora and models. This WP ensures seamless integration of SELEXINI's computational resources in the computational infrastructure of LIS' cluster and CNRS's Jean-Zay cluster. Access will be ensured for all partners, including members external to LIS. We will create, adapt and maintain documentation in the form of wiki pages describing how to install and use shared software (e.g. PyTorch, SKLearn) and how to launch and manage jobs on the cluster (e.g. using slurm). The project's GPU usage will follow a two-tier strategy: models will be developed within the LIS cluster with dedicated project GPUs, and larger-scale experiments will run on the Jean-Zay cluster which requires more mature code.
Deliverables: D0.2.1 shared computing infrastructure; **D0.2.2** documentation for use of all partners

➢ WP0.3: Data and software releases

Software developed for WPs 2-5 will be maintained using version control (e.g. GitLab), released under open licences (e.g. GNU GPL) and publicized using dedicated mailing lists (e.g. corpora) and demonstration papers at conferences. Data, especially the induced semantic lexicon in a standard interchange format and corresponding (automatically) disambiguated corpus and pre-trained language models, will be made available in a linguistic data repository (e.g. Ortolang, CLARIN) under open licenses (e.g. CC-BY).
Deliverables: D0.3.1 public documented repositories for software of each WP; **D0.3.2** yearly releases of the induced semantic lexicon on a linguistic data repository using an interoperable format.

➢ WP0.4: Dissemination and publications

We will communicate to the scientific community and general public on a website, using a collaborative CMS (e.g. wiki). Outcomes will be published in national (e.g. TALN) and international conferences (e.g. *ACL, EMNLP, COLING) and journals (e.g. CoLi, TACL), favouring open access.
Deliverables: D0.4.1 public website; **D0.4.2** 2 papers/year in conferences; **D0.4.3** 2 impact journal articles

WP0: Risks and fallback solutions

No major risks are foreseen. Partners have already collaborated in the past (Sec. II.a), are familiar with most communication tools, data and software release platforms, and their record of publications attests successful past collaboration. If technical restrictions hinder access to computational resources, we can use local infrastructures with copies of the resources, or turn to national clusters such as CNRS's Jean Zay.

WP1: Corpus preparation, lexicon design, corpus-lexicon interface

Partner in charge: LIS (C. Ramisch)
Involved partners: LIFAT (A. Savary, J.-Y. Antoine), ATILF (M. Constant, A. Polguère), LLF (M. Candito)
Overview: WPs 2-5 require large amounts of pre-processed textual data. Hence, WP1 groups most technical and engineering subtasks to gather a large accessible base of French text, preprocess it with state-of-the-art tools, and extract the pre-lexicon. The **input** consists of raw corpora, in-house and off-the-shelf tools. The **output** is a pre-lexicon of automatically extracted single- (e.g. *steal*, *fly*) and multiword lemmas (*commit theft*, *take off*) not yet disambiguated for senses (WP2), associated with corpus occurrences, Wiktionary entries, morpho-syntactic structures and pre-trained contextual embeddings.
Objectives:

- Collect, preprocess, document and index a large French corpus
- Define a lexicon format combining continuous and human-readable representations
- Extract a pre-lexicon of lemmas (not grouped into senses) aligned to Wiktionary and embeddings

➢ WP1.1: Corpus collection, documentation and preprocessing

As done by Le et al. (2020), we will gather copyright-free corpora from Wikipedia, CommonCrawl, news, and Web as Corpus (Wacky), aiming at 10 billion tokens. We will document potential biases (Bender & Friedman 2018) and ensure representativity of evaluation domains (WP4 and WP5). The corpus will be cleaned and formatted using off-the-shelf tools (e.g. jusText, Onion). The partners' expertise enables quick enrichment of the corpus with POS tags, lemmas, dependency syntax (Scholivet et al., 2019), "deep" syntax (Candito et al., 2017), named entities and MWEs (Al Saied et al., 2018) using in-house supervised models learnt from annotated corpora. These rich and diverse annotations constitute clues for the extraction of the

pre-lexicon (WP1.3) and for clustering (WP2). For instance, deep syntax can help account for syntactic variation in a verb's arguments (e.g. in both *many children* ***watched*** *the show* and *the show that seems to have been* ***watched*** *by many children*, the verb has the same deep syntactic subject (*many children*) and same deep object (*the show*). The corpus will be available to all partners through an API (e.g. NoSketch).
Deliverables: D1.1.1 10-billion-tokens preprocessed French corpus released under an open license; **D1.1.2** Wiki pages to document preprocessing (links to data sources, tools, models, tagsets, and their versions).

➢ WP1.2: Creation of a model for contextual representations

Lexicon entries such as senses, frames and arguments will be induced from vectors representing lemmas' occurrences. In this WP, we will prepare a model to extract continuous contextual representations for corpus occurrences in a manner that will best suit the needs of WP2. Therefore, we will adapt a current mainstream language modelling architecture based on transformers, such as BERT. Either we will fine-tune an existing pre-trained model for French such as FlauBERT (Le et al. 2020) on a masked language modelling task on our corpus, or we will train it from scratch using a distilled model. The first challenge here is to deal with MWEs and wordpiece tokens,[3] both of which challenge the assumption of 1 word = 1 vector. The second challenge is to ensure versioning and synchronisation between the induced semantic lexicon and the language model supporting its contextual representations, especially upon public release of the lexicon.
Deliverables: D1.3.1 a pre-trained model yielding contextual embeddings for corpus occurrences.

