Verbal MWE annotation and semantic models for MWE identification
STSM Report - Cost action IC1207 PARSEME

Applicant: Carlos Ramisch - LIF & Aix Marseille Université (France)
Host: Aline Villavicencio, Federal Univ. of Rio Grande do Sul (Brazil)

July 11 to September 6, 2016

1 Reminder of STSM Goals

This report describes my PARSEME STSM application for an 8-week stay, which took place from July 11 to September 6, 2016 at the Federal University of Rio Grande do Sul (Brazil) to visit Aline Villavicencio and her research group. The goal of the STSM was to work in two relevant topics for PARSEME IC1207 Cost action: make significant progress on the annotation of verbal MWEs in Brazilian Portuguese and work on the development of a general-purpose MWE identification tool including semantic features.

The mission took place in the context of an ongoing and long-term collaboration between the candidate and the host research group. The expected concrete outcomes of the STSM were a large corpus and related tools containing verbal MWE annotation in Brazilian Portuguese and an MWE identification tool, open source and integrated into the mwetoolkit. Both outcomes are relevant for the PARSEME shared task on verbal MWE detection, currently in preparation.

This report describes the carried out activities, some goal adaptations and the concrete outcomes. It also discusses ongoing collaboration and expected future results of this STSM.

2 STSM activities

The goal of the STSM was twofold: make significant progress on the annotation of verbal MWEs in Brazilian Portuguese and work on the development of a general-purpose MWE identification tool including semantic features. Some of these goals were achieved, others are in progress, specially due to changes in the schedule of the PARSEME shared task (henceforth ST). We also carried out activities that are relevant to PARSEME but that were not initially planned. Most of the activities below were carried out in collaboration with Aline Villavicencio and Silvio Cordeiro, who is doing his PhD under the joint supervision of Aline, Alexis Nasr and myself. I have also interacted with Brazilian colleagues Marco Idiart, Viviane Moreira and Maria José Finatto, the PhD students Rodrigo Wilkens and Alex Salle, and undergraduate students.

2.1 Scientific visit to São Carlos

One of the activities carried out during the STSM was a scientific visit to the city of São Carlos, in Brazil. This visit was funded by the AIM-WEST project[1] and by the University of São Paulo (USP). This visit was not initially planned, but included later in the STSM schedule as it is relevant for PARSEME as well. It took place from August 29 to September 2 (5 days) and covered activities in two local universities.

The first university visited was Universidade Federal de São Carlos (UFSCar). The contact person there is Helena Caseli, with whom I have worked in the past and also work currently in AIM-WEST project. I have given a seminar on parsing complex conjunctions and determiners in French (presented at ACL 2015). The seminar was open to all students and researchers of the university.

We have worked together on an extension to English of our system for never-ending multiword expressions learning (NEMWEL), initially developed by a co-supevised student Alexandre Rondon (presented at MWE 2013).

I have also interacted with Natalie Vargas, MsC student under the supervision of Helena Caseli. We have decided that I would co-supervise the masters thesis. The thesis will be about bilingual discovery of MWEs in parallel corpora, specially focusing on light verb constructions.

The second institution visited was University of São Paulo (USP). There, I have been a member of the MsC thesis proposal committee of Franciele Vargas. Her work was on ontology learning from opinion texts (reviews) for automatic ontology-based information extraction and summarisation.

I have also given a mini-course on the mwetoolkit. This was attended by around 40 students and researchers. It had 2 parts of 3h each. This was adapted from the PARSEME tutorial given in Malta and Iasi as part of WG2 activities. During this occasion, I also presented the PARSEME ST and recruited an extra annotator for Portuguese, Amanda Carneiro, a linguistics student.

I have also met individually around 8 MsC, PhD students and post-docs working at NILC (Núcleo Interinstitutcional de Linguística Computacional). They explained me their projects, ranging from twiter normalization to semantic parsing and word sense disambiguation. This was a wonderful opportunity to spot possible future interactions.

We intend to continue this cooperation via an application to a French-Brazilian CAPES-COFECUB project.[2] We would like to structure this proposal around the theme of lexical-semantic resources for French and Portuguese. We will continue to discuss ideas for synergy and follow-up work during the forthcoming joint AIM-WEST and PARSEME-FR workshop in Grenoble on October 3-4, 2016.[3].

