

# IC1207 COST Action PARSEME

## PARSing and Multi-word Expressions

Towards linguistic precision and computational efficiency  
in natural language processing

1st PARSEME Training School in Prague  
19 January 2015  
Prague, Czech Republic

# IC1207 COST Action: **PARSEME**



scientific network  
30 COST countries  
2 non-COST institutions  
meetings, short-time missions,  
workshops, training schools

## Duration

4 years: 8 March 2013 – 7 March 2017

# People

- **200 members** (official and unofficial),
- **29 languages** from 9 language families,
- multidisciplinary experts: linguists, computational linguists, computer scientists, psycholinguists, industrials, . . . ,
- grammatical frameworks: CCG, DG, GG, HPSG, LFG, TAG, . . . ,
- methodological trends: symbolic, statistical, hybrid,
- early-stage researchers ( $< PhD + 8$ ): **56%**.

## Participation rules

- **All relevant researchers** from all ▶ member countries are **eligible** for funding.
- The Management Committee selects those candidates who will be **entitled** to funding.

# WG4: Annotating MWEs in Treebanks

## Topics

- **MWE-aware** methodologies of **treebank construction**,
- Optimal usability of **MWEs in parsing**.

## WG leader and vice-leader

**Victoria Rosén**, Bergen, Norway

**Petya Osenova**, Sofia, Bulgaria



# WG3: Statistical, Hybrid and Multilingual Processing of MWEs

## Topics

- Combining symbolic and statistical methods into **hybrid models**,
- **Efficiency** and **accuracy** of hybrid parsing methods,
- Applying hybrid models to **MWE processing**,
- Use of **unannotated data** to improve methods based on annotated data,
- Hybrid processing methods and **multilingual applications**.

## WG leader and vice-leader

**Michael Rosner**, Msida, Malta

**Matthieu Constant**, Marne-la-Vallée, France



# WG2: Parsing Techniques for MWEs

## (Symbolic parsing and MWEs)

### Topics

- Potential of different **linguistic frameworks** (CCG, DG, GG, HPSG, LFG, TAG, ...) with respect to parsing MWEs,
- Enhancing parsing **efficiency**,
- Reducing the **cost of grammar production**.

### WG leader and vice-leader

**Yannick Parmentier**, Orléans, France

**Jakub Waszczuk**, Blois, France



# WG1: Lexicon/Grammar Interface

## Objectives

- **Linguistic properties of MWEs**, in particular at the lexical and syntactic level,
- Usability of MWE **lexicons** and valence dictionaries **in parsing**,
- **Interoperability** of lexicons and the reduction of their **production cost**.

## WG leader and vice-leader

**Manfred Sailer**, Frankfurt am Main, Germany

**Gyri Smørdal Losnegaard**, Bergen, Norway



## Homepage

### Upcoming events:

- 1st Training School, 19-23 January 2015, Prague, Czech Republic
- 4th general meeting, 19-20 March 2015, Valletta, Malta
- MUMTTT workshop at EUROPHRAS, 1-2 July 2015, Málaga, Spain
- 5th general meeting, 23-24 September 2015, Iasi, Romania

## PARSEME (PARSING and Multi-word Expressions)

### Towards linguistic precision and computational efficiency in natural language processing



The IC1207 COST Action, **PARSEME**, is an interdisciplinary scientific network devoted to the role of **multi-word expressions (MWEs)** in  **parsing**.

- It gathers interdisciplinary experts (linguists, computational linguists, computer scientists, psycholinguists, and industrialists) from 30 countries who have signed the Memorandum of Understanding.
- It represents 29 languages and 6 dialects from 10 language families<sup>1</sup>.
- It covers different parsing frameworks: CCG (Combinatory Categorical Grammar), DG (Dependency Grammar), GG (Generative Grammar), HPSG (Head-driven Phrase Structure Grammar), LFG (Lexical Functional Grammar), TAG (Tree Adjoining Grammar), ...
- It addresses different methodologies (symbolic, probabilistic and hybrid parsing) and language technology applications (machine translation, information retrieval, ...).

Duration: **8 March 2013 – 7 March 2017**

PARSEME is structured in 4 Working Groups. Its main activities include scientific meetings, short-term scientific missions, workshops and training schools. It promotes gender balance and it adheres to COST policy in promoting early-stage researchers.

