Representation and Processing of Composition, Variation and Approximation in Language Resources and Tools
Towards an accreditation to supervise research
Vers une habilitation à diriger des recherches (HDR)

Agata Savary

Laboratoire d’informatique
Université François Rabelais Tours, Blois

March 27, 2014
Compositionality – controversial notion

Key notion in linguistics, philosophy, logic and computer science.

*The possibility for us to understand sentences which we have never heard before is evidently based on the fact that we construct the sense of a sentence from parts which correspond to the words.*

(Frege, XIX c.)

A compound expression is *compositional* if its meaning is a function of the *meanings of its parts* and of the *syntactic rule* by which they are combined.

(Parthe et al., 1990) *horse races vs. race horses*
Compositionality – controversial notion

Key notion in linguistics, philosophy, logic and computer science.

The possibility for us to understand sentences which we have never heard before is evidently based on the fact that we construct the sense of a sentence from parts which correspond to the words. (Frege, XIX c.)

A compound expression is compositional if its meaning is a function of the meanings of its parts and of the syntactic rule by which they are combined. (Partee et al., 1990) horse races vs. race horses

Compositionality is a property of a grammar. (Kracht, 2007)
Compositionality – controversial notion

Key notion in linguistics, philosophy, logic and computer science.

*The possibility for us to understand sentences which we have never heard before is evidently based on the fact that we construct the sense of a sentence from parts which correspond to the words.*

(Frege, XIX c.)

*A compound expression is compositional if its meaning is a function of the meanings of its parts and of the syntactic rule by which they are combined.*

(Partee et al., 1990) horse races vs. race horses

**Compositionality is a property of a grammar.** (Kracht, 2007)

Benefits for modeling and computation

Preventing a combinatorial explosion of lexicalized cases.
Non-compositionality of compounds

Semantic non-compositionality

*Cordon bleu* 'expert cook’ is not a blue cord.

Morphosyntactic non-compositionality *(Savary et al., 2007)*

- *chief justices* vs. *lord justices, lords justice, lords justices*
- *[czerwony pajak]*_{mascAnim}^mascHum ’red spider (ex-communist)’
Non-compositionality of compounds

Semantic non-compositionality

- *Cordon bleu* 'expert cook' is not a blue cord.

Morphosyntactic non-compositionality (Savary et al., 2007)

- *chief justices* vs. *lord justices*, *lords justice*, *lords justices*
- *[czerwony pajak]<sub>mascAnim</sub>]<sub>mascHum</sub> 'red spider (ex-communist)'

Lexicalization

An expression $E$ has a meaning, a reference or inflectional properties that are not totally **compositional** $\Rightarrow$ $E$ has to be explicitly mentioned and described in a lexicon.
“Frozenness” – a measure of non-compositionality

“Frozenness” (G. Gross 1988; Sag et al. 2002; Mel’čuk, 2010)

Blocking the linguistic transformations typical for a syntactic structure under study:

- **French**: *Luc a pris un train de campagne* ⇒ *Luc a pris un train.*
  
  ’Luc took a suburb train ⇒ Luc took a train’

- **French**: *Le gouvernement a pris un train de mesures* ≠ *Le gouvernement a pris un train.*
  
  ’The government took a “train of measures”. ≠ The government took a train’.

Degree of “frozenness” (G. Gross 1990)
Linguistic variation

Types of variants (Jacquemin 2001; Savary & Jacquemin, 2003)

- graphical variants
  - behavioral model → behavioural model

- morphological variants
  - image converter → image conversion

- semantic variants
  - automobile cleaning → car washing

- syntactic variants
  - processing of cardiac image → image processing
The same concept has different surface realizations in texts

Example in IR:

- document phrase:
  - the philosophy and implementation of an experimental interface

- terms (for extraction or indexation):
  - *interface philosophy*,
  - *interface implementation*,
  - *philosophy implementation*. 
Contents

1. Composition and Variation – an Introduction
2. Multi-Word Expressions
3. Compound Named Entities and Beyond
4. Finite-State Methods for Word and Tree Approximation
5. Conclusions and Perspectives
6. Research Framework and Management
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*.
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*.

