Multilingual linguistic workflows.  
A case study for two languages  

Dan Cristea$^{1,2}$ and Ionuţ Pistol$^1$  
{dcristea, ipistol}@info.uaic.ro  

1 “Alexandru Ioan Cuza” University of Iaşi – Faculty of Computer Science  
2 Romanian Academy – Institute for Computer Science
The demands of multilinguality in language technology

– Necessity of reusing the processing modules performing specific linguistic tasks for different languages

– The language a module is able to interpret becomes commanded by the resources/parameters it is fuelled with

– Interoperability requirements: a standardization of the processing steps and an efficient building and execution of workflows
The Formats and Modules Hierarchy (FMH)

• A model to interpret XML meta-data
• An environment allowing automatic configuration of processing chains

• A project partially motivated by multilinguality!
A FMH is a directed acyclic graph configuring the metadata of linguistic annotations in a hierarchy of XML schemes.

- each *node* corresponds to a distinct *XML annotation scheme*
- edges correspond to different types of processing
The subsuming relation: a lattice-like structure

• Node (scheme) $A$ subsumes node (scheme) $B$ iff:
  – any element (tag-name) of $A$ is also in $B$;
  – any attribute in the list of attributes of a tag-name in $A$ is also in the list of attributes of the same tag-name of $B$
The subsuming relation: an example

A

<document>
  <tok id="1">This</tok>
  <tok id="2">is</tok>
  <tok id="3">a</tok>
  <tok id="4">sentence</tok>
  <tok id="5">.</tok>
</document>

B

<document>
  <seg id="s1">
    <tok id="t1" pos="p">This</tok>
    <tok id="t2" pos="v">is</tok>
    <tok id="t3" pos="d">a</tok>
    <tok id="t4" pos="n">sentence</tok>
    <tok id="t5" pos="m">.</tok>
  </seg>
</document>
Processing operators

• A FMH model includes three types of operators:
  – pipelining (along subsuming relations)
  – simplification (opposite subsuming relations)
  – merging
Pipelines

A process $p$ should be able to transform a file observing the restrictions of the scheme $A$ into one including the supplementary information as given by the scheme $B$.
Simplification

- A file $f_B$ is simplified to a file $f_A$ if:
  - $A$ and $B$ are nodes in the subsuming relation ($A$ subsumes $B$)
  - all additional information that exist in $f_B$ and does not exist in $f_A$ is removed.
Nodes $A$, $B$, $C$... merge to node $D$ if a file $f_D$ can be obtained my merging (putting together) all annotation in the files $f_A$, $f_B$, $f_C$...
Flows

- Flows are paths in the graph linking a **start node** to a **destination node**
- Operators are applied to files observing start node restrictions and transforming them onto output files observing destination node restrictions
Instantiations

• Flows are instantiated with respect to specific conditions:
  – language
  – cost issues
  – IPR
  – etc.

• Instantiations act as filtering/disambiguation conditions on flows
Second part

A use case
The problem

• Build a gold corpus of RO texts annotated for semantic roles, by importing them from EN (in parallel texts).
  – find an instantiation of a flow linking the start node to the destination node, such that, given a parallel EN-RO document in the input, obtain the RO sentences annotated with specific SR elements.
Example

<roleset id="take.16" name="take by surprise">
  <roles>
    <role n="0" descr="causer of surprise" />
    <role n="1" descr="experiencer of surprise" />
    <role n="2" descr="surprise" />
  </roles>
  <example name="surprise!">
    <text>...the size of the cut took many Western banks by surprise.
    </text>
    <sr n="0">the size of the cut </sr>
    <rel>took </rel>
    <sr n="1">many Western banks </sr>
    <sr n="2" f="by">surprise </sr>
  </example>
</roleset>
Importing semantic role labels: the model
Importing semantic role labels: a node

$2\text{TXT}^{\text{EN,RO},S}$: two versions of the same text in EN and RO (parallel translations; $S$ means identical semantic content). Markings make explicit the two languages.
<parallel>
  <sentence lang="EN">…
    <w id="110">the</w>
    <w id="111">size</w>
    <w id="112">of</w>
    <w id="113">the</w>
    <w id="114">cut</w>
    <w id="115">took</w>
    <w id="116">many</w>
    <w id="117">Western</w>
    <w id="118">banks</w>
    <w id="119">by</w>
    <w id="120">surprise</w>
    <w id="121">.</w>
  </sentence>
  <sentence lang="RO">…
    <w id="210">dimensiuneae</w>
    <w id="211">reducerii</w>
    <w id="212">a</w>
    <w id="213">luat</w>
    <w id="214">multe</w>
    <w id="215">bănci</w>
    <w id="216">vestice</w>
    <w id="217">prin</w>
    <w id="218">surprindere</w>
    <w id="219">.</w>
  </sentence>
</parallel>
Importing semantic role labels: a manual annotation edge

SEM-LAB\textsubscript{EN} – an unrevealed (perhaps manual) process which accomplishes semantic roles labelling on English texts. The language EN is an instantiation condition, therefore only sentences marked with the attribute-value pair language=”EN” will be considered.
Importing semantic role labels: a node

SRL: the text includes semantic role markups: SR elements surrounding constituents of sentences.
<parallel>
  <sentence lang="EN">…
    <sr n="1"><w id="110">the</w>
    <w id="111">size</w>
    <w id="112">of</w>
    <w id="113">the</w>
    <w id="114">cut</w></sr>
    <rel><w id="115">took</w></rel>
    <sr n="2"><w id="116">many</w>
    <w id="117">Western</w>
    <w id="118">banks</w></sr>
    <sr n="3"><w id="119">by</w>
    <w id="120">surprise</w></sr>
    <w id="121">.</w></sentence>
  <sentence lang="RO">…
    <w id="210">dimensiun</w>
    <w id="211">reducerii</w>
    <w id="212">a</w>
    <w id="213">luat</w>
    <w id="214">multe</w>
    <w id="215">bândci</w>
    <w id="216">vestice</w>
    <w id="217">prin</w>
    <w id="218">surprindere</w></sentence>
</parallel>
Importing semantic role labels: a pipeline edge

ALIGN\textsubscript{EN,RO,S} – alignment at word level of sentences belonging to EN and RO and having the same content, S.
Importing semantic role labels: a node

$\text{WA}_S$: scheme containing word aligning markups between two texts representing the same content $S$. 