Its target groups are the Action's members, early-stage researchers, Master and PhD students, related Actions, projects and networks, other NLP researchers, language industry professionals, language resource providers, European institutions, foreign language teachers and learners, as well as the general public.

PARSEME welcomes **new members** from any member countries or from new COST countries.

### Latest Article

- MUMTTT Workshop at EUROPHRAS-15
- Book project: Multiword Expressions: Insights from a Multi-lingual Perspective
- 3rd call for meeting organizers
- 2nd call for meeting organizers
- MC meeting, 10 March 2014, Athens, Greece

### Events

MUMTTT Workshop at EUROPHRAS-15  
5th general meeting, 23-24 September 2015, Iasi, Romania  
4th general meeting, 19-20 March 2015, Valletta, Malta  
Frankfurt Workshop on MWEs, 8-9 September 2014, Frankfurt, Germany  
3rd general meeting, 8-10 September 2014, Haifa, Israel (relocated to Frankfurt)

### News

3rd call for meeting organizers  
2nd call for meeting organizers  
MWE Workshop accepted to EACL-2014  
1st call for meeting organizers  
Website online



# WG members list

PARSEME

[Submit](#)[Contact](#)[Sign In](#)

This list contains the members of [PARSEME Working Groups](#). To submit a new application for a membership, click on [Submit](#). The list of the [Management Committee members](#) is available in the official [COST](#) pages of the action.

Member	Country	Affiliation	Languages	ESR	WG1	WG2	WG3	WG4
<a href="#">Ahmad Aghaebrahimian</a>	Czech Republic	Charles university, Faculty of Mathematics and physics, Institute of formal and applied linguistics	English	✓		✓	✓	
<a href="#">Zeljko Agic</a>	Croatia	University of Zagreb		✓			✓	✓
<a href="#">Mehmet AKTAS</a>	Turkey	Yildiz Technical University	English			✓		
<a href="#">Marco Angster</a>	Italy	Free University of Bolzano/Bozen	Italian, German, languages of Europe (cross-linguistically)	✓	✓			
<a href="#">Giuseppe Attardi</a>	Italy	università di Pisa				✓		
<a href="#">Sascha Bargmann</a>	Germany	University of Frankfurt	English, German	✓	✓	✓		
<a href="#">Mia Batinić</a>	Croatia	University of Zadar	Croatian, Italian	✓	✓			
<a href="#">Eduard Bejcek</a>	Czech Republic	Charles University in Prague, Institute of Formal and Applied Linguistics	Czech	✓	✓			✓
<a href="#">Matea Birić</a>	Croatia	Institute of Croatian Language and Linguistics	Croatian		✓			
<a href="#">Philippe Blache</a>	France	LPL, CNRS & Aix-Marseille Université				✓		
. . .								
<a href="#">Jan Šnajder</a>	Croatia	University of Zagreb, Faculty of Electrical Engineering and Computing	Croatian	✓	✓	✓		
<b>Total:</b>	<b>146</b>	<b>30</b>		<b>82</b>	<b>89</b>	<b>53</b>	<b>53</b>	<b>45</b>

# We welcome new members

## PARSEME

[Home](#) [The Action](#) [Organization](#) [Participants](#) [Events](#) [STSM Grants](#) [Related Links](#) [Downloads](#) [Contact](#) [Publications](#) [Search](#)

### How to join us

How to join us

Mailing Lists

*"MY PROFESSIONAL INTERESTS ARE RELEVANT TO PARSING AND MULTI-WORD EXPRESSIONS. I HAVE HEARD ABOUT PARSEME. HOW CAN I JOIN THE ACTION?"*




The principle of any COST action is to be open to new members. Here is the lightweight admission procedure:

1. Check if your **country** is an **official member of PARSEME**.
2. If your country **is not** an official member, contact your [COST National Coordinator](#). He/she will be in charge of signing the action's [Memorandum of Understanding](#) on behalf of your country. You may become its official representative at the [Management Committee](#). If your country joins the action after November 2013 it will need a formal agreement from the Management Committee.
3. If your country already **is** an official member, read the scientific program of the action (see the [Memorandum of Understanding](#), pp. 12-17) and the description of the [Working Groups](#) (WGs). Choose one or more Working Groups which you would like to participate in.
4. Fill in the [application form](#) with the following data:
  - name,
  - country,
  - affiliation,
  - personal webpage address (if any),
  - Early Stage Researcher status (are you a PhD student or have you received your PhD later than 6 years before the beginning of your involvement in the action?),
  - female/male status (for the sake of gender balance reporting),
  - languages under study,
  - numbers of the Working Groups which you would like to join, a short scientific **statement of interest** (up to **2500 characters**) describing your previous and planned contributions to WG-related topics.

