Fifth CAMELEON Workshop

October 24, 2014, Porto Alegre

CAMELEON

Collaborative and **A**utomatic **ME**thods for the Multilingualisation of **LE**xica and **ON**tologies

http://cameleon.imag.fr

Project overview

- Funding from CAPES (Brazil) and COFECUB (France) -
- 4-years: Jan 2011 to Dec 2014
- Research and student missions in both directions
- Part of LICIA international laboratory
- Yearly CAMELEON workshops

Partners

- LIG: GETALP and EXMO teams
- Univ. Toulouse: MELODI
- Aix Marseille Univ.: TALEP
- UFRGS: INF and Letras
- UFSCar: Departamento de Computação

Keywords

- Natural Language Processing
- Semantic Web
- Lexical Resources
- Ontologies
- Multilingualism
- Machine Translation
- Information Retrieval
- Lexical Acquisition
- Ontology Matching
- Multiword Expressions

Outcomes: Formation

- 5 PhD students (2 cotutelles), 1 post-doc
- 15 Researcher missions
- Co-supervision of IC, TCC and Master students
- Workshops, tutorials and mini-courses

Outcomes: Tools and Resources

- `mwetoolkit`
- CAMELEON corpora
- Focused crawlers
- SRL lexicon for Portuguese
- SMT systems
- JeuxDeMots Português

Publications I

- Ramisch, C., Villavicencio, A., and Boitet, C. (2010d). Web-based and combined language models: a case study on noun compound identification. In Huang, C.-R. and Jurafsky, D., editors, *Proc. of the 23rd COLING (COLING 2010) — Posters*, pages 1041–1049, Beijing, China. The Coling 2010 Organizing Committee
- Ramisch, C., Villavicencio, A., and Boitet, C. (2010b). Multiword expressions in the wild? the mwetoolkit comes in handy. In Liu, Y. and Liu, T., editors, *Proc. of the 23rd COLING (COLING 2010) — Demonstrations*, pages 57–60, Beijing, China. The Coling 2010 Organizing Committee
- Linardaki, E., Ramisch, C., Villavicencio, A., and Fotopoulou, A. (2010). Towards the construction of language resources for Greek multiword expressions: Extraction and evaluation. In Piperidis, S., Slavcheva, M., and Vertan, C., editors, *Proc. of the LREC Workshop on Exploitation of multilingual resources and tools for Central and (South) Eastern European Languages*, pages 31–40, Valetta, Malta. May
- Ramisch, C., de Medeiros Caseli, H., Villavicencio, A., Machado, A., and Finatto, M. J. (2010a). A hybrid approach for multiword expression identification. In *Proc. of the 9th PROPOR (PROPOR 2010)*, volume 6001 of *LNCS (LNAI)*, pages 65–74, Porto Alegre, RS, Brazil. Springer
- Villavicencio, A., Ramisch, C., Machado, A., de Medeiros Caseli, H., and Finatto, M. J. (2010). Identificação de expressões multipalavra em domínios específicos. *Linguamática*, 2(1):15–33
- Ramisch, C., Villavicencio, A., and Boitet, C. (2010c). mwetoolkit: a framework for multiword expression identification. In *Proc. of the Seventh LREC (LREC 2010)*, pages 662–669, Valetta, Malta. ELRA
- de Medeiros Caseli, H., Ramisch, C., das Graças Volpe Nunes, M., and Villavicencio, A. (2010). Alignment-based extraction of multiword expressions. *Lang. Res. & Eval. Special Issue on Multiword expression: hard going or plain sailing*, 44(1-2):59–77
- Araujo, V. D., Ramisch, C., and Villavicencio, A. (2011). Fast and flexible MWE candidate generation with the mwetoolkit. In [Kordoni et al., 2011], pages 134–136
- Duran, M. S., Ramisch, C., Aluísio, S. M., and Villavicencio, A. (2011). Identifying and analyzing Brazilian Portuguese complex predicates. In [Kordoni et al., 2011], pages 74–82

