Design and automatic induction of a multiword expression lexicon at the service of linguistic diversity

PhD position in computational linguistics

- **Application deadline**: May 14, 2020
- **Field**: natural language processing/computational linguistics
- **Supervisors**:
  - Agata Savary (University of Tours/Blois, France)
  - Emmanuel Schang (University of Orléans, France)
  - Laura Kallmeyer (University of Düsseldorf, Germany)
- **Duration**: 3 years, October 2020 to September 2023; additional 1 year in Germany, if funding confirmed
- **Remuneration**: around 1400 €/month (possibly combined with an additional remuneration for part-time teaching)
- **Funding**: Centre-Val de Loire Region
- **Keywords**: multiword expressions, lexicon induction, machine learning, diversity, multiword expression identification, parsing

Candidate's profile

- Master in computer science or computational linguistics
- Age below 30 (condition from the funding body)
- Interests in linguistics
- Foundations of NLP methods and machine learning. Deep learning skills would be a plus.
- Good programming skills and experience in integration of libraries and infrastructures
- Good knowledge of French and English, another language (especially Creole) would be a plus
- Foundations of formal languages and formal grammars
- Good writing skills
- Capacity to work independently and as part of a team
- Blois-Orléans mobility, availability for short-term visits to Düsseldorf (Germany)
- Availability for a 1-year stay in Düsseldorf (if funding is confirmed for the 4th year of the PhD)

Important dates

- Application deadline: May 14, 2020 (or until filled)
- Notification: June 2, 2020
- Position starts: October 2020
- Position ends: September 2023

Application

Candidates should send the following documents in PDF format, in French or in English, to Agata Savary (FirstName.LastName@univ-tours.fr) and Emmanuel Schang (FirstName.LastName@univ-orleans.fr):

- CV
- Cover letter
- Transcript of Master's and Bachelor's grades (translated if not in French or English)

Hosting Institutions

- Main affiliation: University of Tours, LIFAT laboratory, BdTln team, campus in Blois
- Secondary affiliation: University of Orléans, LLL lab, DDL team

Prospect of a French-German co-supervision

The PhD position is meant to be transformed into a French-German co-supervision (*co-tutelle*), provided that the expected funding is granted. The first 3 years of research, funded by France, will be carried out in Blois and Orléans, with short-term visits at the University of Düsseldorf, Germany. The 4th year, funded by Germany, would be spent in Düsseldorf, under the supervision of prof. Laura Kallmeyer.

Context, challenges and positioning

This PhD position concerns the field of computational linguistics (CL), which belongs to a larger domain of artificial intelligence. On the scientific side, CL is concerned with understanding written and spoken natural language (i.e. language spoken by humans, as opposed to formal languages dedicated to machines) from a computational perspective. On the engineering side, CL, also called natural language processing (NLP), aims at building models and software to usefully process and produce language. End-user applications in the domain include machine translation, information retrieval, information extraction, sentiment analysis, speech recognition and synthesis, computer-aided language learning, and many others.

This PhD position addresses one of the major challenges in CL: **multiword expressions** (MWEs). MWEs such as *to pay a visit*, *to pull one's leg* or *the die is cast*, are combinations of words exhibiting unexpected linguistic behavior. Most prominently, they are semantically non-compositional, that is, their meaning cannot be deduced from the meanings of their components and from their syntactic structure in a way deemed regular. For instance, *to pull one's leg* means 'to deceive someone playfully', which has few explicit links with the component words *pull* and *leg*. MWEs can also exhibit syntactic irregularities, e. g. *the die is cast* (‘*a point of no-retreat has been passed*’) allows no inflection of the noun (*#the dies are cast*) and no active voice (*#someone cast the die*). Violating these constraints leads to the loss of the idiomatic reading.

State-of-the-art NLP methods, heavily relying on supervised machine learning, deep learning and distributional semantics, bring substantial contribution to some aspects of semantic non-compositionality (Cordeiro et al. 2019) and syntactic irregularity (Pasquer et al. 2018, Rohanian et al. 2019, Waszczuk et al. 2019) of MWEs. These methods, however, suffer from opacity of models, evaluation biases and insufficient robustness, posing the risk of undermining **intra- and inter-linguistic diversity**. Namely, most language

phenomena, including MWEs, are known to have a so-called **Zipfian distribution**, i.e. few of them occur frequently in texts and there is a long tail of those occurring rarely. Modern NLP algorithms strongly favor the former and often underperform in the latter, since they are conceived and tuned for optimal global performances in which the former dominate (Savary et al. 2019). This sensitivity of NLP to data sparseness endangers the **diversity within a language**. Secondly, recent progress in NLP mainly concerns a few economically dominating languages, especially English, while a vast majority of other languages, e.g. **Creole languages** (Duchier et al. 2012, Petitjean & Schang 2018, Schang 2018), are low-resourced or undocumented.