➢ WP1.3: Extraction and mapping of the pre-lexicon

This subtask aims at (1) extracting from the preprocessed corpus a pre-lexicon of single-word and MWE lemmas, (2) preprocessing existing lexical resources, mainly Wiktionary, building on previous work on (e.g. Sérasset 2012), and (3) mapping the pre-lexicon to Wiktionary. The *pre-lexicon* is defined as a list of single- and multiword lemmas not disambiguated for word senses nor grouped into frames, and associated with their corpus occurrences and Wiktionary entries. Example sentences from Wiktionary will be added to the corpus as gold sense-disambiguated sentences (Segonne et al., 2019), used as seeds for clustering (WP2). Lemmas of *single words* are straightforward to extract and map to Wiktionary given that, at this point, we do not disambiguate senses, but simply link abstract lemmas to occurrences, addressing orthographic variability in noisy pre-processed text. MWEs, however, are trickier to process since they require neutralising morpho-syntactic variability (e.g., discontinuities and inversions). We will rely on the output of automatic MWE identification (WP1.1) and extend it with Wiktionary entries, counting on the team's expertise on modelling MWE variability (Pasquer et al. 2018). The main challenge for MWEs lies in dealing with the variability of corpus-based lemmas and Wiktionary entries. For instance, the Wiktionary entry *aux quatre vents* (lit. *to-the four winds* 'in all directions') may be lemmatized in the corpus as *à le quatre vent* in the corpus, with *aux* 'to-the' separated into two words and the noun in singular. Moreover, the lexicalized elements (Savary et al. 2017) may not match in the corpus and in Wiktionary. Abstract lemmas of MWEs not present in Wiktionary will simply consist of sequences of single-word lemmas in an arbitrary order. In the whole process, frequency will help dealing with noise introduced by automatic tools in WP1.1.
Deliverables: D1.3.1 pre-lexicon of word and MWE lemmas linked to Wiktionary and corpus occurrences.

➢ WP1.4: Design of the final semantic lexicon format

While the pre-lexicon consists in a flat list of lemmas, the final *lexicon* is more complex. A minimal version of the lexicon format includes lemmas, (predicted) morphosyntactic features of single words, and coarse syntactic structures of MWEs (Savary et al. 2019). For both types of units, lexicon entries are linked to their corpus and Wiktionary occurrences. Here, we extend this minimal format as sketched in Figure 1 (Sec. I.b) to account for induced properties of lexicon entries (WP2). Notably, lemmas should be assigned to induced senses and semantic frames, including the semantic arguments linked to predicative units. Three challenges are to: (a) represent MWEs in the same framework as single, predicative and non-predicative entries, (b) allow for some flexibility in the granularity of induced senses and frames, and (c) link continuous (induced) representations with their interpretatable counterparts, e.g. embeddings (WP1.2) with textual definitions (WP3.1), induced frames in the form of clusters (WP2.2) with human-readable descriptions (WP3.2).
Deliverables: D1.2.1 lexicon format specification; **D1.2.2** release procedure for alpha versions of the lexicon

[3] BERT's tokenizer decomposes (rare) words into sub-lexical units for which representations are pre-trained.

➢ WP1: Risks and fallback solutions

This WP is mostly technical and does not involve major scientific risks. Its most exploratory aspect concerns mapping MWEs in the corpus to Wiktionary in WP1.3, since their forms in the lexicon may differ from their occurrences. As a fallback solution, we consider focusing on more fixed MWEs, leaving syntactically flexible ones as future work. Moreover, if we fail to retrieve variable occurrences, we can focus on frequent MWEs and their "standard" occurrences. Another technical challenge in WP1.4 is to deal with the size and stability of the reference data for induced frames with a flexible degree of granularity, especially if lexicon induction is seen as an iterative process. Clustered occurrences and their contextual embeddings will need to refer to stable corpus, sentence and/or token identifiers. To ensure this, strong interaction will occur between WP1 and WP2, with a rigorous versioning and release policy.

WP2: Weakly supervised induction of lexical senses and semantic frames

Partners in charge: LLF (M. Candito)

Involved partners: LLF (L. Barque, G. Wisniewski), LIS (C. Ramisch, B. Favre, A. Nasr)

Overview: In this WP, we will develop methods to disambiguate and structure the entries of the pre-lexicon. The **input** is the pre-lexicon (lemmas associated with corpus occurrences, some lemmas being associated with wiktionary senses and examples). The **outputs** will consist of **senses**, seen as clusters of occurrences of a given word or MWE lemma, with mapping to Wiktionary senses when possible/appropriate. Senses of predicates (mainly verbs) will be further grouped into coarser **semantic frames**, with their semantic arguments grouped into slots serving as unlabeled **semantic roles**.

For instance, we target to cluster occurrences such as "*the diamonds stolen that night*", "*you should not steal from other people*", "*they may have purloined data*" into a single frame (for a stealing event), with buyers grouped together in one slot ("*you*", "*they*"), another slot grouping the stolen goods ("*the diamonds*" and "*data*"), and a slot for the entity who is stripped of his/her property ("*other people*").

Objectives:

- Develop induction methods via occurrence clustering using Wiktionary entries and examples as seeds
- Include from the beginning both single-word and MWE occurrences to cluster
- Induce a hierarchical lexicon with two levels of granularity: senses and semantic frames, the semantic frames being defined as sets of senses, augmented with sets of semantic roles
- As a by-product, build a pseudo-gold corpus annotated with induced senses, frames and roles.

➢ WP2.1: Supersense tagging and semantic argument identification

WP2 will identify the semantic arguments of predicative lemmas, including most verbs (e.g. *[somebody]* ***gives*** *[something] to [somebody]*), but other parts-of-speech too (e.g. ***destruction*** *of [something] by [something/somebody]*). This subtask aims at (1) **distinguishing predicative from non-predicative nouns**, both for single words and MWEs, and (2) identifying, **in each predicative lemma occurrence, the realized semantic arguments**, using deep grammatical functions as an approximation (cf. deep parsing in WP1.1).

Using deep syntax allows us to partially abstract away from syntactic variation such as active/passive alternation. For instance, in **(1)** *younger children who still do not fully* ***grasp*** *the concept of death*, the occurrence of *grasp* gets a deep syntactic subject and a deep syntactic object (underlined), while in **(2)** *the causes of the fire are not fully* ***grasped*** *yet*, the underlined portion is identified as deep object. For verbal MWEs, situations in which syntactic arguments are part of MWEs must be identified (e.g. *the tables* in *farmers can* ***turn the tables on*** *desertification*), as these should not count as open semantic arguments.

Moreover, we will build on the recently released FrSemCor corpus in which nouns are annotated for coarse semantic types or *supersenses* (Barque et al., 2020), using semi-supervised learning to predict the supersenses for all lemmas in the pre-lexicon (Aloui et al., 2020). Supersenses will help identify predicative nouns and type the nominal arguments of predicates. For instance, in (1) and (2) above, the deep subject (*children*) has the Person supersense, while deep objects (*concept*, *causes*) have the Cognition supersense.