**MWEs – definition criteria**
- being composed of 2 or more words,
- show some degree of morphological, distributional or semantic non-compositionality,
- have unique and constant references.
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*.

MWEs – definition criteria

- being composed of 2 or more words,
- show some degree of morphological, distributional or semantic non-compositionality,
- have unique and constant references.

Pragmatic definition *(Savary, 2005)*

MWE = a sequence of graphical items which, for some application-dependent reasons, has to be listed, described and processed as a unit.
Multi-Word Expressions

MWEs – basic facts

- **prevalence** (40% of text items belong to MWEs),
- **idiosyncracies at different levels** (lexicon, grammar, meaning, ...),
- **sparseness** (most MWEs appear rarely in corpora),
- MWEs are under-represented in language resources and tools,
- MWEs are hard to detect, understand, translate, etc.
Idiosyncrasy of MWEs . . .

. . . at different NLP levels

- segmentation:
  - French: *bons* _hommes_ 'fellows'
  - English: *personal computer*
  - English: *put sth. off*

- morphology
  - French: *grand-mères_ ‘grand _sing.masc- mothers _pl. fem’*
  - Polish: *wybory powszechne_ ‘general elections’, *wybór powszechny*

- syntax
  - English: *all of a sudden*
  - English: *he kicked the bucket, *the bucket was kicked by him*

- semantics
  - English: *to spill the beans = to reveal a secret*
MWEs in NLP - State of the art

- ** lexical description of MWEs**  
  **SOA:** (Savary, 2008)

- **DELA e-dictionaries**  (Courtois et al., 1990; Silberztein, 1993a; **Savary, 2000**; Kyriacopoulou et al., 2002; Silberztein, 2005)

- **two-level morphology**  (Beesley & Karttunen, 2003; Karttunen et al., 1992; Karttunen, 1993; Breidt et al., 1996; Oflazer et al., 2004)

- **relational DB**  (Alegria et al., 2004; Itai & Wintner, 2013),

- **parameterized equivalence classes**  (Grégoire, 2010)

- **unification grammars and meta-grammars**  (Sag et al., 2002; Copestake et al., 2002; Villavicencio et al., 2004; Jacquemin, 2001)
MWEs in NLP - state of the art ctd.

- **MWE extraction**
  - **SOA:** (Savary & Jacquemin, 2003)
  - Monolingual: (Smadja, 1992; Daille, 1996; Pecina, 2010; Al-Haj & Wintner, 2010; Ramisch et al., 2010; Davis & Barrett, 2013)
  - Bilingual: (Tsvetkov & Wintner, 2010; Morin & Daille, 2010; Delpech et al., 2012)

- **MWE identification**
  - (NER systems; Vincze et al., 2013)

- **MWE annotation**
  - (Abeillé et al., 2003; Bejček & Straňák, 2010; Laporte et al., 2008a,b; Kaalep & Muischnek, 2008)

- **Parsing and MWEs**
  - (Abeillé & Schabes, 1989; Sag et al., 2002; Copestake et al., 2002; Villavicencio et al., 2004; Nivre & Nilsson, 2004; Attia, 2006; Finkel & Manning (2009a), Wehrli et al., 2010, Constant et al., 2013, Green et al., 2013)
Multiflex – describing the morphosyntax of contiguous MWEs

(Savary, 2005, 2008, 2009; Savary et al., 2007, 2009; Graliński et al., 2010)

Two-layer approach

- single words are analysed and generated by an external module,
- MWE inflection graphs combine single forms into MWE forms

Interoperability constraints for the underlying single words module

- same morphological model for the language under study,
- clear-cut definition of a token,
- generation of inflected forms of simple words.
Multiflex – inflection, agreement and non-compositionality