This PhD thesis will address the above-mentioned challenges by the design and automatic induction of a **syntactic-semantic lexicon of MWEs**. In the context of a quest for diversity, electronic lexicons are complementary to texts because they aim at holistic language modeling, describing possibly many linguistic objects, whereas in texts many phenomena occur rarely or never. While lexical encoding of MWEs has a long-standing tradition (Savary 2008), current lexicons need enhancements and extensions, especially to encode syntactic and semantic properties of MWEs, and to be easily interfaced with NLP tools (Lichte et al. 2019, Lichte & Kallmeyer 2016, Savary et al. 2018).

This PhD thesis will focus on: (i) a design of a MWE lexicon, unified across many languages, including low-resourced one, such as Guadeloupean or Kriol, (ii) weakly supervised automatic induction of a structured MWE lexicon, (iii) filtering the lexicon to balance the Zipfian distribution of MWEs in texts, (iv) integrating the lexicon in NLP tasks such as MWE identification or syntactic/semantic parsing, (v) design of evaluation measures for the promotion of linguistic diversity.

References

- Tatiana Bladier, Jörg Hendrik Janke, Jakub Waszczuk and Laura Kallmeyer (2019): From partial neural graph-based LTAG parsing towards full parsing. Abstract accepted for presentation at the 29th Computational Linguistics in the Netherlands conference CLIN, Groningen, The Netherlands, January 2019.
- Silvio Cordeiro, Aline Villavicencio, Marco Idiart, Carlos Ramisch (2019): Unsupervised Compositionality Prediction of Nominal Compounds, Computational Linguistics, 45(1):1--57, MIT Press.
- Duchier, D., Ekoukou, B. M., Parmentier, Y., Petitjean, S., & Schang, E. (2012). Describing morphologically-rich languages using metagrammars: A look at verbs in Ikota. *Language Technology for Normalisation of Less-Resourced Languages*, 55.
- Timm Lichte and Laura Kallmeyer (2016). Same Syntax, Different Semantics: A Compositional Approach to Idiomaticity in Multi-word Expressions. In *Empirical Issues in Syntax and Semantics* EISS 11.
- Timm Lichte, Simon Petitjean, Agata Savary, and Jakub Waszczuk (2019): Lexical encoding formats for multi-word expressions: The challenge of “irregular” regularities. In Yannick Parmentier and Jakub Waszczuk, editors, Representation and Parsing of Multiword Expressions, pages 41–72. Language Science Press, Berlin.
- Caroline Pasquer, Agata Savary, Jean-Yves Antoine, Carlos Ramisch (2018): If you’ve seen some, you’ve seen them all: Identifying variants of multiword expressions, in the Proceedings of the 27th International Conference on Computational Linguistics (COLING-18), Santa Fe, USA.
- Omid Rohanian, Shiva Taslimipoor, Samaneh Kouchaki, Le An Ha, Ruslan Mitkov (2019): Bridging the Gap: Attending to Discontinuity in Identification of Multiword Expressions. NAACL-HLT (1) 2019: 2692-2698
- Agata Savary, Silvio Cordeiro, Carlos Ramisch (2019) Without lexicons, multiword expression identification will never fly: A position statement, in the Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019), 2 August 2019, Florence, Italy.
- Agata Savary (2008): Computational Inflection of Multi-Word Units, a contrastive study of lexical approaches, in Linguistic Issues in Language Technology 1(2), CSLI, pp. 1-53.

- Simon Petitjean & Emmanuel Schang (2018): Sentential Negation and negative Words in Guadeloupean Creole. In *Negation and Negative Concord, the View from Creoles.* Déprez & Henri (eds).
- Agata Savary, Simon Petitjean, Timm Lichte, Laura Kallmeyer, Jakub Waszczuk (2018) *Object-oriented lexical encoding of multiword expressions: Short and sweet*, CoRR abs/1810.09947
- Emmanuel Schang, Denys Duchier, Brunelle Magnana Ekoukou, Yannick Parmentier, Simon Petitjean (2012) *Describing São Tomense Using a Tree-Adjoining Meta-Grammar*. TAG 2012: pp. 82-89
- Emmanuel Schang (2018). A Metagrammatical Approach to Periphrasis in Gwadloupéyen. *Quaderni di Linguistica e Studi Orientali*, *4*, 131-149.
- Jakub Waszczuk, Rafael Ehren, Regina Stodden, Laura Kallmeyer (2019): A Neural Graph-based Approach to Verbal MWE Identification. MWE-WN@ACL 2019: 114-124

Mailing lists to use:

corpora@uib.no
ln@groupes.renater.fr
elsnet-list@elsnet.org
http://alt.qcri.org/siglex/messboard.php
parseme-all@nlp.ipipan.waw.pl
http://linguistlist.org/
elexis-all@googlegroups.com
multiword-expressions@lists.sourceforge.net

lifat@listes.univ-tours.fr
lift_members@inria.fr