Deliverables: D2.1.1 semi-supervised method for supersense tagging; **D2.1.2** semi-supervised method to identify the semantic arguments of predicative lemma occurrences and their deep grammatical functions.

➢ WP2.2: Induction of senses for non-predicative lemmas via constrained clustering with Wiktionary seeds

We will first focus on non-predicative lemmas output by WP2.1. We aim at **inducing lexical senses** by hard-clustering occurrences of the same lemma, using constraints from Wiktionary senses. The basic system can be based on constrained k-means (e.g. Basu et al., 2002). The encoding of lemma occurrences will make use of the pre-trained language model (WP1.2) to provide transformer-based contextual embeddings, as well as, in the case of nouns, a probability distribution over nominal supersenses (WP2.1). Wiktionary information can be injected as must-link and cannot-link constraints (e.g. the different examples of the same Wiktionary sense must remain in a single induced sense). **Evaluation** and model tuning can be performed using the small noun sense disambiguation set included in FLUE (Le et al., 2020). As **exploratory** work, other semi-supervised methods will be investigated. In particular Zhang et al. (2020) propose a generic framework for deep constrained clustering, allowing to incorporate various formal types of constraints (e.g. instance-difficulty constraints) into a deep embedded clustering algorithm (Xie et al., 2016), in which dense representations are learnt through auto-encoding prior to clustering.

Deliverables: D2.2.1 semi-supervised word-sense induction method and induced senses for non-predicative lemmas as hard clusters of lemma occurrences, with partial mapping to Wiktionary senses.

➢ WP2.3: Induction of senses and semantic frames for verbs

We will induce lexical senses and semantic frames by clustering predicative lemma occurrences, starting with (single- and multiword) verbs. Lexical senses can be identified by constraining a cluster to contain only instances of the same target lemma, whereas semantic frames can be induced either by allowing to group instances of different lemmas or by first inducing lexical senses and then grouping them into frames.

The key difference with respect to WP2.2 will be to leverage the identified semantic arguments (WP2.1). In a baseline version, the representation of instances to cluster can be limited to occurrences with a deep subject and/or a deep object, incorporating a contextual embedding of the target predicate plus a representation of its arguments, e.g. as in (Ustalov et al., 2018). Clusters for semantic roles can then be derived from the clusters for frames. A more elaborated version should allow us to cluster occurrences with a varying number of expressed semantic arguments. As **exploratory work,** we will investigate:

- How to achieve clustering into frames and clustering of deep syntactic arguments into roles using a joint learning objective, so that both clusterings can help each other, favoring their compatibility. For instance, frame-clusters should group items whose arguments belong to the same role-cluster;
- How to better articulate the pretrained model (providing the contextual embeddings, WP1.2) with the objective of representing the predicate-argument structure. Specific self-supervised pre-training strategies could be used, such as learning deep verb+object (resp. verb+subject) representations that best predict the deep subject (resp. the object).

Evaluation can be performed by adding the French FrameNet corpora ASFALDA and CALOR (Sec. I.c) to the set of occurrences to cluster, then comparing the induced clusters to those underlying these FrameNets.

Deliverables: D2.3.1 induction method and induced lexical senses and frames for predicative lemmas + annotated corpus. **D2.3.2** sense, frame and role "pseudo-gold" annotations as a by-product of clustering

➢ WP2.4: Extension to predicative nouns

The next step is to incorporate predicative nouns as instances to cluster. Lexicon-based information on verb/nominalization pairs (e.g. *destroy/destruction*) can be used as weak supervision, overcoming the divergences in verb and nominalization instances (e.g. semantic arguments are more likely to be omitted in nominal instances). Since different lexical senses of a verbal lemma can lead to different nominalizations, we will investigate whether including nominalizations can improve sense induction for verbs.

Deliverables D2.4.1 Enhanced versions of the induced senses and semantic frames (and of the pseudo-gold annotated corpus) including predicative nouns.

➢ WP2: Risks and fallback solutions

The induced clusters can be quite noisy. A fallback can consist in focusing on certain frequency ranges (not too frequent, not too rare) and in using confidence measures to filter out some instances, such as instances with very short or underspecified context. In the case of verbs, a fallback can be to focus on certain types of deep syntactic patterns, in particular strict transitive occurrences.

WP3: Generation of human-readable descriptions for induced frames

Partner in charge: ATILF (M. Constant)
Involved partners: LLF (B. Crabbé, M. Candito), LIFAT (A. Savary), LIS (A. Nasr), ATILF (Y. Parmentier)
Overview: This WP will focus on generating human-readable textual descriptions for the induced frames from WP2. It will take as **input** the pre-lexicon generated in WP1.3 and the occurrence clusters representing induced senses and frames linked to Wiktionary senses, provided as the output of WP2. The output textual descriptions will first consist of specific semantic labels for every slot of the induced frames: e.g. the frame *(you + they) (steal + purloin) (diamonds + data) from (people)* would be labeled *Person steal Something from Person*. Moreover, MWEs not pre-identified in WP1 will be explicitly marked as MWEs and their associated clusters will be reorganised accordingly: ex. assuming that the occurrences of the MWE "commit a theft" *are not* pre-identified as such but are clustered with occurrences of *steal/purloin*, it would be identified as MWE, and the noun *theft* would be removed from the slot Object. Last, we plan to generate lexicographic definitions for induced frames. For example, for the induced frame associated with *steal/purloin*, we would produce a definition like *to take illegally, or without the owner's permission, something owned by someone else* (Wiktionary) in order to help human reading of the lexicon. The next WPs will use these enhanced frames in order to bring more interpretable features for the task of machine reading comprehension (WP4) and more knowledge on MWEs for the task of MWE identification (WP5).
Objectives:

- Develop methods to predict specific semantic labels to every slot of induced frames
- Develop MWE discovery methods within the induced clusters and refine the latter accordingly
- Develop methods to generate textual definitions for every induced frame

➢ WP3.1: Generating interpretable labels for induced semantic roles

Generating descriptions for induced semantic frames consists in predicting a semantic class label for every frame slot. Instead of using semantic roles, we will focus on an easier task, namely generating interpretable semantic classes. For instance, *eat* in the sense of ingesting would be defined as *Person eat Food*, instead of using the generic roles of VerbNet (*Agent eat Patient*) or the frame-specific roles of FrameNet (*Ingestor eat Ingestibles*). Based on the lemma clusters associated to every frame slot coming from WP2, our method will, for each slot, (1) extract a subset of relevant lemmas in order to filter noisy lemmas due to the automatic induction, and then (2) infer a semantic class label in the same line as Flati & Navigli (2012). An interesting path of work would be also to use and adapt the method of Panchenko et al. (2017) to find a plausible hypernym for each slot.
Deliverables: D3.1.1 a method to generate textual descriptions for the slots of induced frames.