<table>
<thead>
<tr>
<th>15 variants</th>
<th>Lemma</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>czerwony pająk</td>
<td>czerwony pająk</td>
<td>sg:nom:m2</td>
</tr>
<tr>
<td>czerwone pająki</td>
<td>czerwony pająk</td>
<td>pl:acc:m2</td>
</tr>
<tr>
<td>czerwonych pająków</td>
<td>czerwony pająk</td>
<td>pl:acc:m1</td>
</tr>
</tbody>
</table>

...
Multiflex – inflection, agreement and non-compositionality

<table>
<thead>
<tr>
<th>15 variants</th>
<th>Lemma</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>czerwony pająk</strong></td>
<td>czerwony pająk</td>
<td>sg:nom:m2</td>
</tr>
<tr>
<td><strong>czerwone pająki</strong></td>
<td>czerwony pająk</td>
<td>pl:acc:m2</td>
</tr>
<tr>
<td><strong>czerwonych pająków</strong></td>
<td>czerwony pająk</td>
<td>pl:acc:<em>m1</em></td>
</tr>
</tbody>
</table>

...
Multiflex – inflection, agreement and non-compositionality

15 variants

- *czerwony pająk*  
  - Lemma: czerwony  
  - Class: adj  
  - Nb: sg  
  - Case: nom  
  - Gen: m2  
  - Deg: pos

- *czerwone pająki*  
  - Lemma: czerwony pająk  
  - Class: subst  
  - Case: nom  
  - Gen: m2

- *czerwonych pająków*  
  - Lemma: czerwony pająk  
  - Nb: pl  
  - Case: acc  
  - Gen: m1

...
Multiflex – syntactic variation & agreement

<table>
<thead>
<tr>
<th>126 variants</th>
<th>Lemma</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan Rodowicz „Anoda”</td>
<td>Jan Rodowicz „Anoda”</td>
<td>sg:nom:m1:offic</td>
</tr>
<tr>
<td>Jan Rodowicz Anoda</td>
<td>Jan Rodowicz „Anoda”</td>
<td>sg:nom:m1</td>
</tr>
<tr>
<td>Jan „Anoda” Rodowicz</td>
<td>Jan Rodowicz „Anoda”</td>
<td>sg:nom:m1</td>
</tr>
<tr>
<td>J. Rodowicz „Anoda”</td>
<td>Jan Rodowicz „Anoda”</td>
<td>sg:nom:m1</td>
</tr>
<tr>
<td>J. Rodowicz</td>
<td>Jan Rodowicz „Anoda”</td>
<td>sg:nom:m1</td>
</tr>
<tr>
<td>„Anoda” Rodowicz</td>
<td>Jan Rodowicz „Anoda”</td>
<td>sg:nom:m1</td>
</tr>
<tr>
<td>Rodowicz</td>
<td>Jan Rodowicz „Anoda”</td>
<td>sg:nom:m1:spok</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Multiflex – syntactic variation & agreement

126 variants

<table>
<thead>
<tr>
<th>126 variants</th>
<th>Lemma</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan Rodowicz „Anoda”</td>
<td>Jan Rodowicz „Anoda”</td>
<td>sg:nom:m1:offic</td>
</tr>
<tr>
<td>Jan Rodowicz Anoda</td>
<td>Jan Rodowicz „Anoda”</td>
<td>sg:nom:m1</td>
</tr>
<tr>
<td>Jan „Anoda” Rodowicz</td>
<td>Jan Rodowicz „Anoda”</td>
<td>sg:nom:m1</td>
</tr>
<tr>
<td>J. Rodowicz „Anoda”</td>
<td>Jan Rodowicz „Anoda”</td>
<td>sg:nom:m1</td>
</tr>
<tr>
<td>J. Rodowicz</td>
<td>Jan Rodowicz „Anoda”</td>
<td>sg:nom:m1</td>
</tr>
<tr>
<td>„Anoda” Rodowicz</td>
<td>Jan Rodowicz „Anoda”</td>
<td>sg:nom:m1</td>
</tr>
<tr>
<td>Rodowicz</td>
<td>Jan Rodowicz „Anoda”</td>
<td>sg:nom:m1:spok</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## MULTIFLEX – nesting

