ANR Project PARSEME-FR
Syntactic Analysis of Multiword Expressions in French

Carlos Ramisch

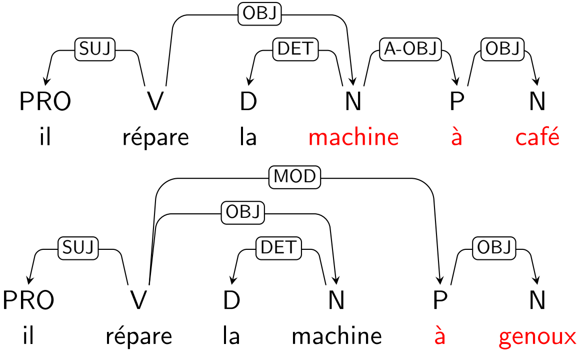
Aix-Marseille Université

3 novembre 2015

Outline

1. Project Overview
2. State of the Art
3. Work Packages

Motivation



Lexico-semantic segmentation ↔ analyse syntaxique

PARSEME

PARSEME

- http://parseme.eu
- European COST Action
- March 2013 – March 2017
- 30 countries, 29 languages and 6 dialects from 10 language families
- 2 yearly workshops, training schools, STSMs
- No money for **actual** research
- 4 Working Groups : state-of-the-art reports, shared task

PARSEME-FR

- National project in France funded by ANR
- Money for **actual** research :-)
- January 2016 – December 2019

Goal

Bridge the gap between linguistic precision and computational efficiency in NLP applications by investigating the syntactic and semantic representation of MWEs in language resources, the integration of MWE analysis in syntactic parsing and its links to semantic processing.

Partners

- Université Paris-Est Marne-la-Vallée, Laboratoire d'informatique Gaspard-Monge (LIGM) [coordinator]
- INRIA, ALPAGE project team
- Université François Rabelais Tours, Laboratoire d'informatique (LI)
- Aix Marseille Université, Laboratoire d'informatique fondamentale (LIF)
- Université d'Orléans, Laboratoire d'informatique fondamentale d'Orléans (LIFO)

Expected Outcomes (1)

A Framework for MWE representation in French language resources

- Fine-grained multidimensional features, symbolic and numerical nature
- Procedures to unify and enrich a lexicon
- Coherence between corpus annotation and lexical entries
- Project the MWE lexicon on any treebank
- Interconnection with symbolic (meta-)grammar

Final Products

1. Guidelines for representing MWEs in linguistic resources
2. Unified and enriched MWE lexicon including linguistic and statistical features
3. Gold standard corpus annotated with all relevant MWE types for French

Expected Outcomes (2)

Comprehensive MWE analysis and syntactico-semantic analysis

- *Orchestration* : where to position MWE analysis in the processing pipeline
  $\Longrightarrow$ Before, during or after parsing ?
- *Algorithms* : adapt parsing algorithms to MWE+syntax representation
- *Semantics* : automatically detect degree of compositionality of MWEs
- *Entity linking* : link MWNE to their pragmatic descriptions in knowledge bases

Final Products

1. MWE-aware surface and deep statistical dependency parsers
2. MWE-aware symbolic parsing environment
3. MWE linker, mapping MWEs to corresponding entries in knowledge bases
4. Web demonstrator integrating final products 1, 2 and 3

Challenges

- Representativeness : constraints, lexicalisation, variability
- Contiguity
- Nesting
- Semantic compositionality
- Lexicon-corpus compatibility
- Unified annotation guidelines
- Orchestration
- Multi-level information integration

Outline

1. Project Overview
2. State of the Art
3. Work Packages

Lexical Resources

- Lexical encoding of MWE properties
  - DELA dictionaries (Courtois and Silberztein 1990)
  - Xerox tools (Beesley and Karttunen 2003)
  - Meta-grammar approaches (Jacquemin, 2001)
  - Morphosyntactic databases (Alegria et al., 2004)
  - Formalisms dedicated to verbs :
    - explanatory combinatorial dictionary (Melcuk et al. 1984, 1988, 1992, 1999)
    - lexicon-grammar (Gross 1994, Leclère 2005)
    - valence dictionaries (Dang et al. 2000, Benesová et al. 2008, Przepiórkowski et al. 2014)
    - ontological approaches (Marjorie McShane and Beale 2005)
    - unification grammar-bound lexicons (Sag et al. 2002, Villavicencio et al. 2004, Samaridi and Markantonatou 2014)