➢ WP3.2: Discovery of multiword expressions

Some occurrences of multiword expressions will not be identified in WP1, which will lead to the generation of frames with erroneous lexical units and possibly too many slots for predicative units, which can be misleading for human readers. For instance, if the MWE *turn the tables* ('change the situation in one's favor') is not pre-identified, the sentence *farmers can* ***turn the tables on*** *desertification* would be analyzed with a frame including three slots (*farmers*, *the tables* and *on desertification*) instead of two (*farmers* and *on desertification*). This WP's objective is to develop methods to discover non-identified MWEs within our clusters, and refine them accordingly. In the case of frames with too many slots, this can be seen as an extension of WP3.1 predicting a semantic label for each slot, by marking those which are part of a non-identified MWE. For instance, the MWE *perdre la face* (lit. *lose the face*, 'lose face'), if not pre-identified as a MWE in WP1.1, would be represented within a frame corresponding to the lemma *perdre* 'lose'. The frame would have an erroneous extra slot corresponding to the lexicalized component *la face* 'the face', to be classified as a MWE component. It is also possible that components of non-identified MWEs are not parts of a generated frame, like the negation *ne plus* 'not anymore' in the MWE *ne plus en pouvoir* (lit. *not anymore of-it can*, 'be exhausted'). The main idea would be to explore the lexical diversity in the syntactic neighbourhood, like it is usually done in MWE discovery (Constant et al. 2017), using statistical association measures based on word occurrence counts, and semantic similarity based on embeddings (Cordeiro et al. 2019). Another interesting feature is that frame clusters may contain different

lexical units, which might have a positive impact to detect non-identified MWEs within the same cluster due to richer contexts. For instance, occurrences of *aider* (help) and *voler au secours* (lit. *fly to-the help* 'help') would be in the same cluster. While WP3.3 is presented here independently of WP2, in practice we will iterate clustering and MWE discovery to avoid error propagation typical of pipeline-like approaches.
Deliverables: D3.2.1 methods to discover new MWEs within clusters of frames. **D3.2.2** enhanced semantic lexicon with explicated MWEs.

➢ WP3.3: Generating human-readable definitions for induced frames

The aim is to generate dictionary-like textual definitions for induced frames based on information extracted in WP2 and WP3.1: distributional representations, lemma clusters, lexical and semantic labels for slots, and mapping to Wiktionary entries when available. When there is one or more Wiktionary senses mapped to the induced unit, we just pick the associated definition(s). In the case of an unknown mapping, the task is more complex. To address this open scientific question, we will build on Mickus et al. (2019) generating definitions using a transformer-based encoder-decoder neural network that will be enriched with the (partial) semantic structure extracted in WP2 and trained on reliable mapped definitions.
Deliverables: D3.3.1 a dataset to train a definition-generation system; **D3.3.2** methods to generate definitions for non-mapped induced frames.

➢ WP3: Risks and fallback solutions

This WP heavily depends on the quality of induced frames (WP2). To prevent bottlenecks, we will employ Agile, using first-iteration results to start with. This WP covers quite challenging tasks: definition generation and MWE discovery. First, definition generation is a novel task (pioneer work in 2017), and preliminary results show rather poor quality. Note though that this difficulty only exists for new units, not mapped to any Wiktionary entry. As a fallback, we may simply try to generate the genus of the definition (e.g. a hypernym) as in Panchenko et al. (2017). Regarding MWE discovery, in which the team has strong expertise, results may add more noise to the clusters. In this case, we would favour precision over recall. The less risky task seems to be the prediction of descriptors for the frame slots that have been partly addressed by previous work (Flati & Navigli 2012; Panchenko et al. 2017), we could easily back off to these solutions.

WP4: Semantic lexicon at the service of interpretability

Partners in charge: LIS (C. Ramisch)
Involved partners: LLF (M. Candito, B. Crabbé), LIS (B. Favre, F. Béchet)
Overview: This WP assesses the usefulness of the induced lexicon in terms of interpretability. We evaluate its impact extrinsically on the downstream task of machine reading comprehension (MRC). Our goal is (1) to devise new strategies to inject (induced) lexical-semantic knowledge into MRC and (2) to assess whether the lexicon helps improve generalization and explainability. E.g. for question q=*who robbed the diamonds?* and passage p=*the theft of the diamonds was committed by the queen*, a system having access to the induced frame [*steal*, *rob*, *commit theft*] could "explain" why the deep subject of *p* is the right answer.
Objectives:

- Enhance the generalization of the high-level of MRC with the help of the induced lexicon
- Assess the interpretability gain for end users from the semantic lexicon use in this task

➢ WP4.1: Frame parsing for machine reading comprehension (MRC)