<table>
<thead>
<tr>
<th>336 variants</th>
<th>Lemma</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>aleja Jana Rodowicza „Anody”</td>
<td>aleja Jana Rodowicza „Anody”</td>
<td>sg:nom:f:offic</td>
</tr>
<tr>
<td>al. Rodowicza</td>
<td>aleja Jana Rodowicza „Anody”</td>
<td>sg:nom:f:neut</td>
</tr>
<tr>
<td>Rodowicza</td>
<td>aleja Jana Rodowicza „Anody”</td>
<td>sg:nom:f:spok</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
MULTIFLEX – nesting

336 variants

<table>
<thead>
<tr>
<th>Lemma</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>aleja Jana Rodowicza „Anody”</td>
<td>sg:nom:f:offic</td>
</tr>
<tr>
<td>al. Rodowicza</td>
<td>sg:nom:f:neut</td>
</tr>
<tr>
<td>Rodowicza</td>
<td>sg:nom:f:spok</td>
</tr>
</tbody>
</table>

$1$ $2$ $3$

- lemma: aleja
- class: subst
  ...

- lemma: Jan Rodowicz „Anoda”
- class: subst
  ...

\[
\begin{align*}
\text{aleja} & \quad Jana Rodowicza „Anody” \\
\text{\$1} & \quad \text{\$2} & \quad \text{\$3} \\
\text{lemma: aleja} & \quad \text{lemma: Jan Rodowicz „Anoda”} \\
\text{class: subst} & \quad \text{class: subst} \\
\text{\ldots} & \quad \text{\ldots}
\end{align*}
\]
Multiflex – applications

Software integration

- Unitex (LGPL)  
  (Paumier, 2008),
- LeXimir  
  (Krstev et al., 2013)
- Toposław  
  (Marciniak et al., 2009b; Sikora & Woliński, 2009)
Multiflex – applications

Software integration

- Unitex (LGPL) (Paumier, 2008),
- LeXimir (Krstev et al., 2013)
- Toposław (Marciniak et al., 2009b; Sikora & Woliński, 2009)

MWE e-dictionaries

<table>
<thead>
<tr>
<th>Dictionary name</th>
<th>Language</th>
<th>Type</th>
<th>Lexicogr. framework</th>
<th>Dictionary size</th>
<th>License</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serbian DELAC</td>
<td>Serbian</td>
<td>general-purpose</td>
<td>LeXimir</td>
<td>11,000</td>
<td>cc-by sa</td>
</tr>
<tr>
<td>Greek DELAC</td>
<td>modern Greek</td>
<td>general-purpose</td>
<td>Unitex</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAWA</td>
<td>Polish</td>
<td>urban proper names</td>
<td>Toposław</td>
<td>9,000</td>
<td>cc-by sa</td>
</tr>
<tr>
<td>SEJF</td>
<td>Polish</td>
<td>general-purpose</td>
<td>Toposław</td>
<td>3,200</td>
<td>cc-by sa</td>
</tr>
<tr>
<td>SEJFEK</td>
<td>Polish</td>
<td>economic terms</td>
<td>Toposław</td>
<td>11,000</td>
<td>cc-by sa</td>
</tr>
</tbody>
</table>
Contents

1. Composition and Variation – an Introduction
2. Multi-Word Expressions
3. Compound Named Entities and Beyond
4. Finite-State Methods for Word and Tree Approximation
5. Conclusions and Perspectives
6. Research Framework and Management
Named and naming entities

real world

(discourse world)

language

*Blois*

named entities

named entities
In NLP naming entities are usually called **named entities**.
Named/Naming Entities and beyond

**NEs – NLP-central objects**
- are/refer to persons, places, objects, events, . . ., crucial for text understanding,
- hard to translate,
- central to IR, IE and QA.

**NEs – controversial objects** (Ehrmann, 2008)
- theoretical studies vs. applicative motivations,
- onomasiological vs. semasiological definitions.
Named/Naming Entities and beyond

**NEs – NLP-central objects**
- are/refer to persons, places, objects, events, . . ., crucial for text understanding,
- hard to translate,
- central to IR, IE and QA.