Publications II

- Duran, M. S. and Ramisch, C. (2011). How do you feel? investigating lexical-syntactic patterns in sentiment expression. In *Proceedings of Corpus Linguistics 2011: Discourse and Corpus Linguistics Conference*, Birmingham, UK
- Ramisch, C. (2012a). A generic framework for multiword expressions treatment: from acquisition to applications. In [con, 2012], pages 61–66
- Ramisch, C., Araujo, V. D., and Villavicencio, A. (2012). A broad evaluation of techniques for automatic acquisition of multiword expressions. In [con, 2012], pages 1–6
- Mangeot, M. and Ramisch, C. (2012). A serious lexical game for building a Portuguese lexical-semantic network. In *Proceedings of the ACL 2012 3rd Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources and their Applications to NLP*, Jeju, Republic of Korea. Association for Computational Linguistics
- Granada, R., Lopes, L., Ramisch, C., Trojahn, C., Vieira, R., and Villavicencio, A. (2012). A comparable corpus based on aligned multilingual ontologies. In *Proceedings of the ACL 2012 First Workshop on Multilingual Modeling (MM 2012)*, Jeju, Republic of Korea. Association for Computational Linguistics
- Ramisch, C. (2012b). Une plate-forme générique et ouverte pour le traitement des expressions polylexicales. In Molina Mejia, J. M. and Schwab, D., editors, *Actes de 14e Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL 2012)*, Grenoble, France
- Villavicencio, A., Idiart, M., Ramisch, C., Araujo, V. D., Yankama, B., and Berwick, R. (2012). Get out but don't fall down: verb-particle constructions in child language. In Berwick, R., Korhonen, A., Poibeau, T., and Villavicencio, A., editors, *Proc. of the EACL 2012 Workshop on Computational Models of Language Acquisition and Loss*, pages 43–50, Avignon, France. ACL

Workshop 2014

	Working session
10:30 - 12:00	1st CAMELEON Report writing session * Discuss and compare the project goals and outcomes * First outline of the final report on Gdocs * Determine the part to be written by each participant
12:00 - 14:00	**LUNCH**
	Session I - Research talks
14:00 - 14:15	Development of a lexical resource annotated with semantic roles for Portuguese Leonardo Zilio
14:15 - 14:30	Multiword expressions and lexical complexity in Portuguese and French Bianca Pasqualini
14:30 - 14:45	Extraction of hierarchical relations from texts Roger Leitzke Granada
14:45 - 15:00	Simplifying Complex Textual Expressions Rodrigo Wilkens
15:00 - 15:30	**COFFEE BREAK**

Workshop 2014

15:30 - 15:50	Token-based identification of MWEs Silvio Ricardo Cordeiro
15:50 - 16:05	Builing a dataset to evaluate MWEs lexical simplification Luiza Hageman
16:05 - 16:20	Web as a Corpus: creation and annotation of very large corpora André Trevisol Trindade
	Session II - Panel and discussion
16:20 - 16:30	Presentation of CAMELEON's follow-up: AIM-WEST Carlos Ramisch
16:30 - 17:00	Project summary, results, final report Animated by Carlos Ramisch and Christian Boitet

Enjoy the workshop!

October 24, 2014, Porto Alegre

Outline

AIM-WEST

A**nalysis and **I**ntegration of **M**ulti-**W**ord **Expressions
in ***S**peech and **T**ranslation*

http://aim-west.imag.fr

Project overview

- Funding from FAPERGS-FAPESP (Brazil) and CNRS-INRIA (France)
- 3 years: Jan 2014 to Dec 2016
- Research and student missions in both directions
- One final workshops

Partners

- LIG, Lattice, LIF (France)
- UFRGS and UFSCar (Brazil)

Keywords

- Natural Language Processing
- MultiWord Expressions
- Machine Translation
- Automatic Speech Recognition
- Language Technology
- Non Compositionality
- Cognitive Lexical Models

Goals

This project aims to investigate techniques, resources and protocols for ***evaluating*** and ***integrating*** models of ***MWE processing*** into ***MT and ASR technology***.

Motivations

- Why multiword expressions (MWEs)?
- Why machine translation (MT)?
- Why automatic speech recognition (ASR)?

Motivations

en source	We only go out **once in a blue moon**
fr MT	Nous allons seulement **une fois dans une lune bleue**
en source	The **dry run** went on smoothly
fr MT	Le **fonctionnement à sec** est allé en douceur
en source	The children **ate** my cookies **up**
fr MT	Les enfants **ont mangé** mes biscuits **jusqu'à**
en source	The boy **is in** his mother's **black books**
fr MT	Le garçon **est dans les livres noirs** de sa mère
en source	MWEs are a **pain in the neck** for computers
fr MT	MWEs sont une **douleur dans le cou** pour les ordinateurs
en source	MWEs are a **tough nut to crack**
fr MT	MWEs sont un **dur à cuire**
en source	I never get on **cable cars**
fr MT	Je ne suis jamais **sur** les **voitures de câble**
en source	He rarely **gets drunk** at work
fr MT	Il **se fait** rarement **bu** au travail

Preliminary experiments [Ramisch et al., 2013]