MRC consists in predicting an answer span *s* over an input passage *p* which answers a question *q* (e.g. *p=the diamonds were stolen last night*, *q=when did the theft take place?, s=last night*). The task, made popular by the English SQuAD datasets, is usually addressed using a variant of BERT fine-tuned on large amounts of *(p,q,s)* pairs as supervision (Devlin et al. 2019). The goal of this WP is to **integrate the induced semantic lexicon into a supervised MRC system**. The hypothesis is that the structure of the induced frames can help improve the quality, generalization and explainability of MRC, as hinted by Guo et al. (2020) for English FrameNet parsing. Evaluation will be performed on French using the subset of SQuAD translated into French, the CALOR-QUEST corpus (Charlet et al. 2020) and/or the recent FQuAD dataset (d'Hoffschmidt et al. 2020). In this first subpackage, we will investigate a standard pipeline architecture. We will first learn an MWE-aware standard frame parser (e.g. Marzinotto et al. 2019), using as supervision the pseudo-gold corpus obtained as a by-product of WP2. Although noisy, this parser should have a much larger domain

coverage than parsers learned on small domain-specific gold data (e.g. ASFALDA, CALOR). Second, we will explore several possibilities to feed predicted semantic frames as input to the MRC system, e.g. using an attention mechanism over the lemmas related to the predicted frames and their arguments (Guo et al. 2020). This should allow less frequent predicative lemmas to benefit from supervision coming from more frequent ones, clustered in the same frame, improving the quality of the MRC system.
Deliverables: D4.1.1 a supervised frame parser learned on the pseudo-gold annotated corpus; **D4.1.2** a supervised MRC system integrating predicted frame representations

➢ WP4.2: Frames as interpretations for machine reading comprehension outputs

MRC is an ideal task to evaluate our semantic lexicon, as it requires semantic knowledge which can be encoded in the lexicon (e.g. *theft* and *stolen* evoke the same *Stealing* frame, the word *when* refers to the *Time* frame argument). Current BERT-based models often obtain impressive performances on leaderboards, but they lack interpretability. The goal of this WP is to explore how the lexicon induced in WP2 and the human-readable labels/definitions predicted in WP3 can provide an extra interpretation layer to MRC predictions. First, we can check the correlation between the predicted frames and the performance of the system on different question types (Charlet et al. 2020) or identify errors coming from inaccurate representations in the lexicon (Aloui et al. 2020). Second, we can use the attention weights on the frame representations to understand how this information is used in predictions. Third, the semantic lexicon can also be used to detect inconsistent predictions, e.g. returning a span corresponding to a frame argument that never co-occurs with the frames detected in the question. Finally, the semantic lexicon can be useful to identify hard test triples (e.g. multi-hop reasoning, comparison, paraphrases, unanswerable questions) and help interpret to what extent the MRC system is able to deal with them (Yang et al. 2018). The overall aim of this WP is an evaluation framework for MRC, generalisable to other tasks, in which (a) the test data can be tailored to study a given phenomenon and (b) the semantic lexicon can be used to explain whether (and how) the MRC system covers that phenomenon (beyond supporting facts in the passage itself).
Deliverables: D4.2.1 a framework to explain MRC predictions/errors in terms of induced frames

➢ WP4.3: Alternatives for integrating frame parsing and machine reading comprehension

The model developed in WP4.1 is a pipeline in which the output of frame parsing is given as an extra input to MRC. This favours the propagation of errors from one task to another and does not benefit from the fact that both tasks are closely related. Therefore, we would like to develop alternative strategies to integrate the induced lexicon within the MRC model. A first possibility is to use the lexicon for data augmentation, automatically creating new in-domain questions using the induced frames, as performed for CALOR-QUEST based on gold frames (Charlet et al. 2020). A second possibility is to use induced frames to generate adversarial examples (Marzinotto et al. 2019). Both these strategies aim at increasing the robustness of the MRC system with respect to the domain of the MRC training data and of the limited available gold data for frame parsing. As a third strategy, we envisage benefitting from the fact that both tasks, frame parsing and MRC, are closely related and could be performed jointly. The semantic lexicon can be used to continue pre-training BERT on frame parsing before fine-tuning it on MRC. Alternatively, we aim at mitigating the error propagation issue by creating a joint frame parsing and MRC architecture in which the intermediate representations for the first task are given as inputs to the second task, as in multi-task stack propagation (Zhang & Weiss 2016). Development of these models will be guided not only by test-set performance, but also by interpretability and diversity metrics developed in WPs 4.2 and 5.1.
Deliverables: D4.3.1 a data augmentation strategy for MRC based on the semantic lexicon; **D4.3.2** a joint multi-task model for frame parsing and MRC; **D4.3.3** a comparative evaluation of these models wrt. D5.1.2

➢ WP4: Risks and fallback solutions

The main risk here is the availability and quality of the induced lexicon during the development of the MRC models. For availability, we can use small gold FrameNet data as alternatives (ASFALDA, CALOR), while the induced lexicon is under development. For quality, if noisy entries in the lexicon prevent it from being useful, we can focus on frames whose quality is above a threshold wrt. gold data. The "interpretable by design" approach allows us to manually check samples of the induced lexicon to identify such risks early.

WP5: Semantic lexicon at the service of diversity

AAPG2021	SELEXINI - Semantic Lexicon Induction for Interpretability and diversity	
Coordinated by :	Marie CANDITO and Carlos RAMISCH	48 months

Partners in charge: LIFAT (A. Savary)
Involved partners: ATILF (M. Constant), LIFAT (J.-Y. Antoine, C. De Runz, A. Soulet), LIS (C. Ramisch)
Overview: This WP is dedicated to evaluating the usefulness of the semantic lexicon in overcoming one of the targeted scientific barriers (Sec. I.b): lack of diversity. We build on measures of linguistic complexity and extend them towards more universal measures of diversity. We focus on the particular phenomenon of multiword expressions (*commit theft*, *take off*) and on their automatic identification (Savary et al. 2017; Al Saied et al. 2019; Ramisch et al. 2020; Pasquer et al. 2020), which is notoriously sensitive to morpho-syntactic and lexical diversity (*thefts committed*; *commit robbery, make robbery, #make theft*). Diversity of MWEs is measured in corpora and in the predictions of MWE identifiers. Thus, we question the mainstream viewpoint on NLP performance evaluation, in that we shift focus to the diversity of phenomena, mitigating for frequency-driven biases present in benchmarks.
Objectives:

- Quantify linguistic diversity (on the example of the multiword expression phenomenon)
- Define evaluation scenarios which favor diversity in MWE identification
- Assess the contribution of the semantic lexicon to increasing the diversity in MWE identification

➢ WP5.1: Quantifying diversity of multiword expressions in a corpus and in system predictions