**NEs – controversial objects** *(Ehrmann, 2008)*
- theoretical studies vs. applicative motivations,
- onomasiological vs. semasiological definitions.

**Beyond NEs**
- mentions (coreference annotation and resolution),
- entities (entity linking).
Most NEs are MWEs

Multi-word NEs in lexicons
- SAWA\textsuperscript{a}: 98\% entries are MWEs,
- Prolexbase: 66\% entries are MWEs.

\textsuperscript{a}Grammatical Lexicon of Warsaw Urban Proper Names

Multi-word NEs in corpora
- National Corpus of Polish: 53\% of the (outermost) NEs are MWEs or ellipses of MWEs.
NEs in NLP - State of the art

Impact on naming entities (names)

Impact on named entities (objects, concepts)

- NE annotation and recognition
  - NKJP 2010
  - PDT 2010
- NE coreference resolution
  - CoNLL 2002-3
  - MUC 1995-7
- NER & coref. res.
  - CoNLL 2012
  - ACE 2004-8

- Interlinked Lexical Ontologies
  - Prolexbase 2005-

- Entity Linking & NE disambiguation
  - TAC 2008-
  - LOD 2009- (DBpedia, YAGO, ...)

- Data linking
NEs in NLP - State of the art

Impact on naming entities (names)

Impact on named entities (objects, concepts)

NER & coref. res.

PDT 2010

MUC 1995-7

CoNLL 2002-3

NKJP 2010

NE annotation and recognition

Interlinked Lexical Ontologies

Prolexbase 2005-

Conclusions

CV

A. Savary

HDR

27/03/2014

22 / 44
National Corpus of Polish

**NKJP (National Corpus of Polish)**

- 1.5-billion ($1.5 \times 10^9$) word corpus,
- 300-million word balanced subcorpus,
- 1-million word manually annotated subcorpus (2 parallel annotators + 1 adjudicator),
- multilevel annotation: segmentation, morphosyntax, WSD, syntactic words, syntactic groups, **NEs**, 
- additional **coreference** level (*Polish Coreference Corpus*),
- distributed under GNU GPL v3 and CC BY v.3.
NEs and mentions in NKJP and PCC – novelty

Common annotation aspects

- recursively **nested** NEs and mentions,
- coordinated and discontinuous NEs and mentions.
NEs and mentions in NKJP and PCC – novelty

Common annotation aspects

- recursively nested NEs and mentions,
- coordinated and discontinuous NEs and mentions.

NE annotation aspects (Savary et al., 2012)

- relative adjectives, personal derivations and derivational bases
  \( \text{amerykański} \leftarrow \text{Stany Zjednoczone} \ '\text{American} \leftarrow \text{United States'} \).
NEs and mentions in NKJP and PCC – novelty

Common annotation aspects
- recursively **nested** NEs and mentions,
- coordinated and discontinuous NEs and mentions.

NE annotation aspects (Savary et al., 2012)
- relative adjectives, personal derivations and derivational bases
  \( \text{amerykański} \leftarrow \text{Stany Zjednoczone} \quad '\text{American} \leftarrow \text{United States'} \).

Coreference annotation aspects (Ogrodniczuk et al., 2013)
- dominant expressions. & semantic heads,
NEs as annotation trees

działkowcy z województw : poznańskiego i bydgoskiego
'garden-owners from the regions : Poznań-adj and Bydgoszcz-adj'
Annotation tools for nested NEs

**TrEd** (Pajas & Štěpánek, 2008)

- customized to constituency trees,
- adjudication.
Annotation tools for nested NEs

**TrEd** *(Pajas & Štěpánek, 2008)*
- customized to constituency trees,
- adjudication.

**SProUT** *(Savary & Piskorski, 2011)*
- customized rule-based NER,
- 78% P, 38% R.
Annotation tools for nested NEs

**TrEd** *(Pajas & Štěpánek, 2008)*
- customized to constituency trees,
- adjudication.

**SProUT** *(Savary & Piskorski, 2011)*
- customized rule-based NER,
- 78% P, 38% R.

**NERF** *(Waszczuk et al., 2013)*
- ML-based NER,
- P 80%, R 74%.
One of the largest multilevel-annotated corpora worldwide