The detailed schedule and extra materials about this visit is available on the AIM-WEST project website.[4]

2.2 PARSEME ST annotation guidelines

During my STSM, I worked with Agata Savary and Silvio Cordeiro on the new dynamic hypertext (HTML) guidelines for the PARSEME ST. We have created a browsable version of the guidelines,

[1] http://aim-west.imag.fr

[2] http://www.capes.gov.br/cooperacao-internacional/franca/cofecub

[3] http://aim-west.imag.fr/joint-aim-west-parseme-fr-workshop-oct-3-4-grenoble-fr/

[4] http://aim-west.imag.fr/aim-west-visit-to-ufscar-and-usp/

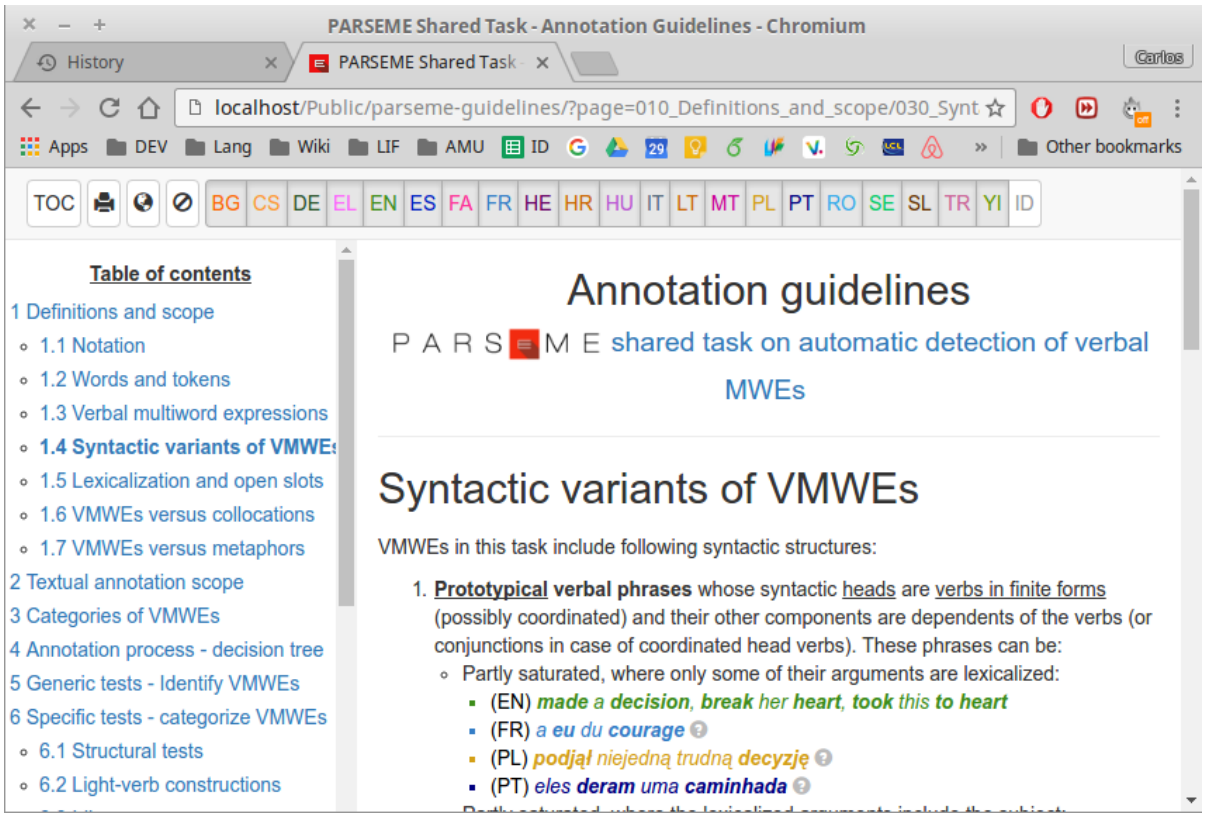

Figure 1: New PARSEME ST annotation guidelines - example page

which includes features like:

- Collapsible examples in different colors in all languages
- Dynamic table of contents
- Hover-like tooltips for glosses and translations
- Hover-like tooltips for footnotes
- Different visualisation modes (full-text printable or chaptered)

We have also improved the guidelines contents. We have unified the notation and formatting of examples. We have integrated all PDF documents into a single website. We have homogenised the structure of linguistic tests. Most of the improvements were implemented by Silvio Cordeiro under my supervision in accordance with my host and Silvio's co-supervisor Aline Villavicencio.