Diversity has been quantified in many domains: ecology, economy, information science, etc. In language processing, related properties such as complexity (Brunato et al. 2016) have been more often addressed, especially for the sake of language learning or text simplification. We are, conversely, interested in diversity (a notion larger than complexity) and in its promotion in language resources and tools. Because this is a relatively new problem framing, we will address a particular linguistic phenomenon which we control and understand. Namely, we will focus on multiword expressions (*commit theft*, *take off*), which have a Zipfian distribution (Williams et al. 2015), and which are known to exhibit idiosyncrasies at the level of lexicon, morphology, syntax and semantics (Constant et al. 2017). Our aim is to characterize a corpus (annotated for morpho-syntax and MWEs) for variety, balance and disparity (Morales 2021) of the vocabulary (*commit theft*, *take off*), morphological features (*plural*, *future*) and syntactic structures (*verb-object*, *verb-particle*) occurring in the MWEs contained therein. By extension, diversity will also be measured in the MWE's semantic slots (*Agent*, *Patient*) and semantic frames (*steal*, *fly*). The diversity of contexts in which MWEs occur will also be formalized. Further, we aim at estimating how representative a corpus is of diversity in language, focusing on rare MWEs (which have a major contribution to lexical diversity) and their rare surface realizations (contributing to morpho-syntactic diversity). Representativeness measures will be inspired from the Good-Turing test or Benford's law, previously applied to knowledge bases (Soulet et al. 2018; Yan et al. 2018). Finally, quantifying diversity within a language will be extended to many languages.
These (application-agnostic) measures will then be used to re-design the evaluation framework for a concrete task: MWE identification, in which lexical and morpho-syntactic diversity of the training and test corpora plays a major role. Corpus representativeness measures will lead to methods of diversity-driven corpus split, over-sampling and augmentation. The resulting corpus splits will be used in evaluation scenarios which will promote MWE identification tools performing well on rare and diverse phenomena, across possibly many languages. We will implement these evaluation scenarios within the multilingual PARSEME shared task for MWE identification (Ramisch et al. 2020).
Deliverables: D5.1.1 MWE-oriented corpus diversity measures; **D5.1.2** corpus representativeness measures with respect to rare and diverse MWEs; **D5.1.3** diversity-oriented corpus splitting/sampling/augmentation methods; **D5.1.4** evaluation scenarios for diversity in the outputs of MWE identifiers.

➢ WP5.2: Diversity-oriented extrinsic evaluation of the semantic lexicon

Here, we will apply the evaluation framework from WP5.1 to show how the semantic lexicon (WPs 1-3) contributes, by data augmentation, to an increased account of diversity in MWE identification in French. We do expect the induced lexicon to be more representative of linguistic diversity than both handcrafted lexicons (here: Wiktionary) and manually annotated corpora (here: the PARSEME corpora). Namely, the lexicon will jointly leverage: (i) the richness of handcrafted lexicons, which often contain both frequent and rare items, (ii) the quality of manual corpus annotation, as PARSEME corpora are used for the pre-lexicon construction, (iii) the size of the corpus from which it will be extracted, more representative of lexical and

morpho-syntactic diversity. Increase in diversity for MWE identification should come from two factors. First, known MWEs (WP1) will be linked to new corpus occurrences in the pseudo-gold sense/frame-annotated corpus (WP2). Second, new MWEs (WP3.3), will also come with pseudo-gold corpus occurrences. This should jointly increase the coverage of rare MWEs and rare realizations of frequent MWEs. To estimate this gain, manually annotated PARSEME corpora (Ramisch et al. 2020) will be extended with fragments of the pseudo-gold corpus containing MWE occurrences. The resulting joint corpus will be first assessed for its representativeness wrt. rare phenomena (D5.1.1). Second, state-of-the art MWE identifiers (Taslimipoor et al. 2020; Pasquer et al. 2020) will be trained on this joint corpus (optimally split with D5.1.2) and their results will be evaluated for diversity (D5.1.3), in comparison to tools trained on the PARSEME corpora only.
Deliverables: D5.2.1 MWE-annotated corpus including gold PARSEME data and pseudo-gold occurrences, optimally split for diversity; **D5.2.2** lexicon assessment in terms of account of MWE diversity.

➢ WP5: Risks and fallback solutions

WP5.1 is relatively independent of previous WPs and can use pre-existing data, e.g. the PARSEME shared task corpora (Ramisch et al., 2020). Most other deliverables of WP5 rely on the results of WPs 1 to 3. To prevent bottlenecks, we will employ Agile, using existing resources and first-iteration results of other WPs.

II. Impact and benefits of the project

The expected broader **scientific** impact of SELEXINI includes:

- **driving mentalities/practices** of the scientific NLP community towards considering interpretability and diversity on par with statistical efficiency while evaluating language technology;
- extending the notion of interpretability of NLP methods to **"interpretability by construction"**: continuous and symbolic representations are jointly induced, stored and applied (Rudin 2019);
- establishing new benchmarks in semantic lexical encoding and text processing, by systematically addressing single words and **multiword expressions** in the same framework;
- enhancing **multilingual perspectives** for sense induction, as the domain is mostly English-centered. We focus on French but our methods will be easy to adapt to other languages: they are language independent or depend on resources existing for many languages (e.g. Wiktionary, parsers).
- paving the way towards NLP applicability in **related scientific domains**, such as digital humanities.
- providing new quantification methods for phenomena subject to theoretical linguistic debates such as discrete vs. continuous nature of lexical meaning, argument vs. adjunct distinction (Przepiórkowski 2016), or language universals vs. linguistic diversity (Evans & Levinson 2009)

Even if SELEXINI does not directly address commercial applicability, we expect a long-term economic impact. Namely, enhanced interpretability and diversity should **reduce the distance to market** for NLP applications. Current benchmark-driven development exhibits strong bias stemming from over-tuning. Thus, state-of-the-art tools do not easily scale up to commercial application, due to lack of sufficiently large annotated domain-specific data. By focusing on statistically underrepresented phenomena, crucial for diversity, we pave the way towards better cross-domain applicability of NLP tools. Also, SMEs wishing to integrate NLP modules in their software need to be able to explain the results and errors of their models, and to customize them on demand. This cannot be achieved with most deep-learning-driven methods unless interpretability and diversity are taken seriously. The by-construction approach to interpretability, proposed by SELEXINI, should thus increase the economic value of NLP applications in the long run.