- 87,000 (gold standard) NEs,
- 180,000 mentions; 109,000 coref. clusters,
- inter-annotator agreement:
  - $F_1 = 0.83$ (NEs),
  - $\kappa = 0.74$ (mentions).
Contents

1 Composition and Variation – an Introduction
2 Multi-Word Expressions
3 Compound Named Entities and Beyond
4 Finite-State Methods for Word and Tree Approximation
5 Conclusions and Perspectives
6 Research Framework and Management
String approximation - state of the art

**String-to-string correction** (Damerau, 1964; Wagner & Fisher, 1974; Lowrance & Wagner, 1975; Du & Chang, 1992)

- **Context**: elementary *edit operations* on letters with costs; allowed *edit sequences*.
- **Input**: two *strings* \( x \) and \( y \).
- **Output**: \( ed(x, y) \) – *edit distance* between \( x \) and \( y \).
String approximation - state of the art

**String-to-string correction** (Damerau, 1964; Wagner & Fisher, 1974; Lowrance & Wagner, 1975; Du & Chang, 1992)

- **Context**: elementary edit operations on letters with costs; allowed edit sequences.
- **Input**: two strings $x$ and $y$.
- **Output**: $ed(x, y)$ – edit distance between $x$ and $y$.

**String-to-language correction** (SOA by Boytsov, 2011; Savary, 2003)

- **Context**: as above.
- **Input**:
  - string language (dictionary) $L$,
  - string $x$,
  - threshold $th$.
- **Output**: strings $y \in L$ such that $ed(x, y) \leq th$. 
Tree approximation - state of the art

Tree-to-tree correction *(Selkow, 1977; Tai, 1979; Zhang & Shasha, 1989)*

- **Context**: elementary edit operations on tree nodes or subtrees; edit sequences.
- **Input**: two trees $x$ and $y$.
- **Output**: $ed(x, y)$ – edit distance between $x$ and $y$. 
Tree approximation - state of the art

**Tree-to-tree correction** (Selkow, 1977; Tai, 1979; Zhang & Shasha, 1989)

- **Context**: elementary edit operations on tree nodes or subtrees; edit sequences.
- **Input**: two trees $x$ and $y$.
- **Output**: $ed(x, y)$ – edit distance between $x$ and $y$.

**Tree-to-language correction** (SOA by Tekli et al., 2011)

- **Context**: as above.
- **Input**:
  - tree language $L$,
  - tree $x$,
  - threshold $th$.
- **Output**: trees $y \in L$ such that $ed(x, y) \leq th$. 
**XMLCorrector**: XML document correction wrt. a DTD

- **Tree-to-tree correction** (Selkow, 1977)
- **String-to-string correction** (Wagner & Fischer, 1974; Du & Chang, 1992)
- **String-to-language correction** (Oflazer, 1996)
**XMLCorrector**: XML document correction wrt. a DTD

- **Tree-to-tree correction** (Selkow, 1977)
- **String-to-string correction** (Wagner & Fischer, 1974; Du & Chang, 1992)
- **Tree-to-language correction** (Amavi et al., 2013)
- **String-to-language correction** (Oflazer, 1996)
Input

- $t$ – XML tree,
- $S$ – a structure description (DTD),
- $th$ – threshold,
- $c$ – intended root node.
XMLCorrector  (Amavi et al., 2013)

**Input**
- $t$ – XML tree,
- $S$ – a structure description (DTD),
- $th$ – threshold,
- $c$ – intended root node.

**Output**
- Node-edit operation sequences allowing to get all trees $t' \in L(S)$ such that $ed(t, t') \leq th$. 
XMLCorrector: example

\[ S = \{ \text{root: } b^* \mid ab*c; \}
\]

\[
\begin{align*}
\text{b: } & \text{ cd;} \\
\text{b: } & \text{ c;} \\
\text{c: } & \epsilon; \\
\text{d: } & \epsilon
\end{align*}
\]

\[ th = 2 \]
XMLCorrector

First full-fledged tree-to-language correction algorithm and implementation

- correction trees, **sequences** and distances returned,
- **all** candidates within a threshold found,
- complexity, correctness and soundness **proofs**,
- **GNU LGPL** license,
- test data available (reproducibility).
Contents