MWE Identification

- MWE extraction from monolingual texts (Smadja 1992, Daille 1996, Pecina 2010, Ramisch 2015)
- MWE extraction from bilingual texts (Tsvetkov and Wintner 2010, Morin and Daille 2010, Delpech et al. 2012)
- MWE lexicon enrichment (Sikora and Wolinski 2009, Krstev et al., 2013)
- MWE identification in running texts (Vincze et al. 2013, Schneider et al. 2014)

MWEs in Treebanks

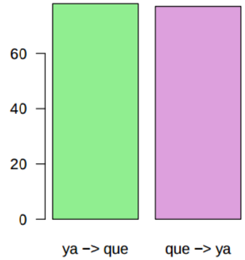
- Prague Dependency Treebank of Czech (Böhmová et al., 2003, Bejcek and Stranák 2010).
- Estonian corpus (Kaalep and Muischnek 2008)
- Hungarian Szeged Treebank (Vincze et al, 2013)

The French Treebank

- Contiguous compounds and compound proper names
- 3-level flat substructures
  - categories
  - morphological tags constituents
  - constituent words
- Consistency problems (Green et al., 2012)

Universal Dependencies

- Contains `mwe`, `compound` and `name` relations
- Guidelines for head-initial MWE annotation
- Consistency problems



MWEs and Parsing

- Gold tokenisation improves parsing (Nivre and Nilsson 2004, Korkontzelos and Manandhar 2010)
- Pre-tokenisation with CRFs and structured perceptron (Schneider 2014, Constant et al. 2013)
- MWEs as dependencies (Erygit et al. 2011, Seddah et al. 2013, Vincze et al. 2013, Candito and Constant 2014, Nasr et al. 2015)
- MWEs as non-terminal nodes (Arun and Keller 2005, Green et al. 2011)
- MWE extraction can use parsing results (Seretan 2011)

Outline

1. Project Overview
2. State of the Art
3. Work Packages

Overview of WPs

- WP 1 : MWE representation and annotation
- WP 2 : MWE lexicon
- WP 3 : MWE-aware statistical dependency parsing
- WP 4 : MWE-aware symbolic parsing
- WP 5 : Parsing-enabled MWE linking
- WP 6 : Integration and dissemination

WP 1 : MWE representation and annotation

- Coordinators : ALPAGE (Marie Candito) and LIGM (Mathieu Constant)
- Goals :
  1. Select the criteria for MWE identification, classification, description
  2. Produce a gold standard corpus
- Outcomes
  1. State-of-the-art report on MWE representation
  2. Guidelines indicating the criteria to identify and classify MWEs
  3. List of properties to be encoded in the lexicon and an annotation scheme
  4. Gold standard corpus manually annotated by experts, including deep MWE annotation

WP 2 : MWE lexicon

- Coordinators : LI (Agata Savary) and LIGM (Mathieu Constant)
- Goal :
  1. Build MWE lexicons including morphological, distributional, syntactic and semantic information
  2. Multiword NEs will be associated with pragmatic information (i.e. linking with the LOD)
- Outcomes
  1. New lexical resource, free license, standard format
  2. Tool to project MWE lexicon on treebanks

WP 3 : MWE-aware statistical dependency parsing

- Coordinators : ALPAGE (Djamé Seddah) and LIF (Alexis Nasr)
- Goals :
  1. Adapt surface dependency parsing algorithms to combined MWE and syntactic representation
  2. Experiment MWE-aware parsing architectures for the full range of MWEs
  3. Design procedures to integrate MWE lexicon online lookup in a statistical dependency parser
- Outcomes
  1. Surface dependency MWE-aware syntactic parsers
  2. Extension to deep syntactic parsing

     neutralization of syntactic variation, compatible with MWE representation

WP 4 : MWE-aware symbolic parsing

- Coordinators : LIFO (Yannick Parmentier) and ALPAGE (Eric de la Clergerie)
- Goals :
  1. Enrich existing formal grammars with MWE syntactic descriptions
  2. Provide NLP applications with extended broad-coverage MWE-aware resources
- Outcomes
  1. Extended grammatical resource, distributed under a free license

WP 5 :

- Coordinators : LIF (Carlos Ramisch) and LI (Agata Savary)
- Goals :
    1. Develop MWE linking system that links MWEs recognized in WP 3/4 to entries in knowledge bases (lexicons and Linked Open Data)
- Outcomes
    1. An MWE linking system for French

WP 6 : Integration and dissemination

- Coordinators : LIGM (Mathieu Constant)
- Goals :
    1. Integrate tools developed in other WPs in a web-based demonstrator
    2. Produce user-friendly multiplatform releases of these tools
- Outcomes
    1. Web demonstrator integrating parsing and linking tools developed in other WPs
    2. Final release of the unified lexical resource with documentation
    3. Final release of the parsing and linking tools with documentation