

IC1207 COST Action PARSEME

PARSIng and Multi-word Expressions

Towards linguistic precision and computational efficiency
in natural language processing

PARSEME scientific meeting
Plenary session
16 September 2013
ICS PAS, Warsaw, Poland

COST



■ The 27 EU Member States

■ EU Acceding & Candidate Countries

- ▶ Croatia
- ▶ Former Yugoslav Republic of Macedonia
- ▶ Iceland
- ▶ Turkey

Other Countries

- ▶ Bosnia and Herzegovina
- ▶ Republic of Serbia
- ▶ Norway
- ▶ Switzerland

COST Cooperating States

- ▶ Israel

- inter-governmental framework (founded in 1971),
- coordination of nationally-funded European research,
- funded by FP7.

What Is a COST Action?

- **bottom-up approach**: the scientific challenges are defined by the researchers,
- objective: to overcome the research **fragmentation** issues,
- COST supports **cooperation** and **dissemination**: meetings, workshops, short-term missions, training schools,
- **no direct research funding**,
- precursor role for **other European programmes**,
- important roles given to **Early-Stage Researchers** (< PhD+8),
- **gender balance** promoted,
- budget: **129,000–156,000 euros** per year for all partners,
- proposal **selectivity**: 6%.

IC1207 COST Action: **PARSEME**

Duration

4 years: 8 March 2013 – 7 March 2017

IC1207 COST Action: PARSEME

Duration

4 years: 8 March 2013 – 7 March 2017

Objectives (cf. Memorandum of Understanding)

- **Outreach:** to put **multilingualism** in focus of linguistic and technological studies.
- **Networking:** to establish a long-lasting **cross-lingual**, **cross-theoretical** and **cross-methodological** NLP research network.
- **Scientific:** to bridge the gap between **linguistic precision** and **computational efficiency** in NLP applications.

Key problem

Multi-Word Expressions

The **prime time** speech by **first lady Michelle Obama** **set** the house **on fire**. She made **crystal clear** which issues she **took to heart**, but she was **preaching to the choir**.

Key problem

Multi-Word Expressions

The **prime time** speech by **first lady Michelle Obama** **set** the house **on fire**. She made **crystal clear** which issues she **took to heart**, but she was **preaching to the choir**.

Facts

- MWEs are prevalent (40% of text items),
- MWEs are complex phenomena involving different levels of language (lexicon, syntax, meaning ...),
- MWEs are still not sufficiently understood,
- MWEs are under-represented in language resources and tools,
- MWEs are hard to detect, understand, translate, etc.

Consortium

- 103 members (official and unofficial),
- 27 COST countries, 3 experts from non-COST countries (USA, Brazil),
- multidisciplinary experts: linguists, computational linguists, computer scientists, psycholinguists, industrials, . . . ,
- different linguistic 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), . . .
- two methodological trends:
 - knowledge-based,
 - data-driven.

Languages

- **26 languages** from 9 language families:
 - **Celtic:** Gaelic,
 - **Germanic:** English, Danish, Dutch, German, Icelandic, Norwegian, Swedish,
 - **Finno-Ugric:** Estonian, Hungarian,
 - **Hellenic:** Greek,
 - **Romance:** French, Italian, Portuguese, Spanish,
 - **Semitic:** Hebrew, Maltese,
 - **Slavic:** Bulgarian, Croatian, Czech, Polish, Serbian, Slovak, Slovenian, Macedonian,
 - **Turkic:** Turkish.
- 6 dialects:
 - British vs. American **English**,
 - Swiss vs. France **French**,
 - European vs. Brazilian **Portuguese**.

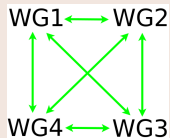
Working Groups

WG1: Lexicon/grammar interface,

WG2: Parsing techniques for MWEs,

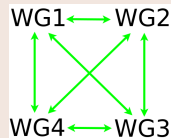
WG3: Hybrid parsing of MWEs,

WG4: Annotating MWEs in treebanks.



Working Groups

- WG1:** Lexicon/grammar interface,
WG2: Parsing techniques for MWEs,
WG3: Hybrid parsing of MWEs,
WG4: Annotating MWEs in treebanks.



Crossing barriers between ...

- different levels of linguistic processing,
- different linguistic frameworks,
- different methodological frameworks.

WG1: Lexicon/Grammar Interface

Challenges

- Simultaneously account for the **fixed** character of MWEs and their similarities to **regular syntactic structures**.
- Represent parsing phenomena at the lexicon level (**agreement**, **discontinuity** and **free word order**)?
- Enrich existing **lexicons and valence dictionaries** with MWEs.
- Design **cost-saving abstract models** of MWEs' properties, automatically **mapped** to different grammar formalisms.