1. Composition and Variation – an Introduction
2. Multi-Word Expressions
3. Compound Named Entities and Beyond
4. Finite-State Methods for Word and Tree Approximation
5. Conclusions and Perspectives
6. Research Framework and Management
Compositional modeling and computation

Advantage

- Preventing a combinatorial explosion of lexicalized cases
  - compositional calculus of emotional valency \( \text{(Tallec et al., 2010)} \),
  - nested description of MWEs in Multiflex \( \text{(Savary et al., 2009)} \).
- Better modeling of semantic relations
  - nested NE annotation in NKJP \( \text{(Savary et al., 2012)} \),
  - nested mention annotation in PCC \( \text{(Ogrodniczuk et al., 2013)} \).
Compositional modeling and computation

Advantage

- Preventing a combinatorial explosion of lexicalized cases
  - compositional calculus of emotional valency \((\text{Tallec et al., 2010})\),
  - nested description of MWEs in Multiflex \((\text{Savary et al., 2009})\).
- Better modeling of semantic relations
  - nested NE annotation in NKJP \((\text{Savary et al., 2012})\),
  - nested mention annotation in PCC. \((\text{Ogrodniczuk et al., 2013})\).

Challenges: MWEs

- MWEs defy compositionality principles \((\text{Savary et al., 2007})\),
- MWEs are usually partly frozen and partly variable,
- heterogeneous properties should be accounted for simultaneously.
Variability – central challenge in NLP

Objective

- Conflate different surface realizations of the same underlying concept.
Variability – central challenge in NLP

**Objective**
- Conflate different surface realizations of the same underlying concept.

**Means**
- lexical and grammatical description,
- algorithmic approximation.
Variability – central challenge in NLP

Objective

- Conflate different surface realizations of the same underlying concept.

Means

- lexical and grammatical description,
- algorithmic approximation.

Multilinguality provides a better understanding of linguistic variability.
Perspectives

Objective

- enhancing and extending language resources and tools,
- integrating language data into Linked Open Data (LOD),
- integrating MWEs in deep parsing,
- taxonomy and benchmarking for tree-to-language correction,
- modeling MWE identification as a tree-to-language correction problem.
Contents

1. Composition and Variation – an Introduction
2. Multi-Word Expressions
3. Compound Named Entities and Beyond
4. Finite-State Methods for Word and Tree Approximation
5. Conclusions and Perspectives
6. Research Framework and Management
External collaborations

- Polish Academy of Sciences, Warsaw
- Gdańsk University of Technology
- University of Poznań
- University of Olsztyn
- University of Belgrade
- Université Paris Est Marne-la-Vallée
- University of Orléans
- Tomsk State University
Event organisation – OC co-chair

Blois, 12–16 July, 2011

- 16th International Conference on Implementation and Application of Automata (CIAA-2011),
- 9th International Workshop on Finite State Methods and Natural Language Processing (FSMNLP-2011),
- 95 participants, 65,000 € budget.
## Participation in funded collaborative projects