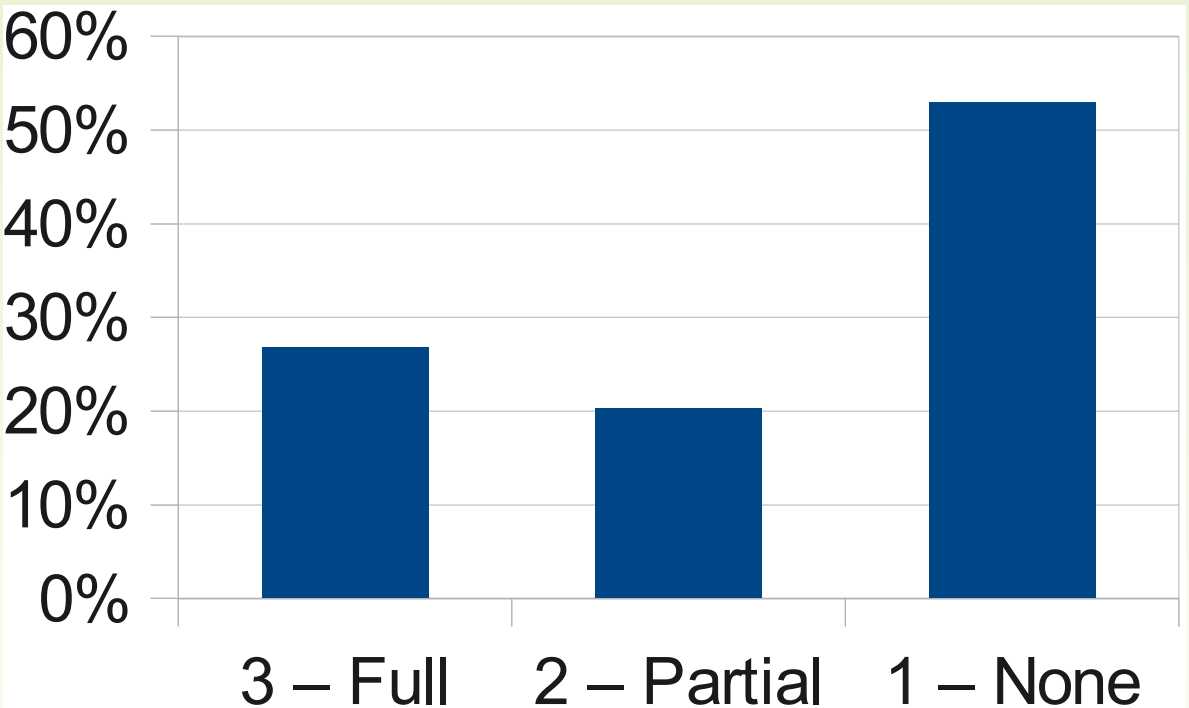
Investigate:

- What proportion of split phrasal verbs is translated correctly/acceptably by current MT paradigms?
- Which MT paradigm, phrase-based or hierarchical, can better handle these constructions?
- What are the main factors that influence translation quality?

Translation examples

	could **boil** *this poem* **down** *to saying*
PBS	*pourriez* **furoncle** *ce poème* **jusqu'** *à dire*
HS	*pourriez* **bouillir** *ce poème* **descendu** *à dire*
	he would **think** *it* **through** *and say*
Both	*il* **pense que** *ça* **à travers** *et dire*
	you couldn 't **figure** *it* **out**
HS	*vous ne pouvais pas le* **comprendre**
PBS	*vous ne pouviez pas le* **découvrir**
	Then we 'll **test** *some other ideas* **out**
Both	*puis nous allons* **tester** *certains autres idées*

Main finding: MT is quite bad



Distribution of annotations - adequacy

Ongoing work

- AIM-WEST kick-off mission in São Carlos
- Gather data sets and corpora
- Develop automatic evaluation metrics
- Improve manual evaluation metrics
- Propose a shared task on MWE translation

Merci beaucoup !

Muito obrigado!

Thank you!

References I


(2012).
Proc. of the ACL 2012 SRW, Jeju, Republic of Korea. ACL.


Araujo, V. D., Ramisch, C., and Villavicencio, A. (2011).
Fast and flexible MWE candidate generation with the mwetoolkit.
In [Kordoni et al., 2011], pages 134–136.


de Medeiros Caseli, H., Ramisch, C., das Graças Volpe Nunes, M., and Villavicencio, A. (2010).
Alignment-based extraction of multiword expressions.
Lang. Res. & Eval. Special Issue on Multiword expression: hard going or plain sailing, 44(1-2):59–77.