WG leader

Manfred Sailer
Frankfurt, Germany



WG2: Parsing Techniques for MWEs

Challenges

- Design **interoperable MWE representation** for different syntactic frameworks: **CCG, DG, GG, HPSG, LFG, TAG,**
- Reduce the **cost of grammar production**.
- Enhance **parsing speed and precision** by reducing spurious ambiguity in MWEs.
- Express the **semantics of MWEs** in parse structures.

WG leader

Yannick Parmentier
Orléans, France



WG3: Hybrid Parsing of MWEs

Challenges

- Cope with **long-distance relations and discontinuities** in probabilistic parsing.
- Integrate high-quality **language resources in probabilistic parsing** (MWE-oriented reranking of state-of-the-art parsers).
- Enhance **knowledge-based parsing** of MWEs with **probabilistic scores**.
- Enhance **supervised methods** (using scarce annotated corpora) with **unsupervised** ones (using unannotated corpora).

WG leader

Michael Rosner
Msida, Malta



WG4: Annotating MWEs in Treebanks

Challenges

- **Annotation guidelines** for representing MWEs in constituency and dependency treebanks.
- Best practices for automatically **extracting lexicons and probability scores** addressed in other WGs.

WG leader

Victoria Rosén
Bergen, Norway



Managment

Management Committee (MC)

- up to **2 members** and **2 substitutes** per party (country).

Management

Management Committee (MC)

- up to **2 members** and **2 substitutes** per party (country).

Grant Holder

- Institute of Computer Science, Polish Academy of Sciences (**IPIPAN**), Warsaw, Poland,
- Scientific Representative: **Adam Przepiórkowski**,
- Secretary: **Beata Wójtowicz**
- roles: reimbursement, reporting, secretariat, coordination, ...



Financial Rapporteurs

- **Shuly Wintner** (Israel) and **Michael Rosner** (Malta),
- roles: expenditure verification.



Steering Committee (SC)

- Chair: **Agata Savary** (France)
- Vice-Chair: **Adam Przepiórkowski** (Poland),
- WG Leaders (to be validated by WGs):
 - WG1: **Manfred Sailer** (Germany)
 - WG2: **Yannick Parmentier** (France)
 - WG3: **Michael Rosner** (Malta)
 - WG4: **Victoria Rosén** (Norway)
- WG **Vice-Leaders** to be appointed
- Representative of Early-Stage Researchers (ESRs): **Pavel Straňák** (Czech Republic), subst. **Veronika Vincze** (Hungary)
- Coordinator of Short Term Scientific Missions (STSMs): **Cvetana Krstev** (Serbia)
- Dissemination Coordinator: **Miriam Butt** (Germany)



Budget – Year 1

Budget year 1

1 June 2013 – 31 May 2014

Allocated budget

161,400€

Budget – Year 1

Budget year 1

1 June 2013 – 31 May 2014

Allocated budget

161,400€

Travel reimbursement rules

- receive a formal invitation via e-COST,
- pay your travel first, get reimbursed on return,
- researchers from all member countries are **eligible** for reimbursement,
- the MC chair selects those **entitled** to reimbursement.

1st year objectives

Outreach

- new countries & members,
- external communication means.

1st year objectives

Outreach

- new countries & members,
- external communication means.

Networking

- getting to know each other,
- structuring the community around the Working Groups,
- **workplan** of each WG,
- internal communication means.

1st year objectives

Scientific objectives

- better understanding of:
 - linguistic properties of MWEs,
 - the potential of frameworks and methodologies wrt. parsing MWEs,
 - challenges behind the MWE annotation in treebanks,
- studies towards:
 - enhanced usability of MWE lexicons and valence dictionaries in parsing,
 - a better coverage of MWEs in LRT,
 - evaluation capacity for MWE parsing.

1st year outcomes

- contrastive **state-of-the-art surveys** in all WGs,
- detailed **scientific program** of each WG,
- **website**,
- **mailing lists** and internal website,
- workshop proceedings,
- publications & technical documents,
- annual report.

Joining a Working Group

Lightweight procedure


- Required data:
 - name, country, affiliation, personal webpage address, female/male status,
 - **ESR status** (are you a PhD student or have you received your PhD later than 7 March 2005?),
 - a short scientific **statement of interest** (up to 1/2 page) describing your previous and planned contributions to WG-related topics.
- Send the data to all relevant **WG leaders**.
- The data will appear on the **website**.
- Your email is added to the **WG mailing list**.
- WG membership influences the **reimbursement** policy.