We dedicate large parts of WP0 to **dissemination**, which is crucial for perpetuating the outcomes of a project. We value **Open Access** for the benefit of both academics and end-users. Data and software will be released under open licenses, clearly defining how to integrate them in industry applications. High-rank conferences and journals, and European and French infrastructures (e.g. CLARIN and Ortolang) will be favored, for visibility. The outcomes of the project will be presented to a **larger public**. We will disseminate results to non-experts, notably young public, by actions such as young researchers' days, science days, etc.

III. References related to the project

Aloui C., **Nasr A.**, **Barque L.**, **Ramisch C.** (2020). *SLICE: Supersense-based Lightweight Interpretable Contextual Embeddings*. In COLING 2020.

Al Saied H., **Candito M.**, **Constant M.** (2019). *Comparing linear and neural models for competitive MWE identification*, In NoDaLiDa 2019.

Amplayo R. K., Hwang S., Song M. (2018). *AutoSense Model for Word Sense Induction*. In CoRR, abs/1811.09242.

Amrani A., Golberg Y. (2019). *Towards better substitution-based word sense induction*. In CoRR, abs/1905.12598.

Banarescu L., Bonial C., Cai S., Georgescu M., Griffitt K., Hermjakob U., Knight K., Koehn P., Palmer M., Schneider N. (2013). *Abstract Meaning Representation for Sembanking*. In LAW 2019, pp. 178–186.

Barque L., Haas P., Huyghe R., Tribout D., **Candito M.**, **Crabbé B.**, Segonne V. (2020). *FrSemCor: Annotating a French corpus with supersenses*. In LREC 2020, pp. 5912–5918.

Basu S., Banerjee A., Mooney R. (2002). *Semi-supervised clustering by seeding*. In ICML 2002.

Bonfante G., Guillaume B. (2018). *Non-size increasing graph rewriting for natural language processing*. In Mathematical Structures in Computer Science, 28(8), pp. 1451–1484.

Bonial C., Stowe K., Palmer M. (2013). *Renewing and Revising SemLink*. In LDL-2013, pp. 9-17.

Brunato D., Dell'Orletta F., Venturi G., François T., Blache P. (2016) *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, The COLING 2016 Organizing Committee.

Candito M., Guillaume B., Perrier G., Seddah D. (2017). *Enhanced UD Dependencies with Neutralized Diathesis Alternation*. In Depling 2017.

Candito M., Constant M., Ramisch C., Savary A., Guillaume B., Cordeiro S., **Parmentier Y.** (2020). *A French corpus annotated for MWEs and named entities*. In Journal of Language Modeling, 8(2), pp. 415–479.

Constant M., Eryiğit G., Monti J., van der Plas L., **Ramisch C.**, Rosner M., Todariscu A. (2017). *Multiword Expression Processing: A Survey*. In Computational Linguistics, 43(4), pp. 837–892.

Charlet D., Damnati G., **Béchet F.**, Marzinotto G., Heinecke J. (2020). *Cross-lingual and Cross-domain Evaluation of Machine Reading Comprehension with Squad and CALOR-Quest Corpora*. In LREC 2020, pp. 5491–5497.

Cordeiro S., Villavicencio A., Idiart M., **Ramisch C.** (2019). *Unsupervised Compositionality Prediction of Nominal Compounds*. Computational Linguistics, 45(1), pp. 1–57.

D'Hoffschmidt M., Vidal M., Belblidia W., Brendlé T. (2020). *FQuAD: French Question Answering Dataset*. In CoRR, abs/2002.06071.

Danlos L., Pradet Q., **Barque L.**, Nakamura T., **Constant M.** (2016). *Un Verbenet du français*. In Traitement Automatique des Langues, 57 (1), pp.25.

Del Corro L., Gemulla R., Weikum G. (2014). *Werdy: Recognition and Disambiguation of Verbs and Verb Phrases with Syntactic and Semantic Pruning*. In EMNLP 2014, pp. 374–385.

Delli Bovi C., Camacho-Collados J., Raganato A., Navigli R. (2017). *EuroSense: Automatic Harvesting of Multilingual Sense Annotations from Parallel Text*. In ACL 2017, pp. 594-600.

Devlin J., Chang M.-W., Lee K., Toutanova K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. In NAACL 2019, pp. 4171–4186.

Djemaa M., **Candito M.**, Muller P., Vieu L. (2016). *Corpus annotation within the French FrameNet: methodology and results*. In LREC 2016.

Evans N., Levinson S. (2009). *The myth of language universals: Language diversity and its importance for cognitive science*. Behavioral and Brain Sciences 32.

Flati T., Navigli R. (2013). *SPred: Large-scale Harvesting of Semantic Predicates*. In ACL 2013, pp. 1222–1232.

Guo S., Li R., Tan H., Li X., Guan Y., Zhao H., Zhang Y. (2020). *A Frame-based Sentence Representation for Machine Reading Comprehension*. In ACL 2020.

Kallmeyer L., QasemiZadeh B., Cheung J. C. K. (2018). *Coarse Lexical Frame Acquisition at the Syntax–Semantics Interface Using a Latent-Variable PCFG Model*. In *SEM 2018, pp. 130–141.

Le H., Vial L., Frej J., Segonne V., Coavoux M., Lecouteux B., Allauzen A. **Crabbé B.**, Besacier L., Schwab D. (2020). *FlauBERT: Unsupervised Language Model Pre-training for French*. In LREC 2020, pp. 2479–2490.

McShane M., Nirenburg S., Beale S. (2015). *The Ontological Semantic treatment of multiword expressions.* In Lingvisticæ Investigationes, 38(1), pp. 73–110.

Marzinotto G., Auguste J., **Béchet F.**, Damnati G., **Nasr A.** (2018) *Semantic Frame Parsing for Information Extraction : the CALOR corpus*. In LREC 2018.

Marzinotto G., Damnati G., **Béchet F.**, **Favre B.** (2019). *Robust Semantic Parsing with Adversarial Learning for Domain Generalization*. In NAACL 2019, pp. 166–173

Mickus T., Paperno D., **Constant M.** (2019). *Mark my Word: A Sequence-to-Sequence Approach to Definition Modeling*. In First NLPL Workshop on Deep Learning for Natural Language Processing.