Duran, M. S. and Ramisch, C. (2011).
How do you feel? investigating lexical-syntactic patterns in sentiment expression.
In *Proceedings of Corpus Linguistics 2011: Discourse and Corpus Linguistics Conference*, Birmingham, UK.


Duran, M. S., Ramisch, C., Aluísio, S. M., and Villavicencio, A. (2011).
Identifying and analyzing Brazilian Portuguese complex predicates.
In [Kordoni et al., 2011], pages 74–82.

References II

Granada, R., Lopes, L., Ramisch, C., Trojahn, C., Vieira, R., and Villavicencio, A. (2012).
A comparable corpus based on aligned multilingual ontologies.
In *Proceedings of the ACL 2012 First Workshop on Multilingual Modeling (MM 2012)*, Jeju, Republic of Korea. Association for Computational Linguistics.

Kordoni, V., Ramisch, C., and Villavicencio, A., editors (2011).
Proc. of the ACL Workshop on MWEs: from Parsing and Generation to the Real World (MWE 2011), Portland, OR, USA. ACL.

Linardaki, E., Ramisch, C., Villavicencio, A., and Fotopoulou, A. (2010).
Towards the construction of language resources for Greek multiword expressions: Extraction and evaluation.
In Piperidis, S., Slavcheva, M., and Vertan, C., editors, *Proc. of the LREC Workshop on Exploitation of multilingual resources and tools for Central and (South) Eastern European Languages*, pages 31–40, Valetta, Malta. May.

Mangeot, M. and Ramisch, C. (2012).
A serious lexical game for building a Portuguese lexical-semantic network.
In *Proceedings of the ACL 2012 3rd Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources and their Applications to NLP*, Jeju, Republic of Korea. Association for Computational Linguistics.

References III


Ramisch, C. (2012a).
A generic framework for multiword expressions treatment: from acquisition to applications.
In [con, 2012], pages 61–66.


Ramisch, C. (2012b).
Une plate-forme générique et ouverte pour le traitement des expressions polylexicales.
In Molina Mejia, J. M. and Schwab, D., editors, *Actes de 14e Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL 2012)*, Grenoble, France.


Ramisch, C., Araujo, V. D., and Villavicencio, A. (2012).
A broad evaluation of techniques for automatic acquisition of multiword expressions.
In [con, 2012], pages 1–6.


Ramisch, C., Besacier, L., and Kobzar, O. (2013).
How hard is it to automatically translate phrasal verbs from English to French?
In Mitkov, R., Monti, J., Pastor, G. C., and Seretan, V., editors, *Proc. of the MT Summit 2013 MUMTTT workshop (MUMTTT 2013)*, pages 53–61, Nice, France.

References IV


Ramisch, C., de Medeiros Caseli, H., Villavicencio, A., Machado, A., and Finatto, M. J. (2010a).
A hybrid approach for multiword expression identification.
In *Proc. of the 9th PROPOR (PROPOR 2010)*, volume 6001 of *LNCS (LNAI)*, pages 65–74, Porto Alegre, RS, Brazil. Springer.


Ramisch, C., Villavicencio, A., and Boitet, C. (2010b).
Multiword expressions in the wild? the mwetoolkit comes in handy.
In Liu, Y. and Liu, T., editors, *Proc. of the 23rd COLING (COLING 2010) — Demonstrations*, pages 57–60, Beijing, China. The Coling 2010 Organizing Committee.


Ramisch, C., Villavicencio, A., and Boitet, C. (2010c).
mwetoolkit: a framework for multiword expression identification.
In *Proc. of the Seventh LREC (LREC 2010)*, pages 662–669, Valetta, Malta. ELRA.


Ramisch, C., Villavicencio, A., and Boitet, C. (2010d).
Web-based and combined language models: a case study on noun compound identification.
In Huang, C.-R. and Jurafsky, D., editors, *Proc. of the 23rd COLING (COLING 2010) — Posters*, pages 1041–1049, Beijing, China. The Coling 2010 Organizing Committee.

References V

Villavicencio, A., Idiart, M., Ramisch, C., Araujo, V. D., Yankama, B., and Berwick, R. (2012).
Get out but don't fall down: verb-particle constructions in child language.
In Berwick, R., Korhonen, A., Poibeau, T., and Villavicencio, A., editors, *Proc. of the EACL 2012 Workshop on Computational Models of Language Acquisition and Loss*, pages 43–50, Avignon, France. ACL.

Villavicencio, A., Ramisch, C., Machado, A., de Medeiros Caseli, H., and Finatto, M. J. (2010).
Identificação de expressões multipalavra em domínios específicos.
Linguamática, 2(1):15–33.