Your application will be evaluated by the Steering Committee. If you become a member, note that all the above data, for networking purposes, will appear on the [list of participants](#) and the member's emails will be added to the [WG mailing list](#) (additionally to the parseme-all list).

# Mailing lists

@chopin.ipipan.waw.pl

List	Members	Admin	
<b>parseme-all</b>	all members (> 200)	Adam P.	Subscribing confirmation required
<b>parseme-mc</b>	MC members and substitutes, official emails		
<b>parseme-steer</b>	SC members		
<b>parseme-wg1</b>	WG members	WG leaders	
<b>parseme-wg2</b>			
<b>parseme-wg3</b>			
<b>parseme-wg4</b>			
<b>parseme-esr</b>	ESR members	ESR repr.	

## Future events

4th general meeting

University of **Malta**  
**19-20 March 2015**



## Future events

### 4th general meeting

University of **Malta**  
19-20 March 2015



### MUMTTT workshop

► Multi-word Units in Machine Translation and Translation Technology

Co-located with the **EUROPHRAS 2015** conference  
Universidad de **Málaga**  
1-2 July 2015



## Future events

### 4th general meeting

University of **Malta**  
19-20 March 2015



### MUMTTT workshop

► Multi-word Units in Machine Translation and Translation Technology

Co-located with the **EUROPHRAS 2015** conference  
Universidad de **Málaga**  
1-2 July 2015



### 5th general meeting

University of **Iasi**, Romania  
23-24 September 2015

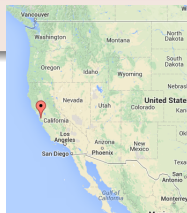
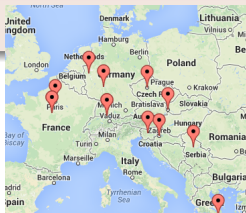


# Short-Term Scientific Missions

## Open call for STSMs

- Duration: 1 week – 3 months (6 months for ESRs)
- From one country of the network to another
- Maximum funding: 2500€
- Best period (budget-wise): April–May
- STSM coordinator: **Cvetana Krstev**, Belgrade, Serbia

## STSMs in 2013-2014



# Survey on MWE lexicons and treebanks (WG1 & WG4)

## Objective

- List of existing MWE **lexicons/treebanks** and their features

## Methods

- Searching existing infrastructures and lists (CLARIN, META-SHARE, <http://multiword.sourceforge.net>, ... )
- ▶ **Public webform** (contributions still welcome)
- Minimal responses per email ([parseme-survey@chopin.ipipan.waw.pl](mailto:parseme-survey@chopin.ipipan.waw.pl))

## Results

- Available in a ▶ **public table**: 84 resources and tools

## People

**Gyri Smørdal Losnegaard & Federico Sangati**



# MWE templates for particular languages (WG1)

## Objective

- Develop a **cross-language classification** of MWEs
- Point at universal and language-specific properties of MWE

## Method

- [Wiki space](#) with one page per language:
  - Fixedness/flexibility of MWE parts (NP, PP, VP, AP, ...)
  - MWEs by syntactic structure (nominal, verbal, ...)
  - MWEs by idiomaticity (lexical, syntactic, semantic, pragmatic, statistical idiomaticity)

## People

**Manfred Sailer** and one/two **contributors** per language

# Book project (WG1)

## Basic data

- Title: *Mutliword Expressions: Insights from a Multi-lingual Perspective*
- Deadline for 2-page proposals: 10 January 2015
- Discussions on selected proposals in **Malta**, 19-20 March 2015

## Topics (see ▶ call )

- MWE classification and tests for them
- Special types of MWEs
- Cross-linguistic comparison of MWE types

## Coordinators

**Manfred Sailer** and **Stella Markantonatou**

# Survey on hybrid processing of MWEs (WG3)

## Objective

- Classification scheme for MWE processing models
- 3 SOA surveys of existing MWE processing methods and their classification in the scheme
  - discovery
  - translation
  - parsing