Mailing lists

@chopin.ipipan.waw.pl

List	Members	Admin
parseme-all	all members (94)	
parseme-mc	MC members and substitutes, official emails	Adam P.
parseme-steer	SC members	
parseme-wg1		WG leaders
parseme-wg2	WG members	
parseme-wg3		
parseme-wg4		

Subscribing confirmation required



Websites

The screenshot shows the COST website's 'Domains and Actions' page for the ICT domain. It features a navigation menu with 'About COST', 'Domains and Actions', 'Participate', 'Events', and 'Media'. A sidebar on the left lists various scientific domains. The main content area is titled 'ICT COST Action IC1207' and includes a table of participating countries with columns for 'Country', 'Date', and 'Status'. A 'Participations' table lists countries like Bulgaria, Croatia, Czech Republic, Denmark, Finland, and France with their respective dates and confirmed status.

Country	Date	Status
Bulgaria	06/03/11	Confirmed
Croatia	29/03/11	Confirmed
Czech Republic	26/03/11	Confirmed
Denmark	19/03/11	Confirmed
Finland	04/10/11	Confirmed
France	17/03/11	Confirmed



The screenshot shows the PARSEME COST Action website. It features a navigation menu with 'Home', 'The Action', 'Organization', 'Participants', 'Events', 'STM Grants', 'Related Links', 'Downloads', 'Contact', and 'Publications'. The main content area is titled 'The PARSEME COST Action' and includes a search bar. The text describes the action's goal: 'PARSEME: PARSING and Multi-word Expressions. Towards linguistic precision and computational efficiency in natural language processing'. It mentions that the action aims to increase and enhance the support of the European multilingual heritage from Information and Communication Technologies (ICT). A 'Latest Article' sidebar on the right lists a kick-off meeting in Brussels (8 March 2011) and a workshop on Multi-Word Expressions (MWEs).

COST pages

- http://www.cost.eu/domains_actions/ict/Actions/IC1207
- abstract, countries, MC, MoU

Website (under development)

- www.parseme.eu
- webmaster: **Maïke Müller** (Konstanz)

Short-Term Scientific Missions

- Duration: 1 week – 3 months (6 months for ESRs)
- Maximum funding: 500€ (travel), 160€ (daily allowance),
- Funding limit: 2500€ per mission,
- Priority to ESRs,
- Current submissions: 2,
- Available budget: >9 STSMs,
- Open **call for STSM proposals**,
- Application deadline: 10 January (of before),
- STSM coordinator: **Cvetana Krstev**.



Future events (to be confirmed)

Next general meeting

Institute for Language and Speech Processing,
Athena Research Center,
Athens, Greece
10-11 March 2014

Future events (to be confirmed)

Next general meeting

Institute for Language and Speech Processing,
Athena Research Center,

Athens, Greece
10-11 March 2014

MWE Workshop

- Endorsed by **EACL 2014 in Gothenburg, Sweden,**
- **26-27 April 2014,**
- Submission deadline: **23 January, 2014,**
- Special track on parsing and MWEs,
- Reimbursement: 20 participants (esp. ESRs), 1-2 invited speakers.

Future events – Year 2

Training School

- July 2014,
- call for organizer,
- Local Organizer Support available.

Future events – Year 2

Training School

- July 2014,
- call for organizer,
- Local Organizer Support available.

General meetings 3-4

- autumn 2014,
- spring 2015,
- call for organizers,
- Local Organizer Support available.

Warsaw meeting

Scientific content

- scientific presentations,
- validating WG leaders,
- WG discussions.

Warsaw meeting

Scientific content

- scientific presentations,
- validating WG leaders,
- WG discussions.

Outcome

- scientific **workplan** for years 1–2 (and more), for each WG,
- main **input**: your current research,
- **added value**: overcoming fragmentation (new collaborations, sharing knowledge, sharing LRTs, ...).

Warsaw meeting schedule

Monday

14:00-15:00	plenary session
15:00-18:00	WG2 meeting
19:00-	dinner

Warsaw meeting schedule

Tuesday

9:30-13:00	WG1 meeting
13:00-14:30	lunch
14:30-17:30	WG3 meeting
17:30-18:30	MC meeting

Wednesday

9:30-13:00	WG4 meeting
13:00-14:30	lunch
14:30-15:30	plenary session (summary)
15:30-16:30	SC meeting

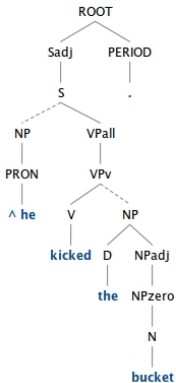
Organization matters

- dinner tonight,
- bring your **badges** to the dinner and to lunches,
- ...

Questions?

Thank you

C-structure



F-structure