Morales P. L., Lamarche-Perrin R., Fournier-S'niehotta R., Poulain R., Tabourier L., Tarissan F. (2021) Measuring Diversity in Heterogeneous Information Networks, in Theoretical Computer Science, Elsevier.

Narayan S., Cohen S. (2015) *Diversity in Spectral Learning for Natural Language Parsing*. In EMNLP 2015.

Navigli R., Ponzetto S. (2012) *Babelnet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network*. In Artificial Intelligence (193), pp. 217–250.

Navigli R., Jurgens D., Vannella D. (2013). *SemEval-2013 Task 12: Multilingual Word Sense Disambiguation*. In SemEval 2013.

Panchenko A., Ruppert E., Faralli S., Ponzetto S. P., Biemann C. (2017). *Unsupervised Does Not Mean Uninterpretable: The Case for Word Sense Induction and Disambiguation*. In EACL 2017, pp. 86–98.

Palumbo E., Mezzalira A., Marco C., Manzotti A., Amberti D. (2020). *Semantic Diversity for Natural Language Understanding Evaluation in Dialog Systems*. In COLING 2020, pp. 44-49.

Pasquer C., **Savary A.**, **Antoine J.-Y.**, **Ramisch C.** (2018). *If you've seen some, you've seen them all: Identifying variants of multiword expressions*. In COLING 2018.

Polguère A. (2014). *From Writing Dictionaries to Weaving Lexical Networks*. In International Journal of Lexicography, 27(4).

Adam Przepiórkowski. (2016). *Against the argument–adjunct distinction in Functional Generative Description*. In Prague Bulletin of Mathematical Linguistics, 106:5–20.

QasemiZadeh B., Petruck M., Stodden R., Kallmeyer L., **Candito M.** (2019). *SemEval-2019 task 2: Unsupervised lexical frame induction*. In SemEval 2019, pp. 16–30.

Ramisch C., **Savary A.**, Guillaume B., Waszczuk B., **Candito M.**, et al. (2020). *Edition 1.2 of the PARSEME Shared Task on Semi-supervised Identification of Verbal Multiword Expressions*. In MWE-LEX 2020, pp. 107—118

Raganato A., Camacho-Collados J., Navigli R. (2017). *Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison*. In *EACL*, pp. 99–110.

Rogers A., Kovaleva O., Rumshisky A. (2020). *A Primer in BERTology: What we know about how BERT works*. In TACL 8, pp. 842–866.

Rothe S., Schütze H. (2015) *AutoExtend: Extending Word Embeddings to Embeddings for Synsets and Lexemes*. In ACL-IJCNLP 2015, pp. 1793–1803.

Rudin C. (2019) *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead*. In *Nature Machine Intelligence, 1*, pp. 206–215.

Sagot B., Fišer D. (2008). *Building a free French Wordnet from multilingual resources*. In OntoLex 2008.

Savary A., Ramisch C., Cordeiro S., Sangati F., Vincze V., QasemiZadeh B., **Candito M.**, et al. (2017). *The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions*. In MWE 2017, pp. 31–47.

Savary A., Cordeiro S., **Ramisch C.** (2019). *Without lexicons, multiword expression identification will never fly: A position statement*. In MWE-WN 2019, pp. 79–91.

Schneider N., Hovy D., Johannsen A., Carpuat M. (2016). *SemEval-2016 Task 10: Detecting Minimal Semantic Units and their Meanings (DiMSUM)*. In SemEval 2016.

Scholivet M., Dary F., **Nasr A.**, **Favre B.**, **Ramisch C.** (2019). *Typological Features for Multilingual Delexicalised Dependency Parsing*. In NAACL-HLT 2019, pp. 3919-3930.

Segonne V., **Candito M.**, **Crabbé B.** (2019). *Using Wiktionary as a resource for WSD : the case of French verbs*. IWCS 2019.

Sérasset G. (2012). *Dbnary: Wiktionary as a LMF based multilingual RDF network*. In LREC 2012.

Soulet A., Giacometti A., Markhoff B., Suchanek F. M. (2018). *Representativeness of knowledge bases with the generalized Benford's law*. In ISWC 2018, pp. 374-390.

Taslimipoor S., Bahaadini S., Kochmar E. (2020). *MTLB-STRUCT @Parseme 2020: Capturing Unseen Multiword Expressions Using Multi-task Learning and Pre-trained Masked Language Models*, In MWE-LEX 2020.

Ustalov D., Panchenko A., Kutuzov A., Biemann C., Ponzetto S. P. (2018). *Unsupervised semantic frame induction using triclustering*. In ACL 2018.

Ustalov D., Panchenko A., Biemann C., Ponzetto S. P. (2019). *Watset: Local-Global Graph Clustering with Applications in Sense and Frame Induction*. In Computational Linguistics, 45(3):423–479.

Vial L., Lecouteux B., Schwab D. (2019). *Sense Vocabulary Compression through the Semantic Knowledge of Wordnet for Neural Word Sense Disambiguation*. In CoRR, abs/1905.05677.

Yang Z., Qi P., Zhang S., Bengio Y., Cohen W., Salakhutdinov R. Manning C. (2018). *HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering*. In EMNLP 2018.

Williams J. R., Lessard P. R., Desu S., Clark E. M., Bagrow J. P., Danforth C. M., Dodds P. S. (2015). *Zipf's law holds for phrases, not words*. Scientific Reports, 5.

Wisniewski G., Yvon F. (2019). *How Bad are PoS Tagger in Cross-Corpora Settings? Evaluating Annotation Divergence in the UD Project*. In NAACL 2019, pp. 218–227.

Yan X., Yang S. G., Kim B. J., Minnhagen P. (2018). *Benford's law and first letter of words*. In Physica A: Statistical Mechanics and its Applications, 512, pp. 305-315.

Xie J., Girshick R., Farhadi A. (2016). *Unsupervised deep embedding for clustering analysis*. In ICML 2016.

Zhang H., Basu S., Davidson I. (2020). *A Framework for Deep Constrained Clustering - Algorithms and Advances*. In Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2019. LNCS (11906).

Zhang Y., Weiss D. (2016). *Stack-propagation: Improved Representation Learning for Syntax*. In ACL 2016, pp. 557–1566.