<table>
<thead>
<tr>
<th>Project</th>
<th>Dates</th>
<th>Budget</th>
<th>Coordinator</th>
<th>Funding</th>
</tr>
</thead>
<tbody>
<tr>
<td>PARSEME</td>
<td>2013–2017</td>
<td>680,000 €</td>
<td>A. Savary</td>
<td>COST</td>
</tr>
<tr>
<td>CORE</td>
<td>2011–2014</td>
<td>120,000 €</td>
<td>IPIPAN</td>
<td>NCN</td>
</tr>
<tr>
<td>CESAR</td>
<td>2011–2013</td>
<td></td>
<td>Hungarian</td>
<td>EC (PSP)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Ac. of Sc.</td>
<td></td>
</tr>
<tr>
<td>NEKST</td>
<td>2009–2014</td>
<td>3,500,000 €</td>
<td>IPIPAN &amp; PWr</td>
<td>ERDF</td>
</tr>
<tr>
<td>CODEX</td>
<td>2009–2012</td>
<td>68,336 €</td>
<td>INRIA</td>
<td>ANR</td>
</tr>
<tr>
<td>LUNA.PL</td>
<td>2008–2009</td>
<td></td>
<td>IPIPAN</td>
<td>MNSW</td>
</tr>
<tr>
<td>NKJP</td>
<td>2007–2010</td>
<td>600,000 €</td>
<td>IPIPAN</td>
<td>MNSW</td>
</tr>
<tr>
<td>EmotiRob</td>
<td>2007–2009</td>
<td>85,200 €</td>
<td>Univ. Bretagne</td>
<td>ANR</td>
</tr>
<tr>
<td>Polonium</td>
<td>2007–2008</td>
<td>6,070 €</td>
<td>LI &amp; IPIPAN</td>
<td>PHC EGIDE</td>
</tr>
<tr>
<td>Pavle Savic</td>
<td>2004–2005</td>
<td>5,500 €</td>
<td>LI &amp; Belgrade Univ.</td>
<td>PHC EGIDE</td>
</tr>
<tr>
<td>NomsPropres</td>
<td>2003–2005</td>
<td>94,000 €</td>
<td>LI</td>
<td>RNTL</td>
</tr>
<tr>
<td>Outilex</td>
<td>2002–2006</td>
<td></td>
<td>Paris-Est</td>
<td>RNTL</td>
</tr>
</tbody>
</table>
IC1207 COST action

- scientific network: 30 COST countries
- bottom-up approach,
- 114 members, 4 working groups,
- 29 languages from 9 language families,

Scientific objective

To bridge the gap between linguistic precision and computational efficiency in NLP applications.

Key issue: MWEs and their links to (deep) parsing.
PARSEME (PARsing and Multi-word Expressions)

IC1207 COST action

- scientific network: 30 COST countries
- bottom-up approach,
- 114 members, 4 working groups,
- 29 languages from 9 language families,

Scientific objective

To bridge the gap between linguistic precision and computational efficiency in NLP applications.

Key issue: MWEs and their links to (deep) parsing.
Thank you!

**C-structure**

```
ROOT
  Sadj
    S
      NP
        PRON he
      VPv
        V
          kicked
            D
              the
                NPadj
                    N
                      bucket
```

**F-structure**

```
PRED 'kick<[8:he], [2:bucket]>'
  OBJ
    SPEC 6 DET 7 PRED 'the'
  SUBJ 8 PRED 'he'
  PRED 'bucket'
```
MWE processing and tree-to-language correction

Modelling a MWEs as a tree language

- A MWE can have (infinitely) many potential instantiations
  
  *to count somebody in, he has counted me and Laura in, I have never counted this idot with stange ideas in*
  
- These instantiations form a **tree language** (of which type?).
MWE processing and tree-to-language correction

Modelling a MWEs as a tree language

- A MWE can have (infinitely) many potential instantiations
  
  to count somebody in, he has counted me and Laura in, I have never counted this idot with stange ideas in
  
- These instantiations form a tree language (of which type?).

Recognizing a MWE in a syntax tree

A selected subtree is corrected wrt. the tree language.
MWE processing and tree-to-language correction

Modelling a MWEs as a tree language

- A MWE can have (infinitely) many potential instantiations:
  - *to count somebody in*, *he has counted me and Laura in*, *I have never counted this idiot with strange ideas in*
- These instantiations form a **tree language** (of which type?).

Recognizing a MWE in a syntax tree

A selected subtree is corrected wrt. the tree language.

Applications

- Annotating MWEs in treebanks.
- Recognizing MWEs after parsing.
- Approximate MWE recognition in noisy input.