These novelties were presented to the community before the Annotathon of the final PARSEME meeting.[5] We are currently including examples from all other languages using an online examples spreadsheet. An example page can be seen in Figure 1.

[5] `parsemefr.lif.univ-mrs.fr/guidelines-hypertext`

2.3 PARSEME ST Brazilian Portuguese team

Concerning the Brazilian Portuguese part of the ST, we did not make as much progress as we wanted to, but this is ongoing work. We have selected the corpus to be annotated, which will be a mixture of a local newspaper *Diário Gaúcho* and wikipedia articles. We expect that the local newspaper, with its simple and informal style, will contain many idioms, while wikipedia will contain more literal expressions such as LVCs and IReflVs.

We have also included examples in Brazilian Portuguese in all sections of the new guidelines mentioned above. We have shown the new version of guidelines to Brazilian annotators and we are currently setting up annotator accounts on FLAT to start the annotation process.

2.4 PARSEME WG3 MWE processing survey

Another unscheduled activity that took place during my STSM was working on the PARSEME WG3 survey on MWE processing. This survey is being written since the Frankfurt meeting with co-authors Matthieu Constant, Gülşen Eryigit, Federico Sangati, Johanna Monti, Mike Rosner, Amalia Todirascu and Lonneke van der Plas.

I have considerably revised the MWE Discovery section, including 2016 references, restructuring the section on semantic-based discovery, and adding subsections on evaluation and open issues. I have also added a section on open issues to the MWE Identification section. The paper is currently being revised one last time before the first submission to the Computational Linguistics journal takes place in mid-october.

2.5 Nominal compound semantic processing

During my STSM I assisted to Silvio Cordeiro's PhD thesis proposal defence. His work on nominal compoound compositionality prediction was published in several conferences this year including LREC and ACL.

Together with his Brazilian supervisor Aline Villavicencio, we decided to extend this work to answer many open questions concerning (a) applicability to Portuguese language, (b) influence of corpus sizes (c) influence of composition function, etc.

We have written the structure of a journal paper on these extended experiments. The goal of this paper is to present a method for compositionality prediction of nominal compounds and then evaluate it extensively in Portuguese, French and English.

Many experiments for this extended journal article are now executed and some are still being executed due to bugs and complex parameter interference. The paper's expected submission date is end of November 2016.

2.6 Semantic features for MWE identification

During the STSM I have worked with a MsC intern in France, Manon Scholivet. She has started the internship before my departure and we continued the work remotely with weekly skype meetings. She has developed a CRF-based MWE tagger capable of identifying MWEs in running text based on a supervised learning model. The model for MWE identification is part of the mwetoolkit[6]

[6] `http://mwetoolkit.sf.net`

and was recently made available and documented to the community. We intend to use it in our participating system for the PARSEME ST.

One of the goals of the STSM was to integrate the compositionality measures developed with Silvio Cordeiro into Manon Scholivet's tagger. Unfortunately, this was not possible yet. Instead, Manon has developed a generic mechanism for including external resources resulting from automatic MWE discovery.

We are currently carrying out experiments to integrate both works. A joint paper is planned but some recently discovered bugs have delayed its redaction and submission. We believe that it will be submitted to EACL short papers or ACL long papers.

3 Concrete outcomes

The concrete outcomes planned for this STSM were those listed below. For each outcome, we indicate its status.:

- Large Brazilian Portuguese corpus annotated for verbal MWEs with around 2000 to 3000 MWE token instances, in accordance with general shared task requirements - **ongoing - corpus selected and annotation will take place in Sep-Dec 2016**
- Portuguese-specific guidelines, tokeniser and tools for annotation and adjudication- **guidelines finalised - tokeniser will be reused, tools for annotation are being set up in coordination with ST technical support**
- System for automatic MWE identification including semantic features, integrated and released with the mwetoolkit - **done**
- Poster presentation proposal for following PARSEME general meeting - **done**
- Submission to international conference on the evaluation of the identification system - **ongoing - paper to be submitted as short EACL or long ACL paper**

We also list concrete outcomes that were not initially planned.

- Hypertext dynamic guidelines for PARSEME ST
- Paper on noun compound compositionality prediction planned for submission to journal in the end of November 2016
- A new cooperation at UFSCar, with the joint MsC supervision of Natalie Vargas

4 Host confirmation

The present report was reviewed and accepted by the STSM host Aline Villavicencio.