PARSEME

PARSing and Multi-word Expressions

Towards linguistic precision and computational efficiency in natural language processing

Yannick Parmentier,
Agnieszka Patejuk,
Adam Przepiórkowski,
Agata Savary,
and PARSEME partners

8 March 2013

Grant Holder

Tasks

- financial reporting,
- scientific and administrative secretariat,
- coordination, liaison,
- publication, dissemination

See COST Vademecum (Part B) — Grant System

Grant Holder

Tasks

- financial reporting,
- scientific and administrative secretariat,
- coordination, liaison,
- publication, dissemination

See COST Vademecum (Part B) — Grant System

Candidate

- Institute of Computer Science, Polish Academy of Sciences (**IPIPAN**), Warsaw, Poland,
- Legal Representative: **prof. Jacek Koronacki**,
- Finance Officer: **Bogusław Martyniak**,
- Scientific Representative: **prof. Adam Przepiórkowski**.

Financial Rapporteurs

Tasks

For each Grant Period:

- verify expenditures,
- provide a financial assessment.

Conflict of interests shall be avoided between the MC Chair, the 2 Financial Rapporteurs and the Grant Holder.

Financial Rapporteurs

Tasks

For each Grant Period:

- verify expenditures,
- provide a financial assessment.

Conflict of interests shall be avoided between the MC Chair, the 2 Financial Rapporteurs and the Grant Holder.

Candidates

- **???**
- **???**

Objectives

General aim

Increasing and enhancing the ICT support of the **European multilingual heritage**.

Objectives

General aim

Increasing and enhancing the ICT support of the **European multilingual heritage**.

More detailed objectives

- crossing language barriers,
- enhancing language representativeness,
- reinforcing interactions between theories and methodologies,
- bridging the gap between linguistic precision and computation efficiency in NLP application.

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

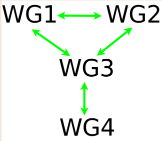
- 75 members (official and unofficial),
- 25 COST countries (20 + Belgium, Bulgaria, Greece, Ireland, Turkey),
- 3 experts from 2 non-COST countries (USA, Brazil),
- multidisciplinary experts: linguists, computational linguists, computer scientists, psycholinguists, industrials, . . . ,
- different linguistic frameworks:
  - **CCG** (Combinatory Categorial Grammar),
  - **DG** (Dependancy Grammar),
  - **HPSG** (Head-driven Phrase Structure Grammar),
  - **LFG** (Lexical Functional Grammar),
  - **TAG** (Tree Adjoining Grammar), . . .
- two methodological trends:
  - knowledge-based,
  - data-driven.

Languages

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

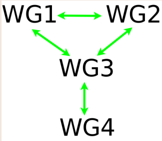
Working Groups

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.

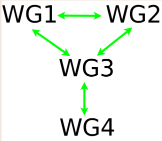


Crossing barriers between . . .

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

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.

Expression of interest in at least 2 WGs from each member (at the full proposal period).

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.

WG2: Parsing Techniques for MWEs

Challenges

- Design **interoperable** MWE **representation** for different syntactic frameworks: **HPSG**, **LFG**, **TAG**, **CCG**, **DG**, ...
- 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.

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).

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.

Bodies and Roles

Management Committee

- up to 2 members and 2 substitutes per country,
- reporting to the Domain Committee,
- meeting at least once a year in member countries,
- see *Rules and Procedures for Implementing COST Actions*.

Bodies and Roles

Management Committee

- up to 2 members and 2 substitutes per country,
- reporting to the Domain Committee,
- meeting at least once a year in member countries,
- see *Rules and Procedures for Implementing COST Actions*.

Steering Committee

- MC Chair and Vice-Chair,
- WG Leaders,
- Representative of Early-Stage Researchers (ESRs),
- Coordinator of Short Term Scientific Missions (STSMs),
- Dissemination Coordinator,
- 4 meetings per year (possibly by video-conference).