## People

**Mike Rosner & Matthieu Constant**

# Classification scheme

<b>Statistical</b>	<b>MWE Resource Creation</b>	<b>MWE Resource Incorporation</b>
<b>Monolingual</b>	Lexicons	Parsing/Simplification Models
<b>Multilingual</b>	Multilingual lexicon	Parsing/Translation/Generation Models

<b>Symbolic</b>	<b>MWE Resource Creation</b>	<b>MWE Resource Incorporation</b>
<b>Monolingual</b>	Lexicons	Parsing/Simplification Models
<b>Multilingual</b>	Multilingual lexicon	Parsing/Translation/Generation Models

- x: role of MWE resources
- y: degree of multilinguality
- z: methodological framework

# Survey on MWE annotation in treebanks (WG4)

iness

Main Page  
Project description  
Participants  
Documentation  
Publications  
Links  
Resources

Treebanks  
Treebank selection  
Parallel Sentences  
Metadata

Tools  
XLE-Web

Text preprocessing  
Word List  
Document  
Variables  
Overview

MWEs in Parseme

Overview of MWE annotation in treebanks

Working Group 4 (WG 4) in Parseme is creating an overview of existing annotation of multiword expressions (MWEs) in treebanks. The table below lists the treebanks that are currently documented or in the process of being documented. Some cells in the table are clickable, leading to wiki pages that provide detailed information about the treebanks and the MWE annotations that they provide. The table has been updated in accordance with the decisions made at the WG 4 sessions at the Frankfurt meeting in September 2014.

If you would like to contribute to this WG 4 subproject by providing MWE information about a treebank that is not already listed in the table, please contact the working group leader [Victoria Rosén](#).

The old table is still available [here](#) if you need to consult it.

Content

Overview of MWE annotation in treebanks

Instructions for providing treebank information

1 The wiki system

1.1 Editing rights

1.2 Saving

2 Editing the table

2.1 The treebank column

2.2 The MWE columns

3. The treebank description page

4. The MWE pages

4.1 Example

4.2 Analysis

4.3 About the analysis

4.4 Searching for the MWE type

Treebank	Language	Nominal MWEs			Verbal MWEs				Prepositional MWEs	Adjectival MWEs	MWEs of other categories	Proverbs
		Named entities	NN compounds	Other nominal MWEs	Phrasal verbs	Light verb constructions	VP Idioms	Other verbal MWEs				
<a href="#">The INESS Norwegian Treebank</a>	Norwegian	YES	N/A	YES	YES	YES	YES	YES	YES	YES	YES	NO
<a href="#">The Alpino Treebank</a>	Dutch	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
<a href="#">DeepBank</a>	English	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
<a href="#">META-NORD</a>												

Sign In: [Local](#) | [eduGAIN](#) | [OpenID](#) | [Clarín IDP](#)

19/27

# Training School programme

<b>Manfred Sailer</b>	Frankfurt	Linguistic properties of MWEs
<b>Shuly Wintner</b>	Haifa	MWEs in linguistic theory (EN/DE/FR) Challenges from other languages (Hebrew) Encoding and applications
<b>Dan Flickinger</b>	Stanford	Introduction to HPSG and challenges from MWEs
<b>Joakim Nivre</b>	Uppsala	Introduction to dependency grammar and parsing Graph-based and transition-based dependency parsing MWEs in dependency parsing Practical lab with MaltParser
<b>Jan Hajič</b> <b>Pavel Straňák</b> <b>Jiří Mírovský</b>	Prague	Treebanking and MWEs

# Prague schedule

	Monday 19 January	Tuesday 20 January	Wednesday 21 January
9:00-9:30	registration	(session 5) <b>Manfred Sailer &amp; Shuly Wintner</b> (2) "MWEs in linguistic theory (based on English/German/French)"	(session 9) <b>Pavel Straňák and Jiří Mírovský</b> (3) "Treebanking and MWEs" (practical lab)
9:30-10:30	(session 1) opening session		
10:30-11:00	coffee break	coffee break	coffee break
11:00-12:30	(session 2) <b>Jan Hajič, Pavel Straňák and Jiří Mírovský</b> (1) "Treebanking and MWEs"	(session 6) <b>Dan Flickinger</b> (2) "Introduction to HPSG and challenges from MWEs"	(session 10) <b>Joakim Nivre</b> (2) "Graph-based and transition-based dependency parsing"
12:30-13:00	lunch	lunch	lunch
13:00-14:00			
14:00-15:30	(session 3) <b>Dan Flickinger</b> (1) "Introduction to HPSG and challenges from MWEs"	(session 7) <b>Pavel Straňák and Jiří Mírovský</b> (2) "Treebanking and MWEs"	(session 11) <b>Manfred Sailer &amp; Shuly Wintner</b> (3) "Challenges from other languages (Hebrew)"
15:30-16:00	coffee break	coffee break	coffee break
16:00-17:30	(session 4) <b>Manfred Sailer &amp; Shuly Wintner</b> (1) "Overview: linguistic properties of MWEs"	(session 8) <b>Joakim Nivre</b> (1) "Introduction to dependency grammar and dependency parsing"	(session 12) <b>Dan Flickinger</b> (3) "Introduction to HPSG and challenges from MWEs"
19:30	Dinner for trainers and local organizers		

# Prague schedule

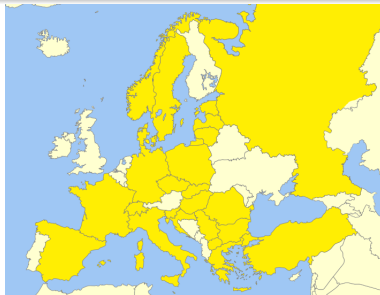
Thursday 22 January	Friday 23 January
(session 13) <b>Dan Flickinger</b> (4) "Introduction to HPSG and challenges from MWEs" (practical lab)	(session 17) <b>Joakim Nivre</b> (4) "MaltParser" (practical lab)
coffee break	coffee break
(session 14) <b>Manfred Sailer &amp; Shuly Wintner</b> (4) "Encoding in a computational lexicon; application to the participants' languages; summary"	(session 18) Cross-module session.
lunch	(session 19) closing session
(session 15) <b>Joakim Nivre</b> (3) "Multiword expressions in dependency parsing"	lunch
coffee break	free sightseeing
(session 16) <b>Pavel Straňák and Jiří Mírovský</b> (4) "Treebanking and MWEs" (practical lab)	



# Trainees

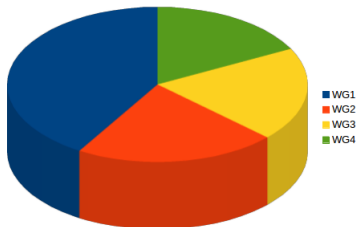
50 trainees from 25 countries

- **103** candidates from **35** countries,
- **43** candidates selected for **funding** on the basis of
  - previous activity in PARSEME,
  - recommendations by the MC representatives,
- **9 self-funded** candidates.

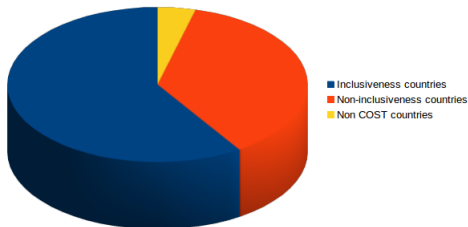


# Participation

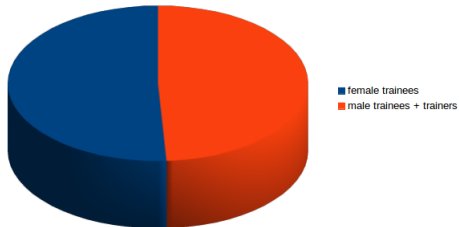
Trainees per WG



Trainees per country type



Gender balance of participants



# Organization matters

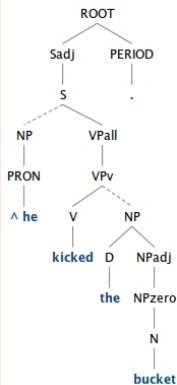
- sign the **attendance list** each day (otherwise no **reimbursement**),
- **lunches** and coffee breaks are funded for all participants,
- **trainers** submit a reimbursement claim via e-COST (as after a meeting),
- funded **trainees** are reimbursed directly (provided that they have signed the list),
- only **funded ESRs** are admitted to **labs** (sessions 9, 13, 16 and 17).

## Round table

# Questions?

Thank you

## C-structure



## F-structure