Distribution of Tasks

Steering Committee **candidates**

- MC Chair and Vice-Chair (*see elections*),
- WG Leaders:
  - WG1: **Manfred Sailer** (Germany), . . .
  - WG2: **Yannick Parmentier** (France), . . .
  - WG3: **Michael Rosner** (Malta), **Matthieu Constant** (France), . . .
  - WG4: **Victoria Rosén** (Norway), . . .
- Representative of Early-Stage Researchers (ESRs): **Pavel Straňák** (Czech Republic), . . .
- Coordinator of Short Term Scientific Missions (STSMs): **Cvetana Krstev** (Serbia), . . .
- Dissemination Coordinator: **Miriam Butt** (Serbia), . . .

Early-Stage Researchers (ESRs) and Gender Balance

COST directives

- ESR: $< PhD + 8$,
- see *COST Strategy towards increased support of early stage researchers*,
- support measures: STSMs, training schools, think tank, conference grants, WG Chair nominations, MC nominations,
- COST family friendly policy.

ESRs and Gender Balance in PARSEME

- Full support to the COST directives,
- Gender balance at the proposal stage (% of women):
  - 50% of initiators,
  - 38% of members,
  - 40% of the proposed MC.
- Gender balance now (% of women):
  - 38% of members,
  - 40% MC members.
- ESRs at the proposal stage:
  - 50% of initiators,
  - 23% of members,
  - 30% of the proposed MC.
- ESRs now: ???

Timetable

Activity	Year 1				Year 2				Year 3				Year 4			
MC meeting	✓				✓				✓				✓			✓
Scientific timetable for each WG	✓															
Public website		✓														
Internal website		✓														
WG meeting	✓				✓				✓				✓			✓
SC meeting	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Internal evaluation		✓			✓				✓				✓			
Training Schools						✓								✓		
Action's Workshop					✓				✓				✓			✓
Open Workshop							✓								✓	
STSMs for ESRs		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
STSMs for senior researchers		✓		✓		✓		✓		✓		✓		✓		
Annual report				✓				✓				✓				✓
Final report																✓

1st year objectives

Networking objectives

- getting to know one another,
- structuring the community around the Working Groups,
- planning the activity of each Working Group.

1st year objectives

Networking objectives

- getting to know one another,
- structuring the community around the Working Groups,
- planning the activity of each Working Group.

Scientific objectives

- better understanding of linguistic properties of MWEs (in particular at lexical and syntactic level),
- enhanced usability of MWE lexicons and valence dictionaries in parsing,
- better coverage of MWEs in linguistic resources,
- a better understanding of the potential of different linguistic frameworks with respect to parsing MWEs,
- evaluation capacity for MWE parsing resources and tools.

1st year outcome

- contrastive multilingual analysis of MWEs' properties ($WG1$),
- contrastive analysis of MWE treatment in different linguistic frameworks ($WG2$),
- contrastive analysis of MWE annotation methods ($WG4$),
- extensions of existing resources with MWEs ($WG1 \leftrightarrow WG2$),
- test beds for parsing with MWEs ($WG1 \rightarrow WG2, WG3$),
- establishing workpackages dedicated to impact (WG1–4),
- Action's website (WG1–4),
- annual report,
- technical documents and publications.

1st year budget

Allocated budget

134 K€ for 20 countries, 5–6 K€ more for each new country

Activity	Details	Draft budget
MC meeting	30 MC members*1 day	22,800
WG meetings	4 repr./country*2 days new countries Local Org. Support 3 non-COST experts	75,540
SC meetings	per videoconference	0
3 STSMs for ESRs	3 * 3 months	7,500
2 STSMs for senior res.	2 * 3 months	5,000
public website	by Dissemination Coord.	2,000
internal website	by Dissemination Coord.	
WG scientific timetables	at Warsaw meetings	
internal evaluation	by the SC and the MC	
annual report	by the Grant Holder	
administration	by the Grant Holder	20,160
bank fees	by the Grant Holder	1000
TOTAL		134,000

Next meeting

Proposal

Institute of Computer Science,

Polish Academy of Sciences,

Warsaw, Poland
16-18 September

Any other business?

- call for STSMs proposals,
- non-COST countries (members or occasional experts?),
- WG organisation (with 2-3 WG memberships per person),
- links with the MWE Workshop community,
- . . .

Any other business?

- call for STSMs proposals,
- non-COST countries (members or occasional experts?),
- WG organisation (with 2-3 WG memberships per person),
- links with the MWE Workshop community,
- . . .