Crosslingual Language Technology – Univ. Hamburg, 7-8 May 2012
<parallel>
  <sentence lang="EN">...<w id="110">the</w> <w id="111">size</w> <w id="112">of</w> <w id="113">the</w> <w id="114">cut</w> <w id="115">took</w> <w id="116">many</w> <w id="117">Western</w> <w id="118">banks</w> <w id="119">by</w> <w id="120">surprise</w> <w id="121">.</w></sentence>
  <sentence lang="RO">...<w id="210" align="110 111">dimensiuneae</w> <w id="211" align="112 113 114">reducerii</w> <w id="212" align="115">a</w> <w id="213" align="115">luat</w> <w id="214" align="116">multe</w> <w id="215" align="118">bănci</w> <w id="216" align="117">vestice</w> <w id="217" align="119">prin</w> <w id="218" align="120">surprindere</w> <w id="219" align="121">.</w></sentence>
</parallel>
Importing semantic role labels: a merging hyper-edge

Markings belonging to both parent nodes are merged
Importing semantic role labels: a node

$\text{WASRL}_S$: scheme containing word-alignment and semantic role markups. Not all sentences contain SR markups.
<parallel>
  <sentence lang="EN">…
    <sr n="1">the  </sr>
    <w id="110">the </w>
    <w id="111">size </w>
    <w id="112">of </w>
    <w id="113">the </w>
    <w id="114">cut </w></sr>
    <rel><w id="115">took </w></rel>
    <sr n="2">many  </sr>
    <w id="116">many </w>
    <w id="117">Western </w>
    <w id="118">banks </w></sr>
    <sr n="3">by  </sr>
    <w id="119">by </w>
    <w id="120">surprise </w></sr>
    <w id="121">.</w>
  </sentence>
  <sentence lang="RO">…
    <w id="210" align="110 111">dimensiuneae </w>
    <w id="211" align="112 113 114">reducerii </w>
    <w id="212" align="115">a </w>
    <w id="213" align="115">luat </w>
    <w id="214" align="116">multe </w>
    <w id="215" align="118">bănci </w>
    <w id="216" align="117">vestice </w>
    <w id="217" align="119">prin </w>
    <w id="218" align="120">surprindere </w></sentence>
</parallel>
Importing semantic role labels: a pipeline edge

\[ \text{IMPORT}_{\text{EN,RO,S}} \] — module importing the semantic roles from one part (EN) onto the parallel part (RO), provided the two parts represent the same content (S). The module presupposes finding SR markups on all language=“EN” sentences.
<parallel>
  <sentence lang="EN">...<br/>
  <sr n="1"><w id="110">the</w><br/>
  <w id="111">size</w><br/>
  <w id="112">of</w><br/>
  <w id="113">the</w><br/>
  <w id="114">cut</w><br/>
  <rel><w id="115">took</w><br/>
  <sr n="2"><w id="116">many</w><br/>
  <w id="117">Western</w><br/>
  <w id="118">banks</w><br/>
  <sr n="3"><w id="119">by</w><br/>
  <w id="120">surprise</w><br/>
  <w id="121">.</w><br/>
  </sentence>
  <sentence lang="RO">...<br/>
  <sr n="1"><w id="210" align="110 111">dimensiuneae</w><br/>
  <w id="211" align="112 113 114">reducerii</w><br/>
  <rel><w id="212" align="115">a</w><br/>
  <w id="213" align="115">luat</w><br/>
  <sr n="2"><w id="214" align="116">multe</w><br/>
  <w id="215" align="118">bănci</w><br/>
  <w id="216" align="117">vestice</w><br/>
  <sr n="3"><w id="217" align="119">prin</w><br/>
  <w id="218" align="120">surprindere</w><br/>
  <w id="219" align="121">.</w><br/>
  </sentence>
</parallel>
Importing semantic role labels: a simplifying edge

SIMPL\textsubscript{RO} – a simplifying operation that prunes off all markings except the SR elements of the sentences marked with language=”RO”
<parallel>
  <sentence lang="EN">…
    <w id="110">the</w>
    <w id="111">size</w>
    <w id="112">of</w>
    <w id="113">the</w>
    <w id="114">cut</w>
    <w id="115">took</w>
    <w id="116">many</w>
    <w id="117">Western</w>
    <w id="118">banks</w>
    <w id="119">by</w>
    <w id="120">surprise</w>
    <w id="121">.</w>
  </sentence>
  <sentence lang="RO">…
    <sr n="1"><w id="210" align="110 111">dimensiun
eae</w>
    <w id="211" align="112 113 114">reduceri</w></sr>
    <rel><w id="212" align="115">a</w>
    <w id="213" align="115">luat</w></rel>
    <sr n="2"><w id="214" align="116">multe</w>
    <w id="215" align="118">bănci</w></sr>
    <w id="216" align="117">vestic
e</w></sr>
    <sr n="3"><w id="217" align="119">prin</w>
    <w id="218" align="120">surprindene</w></sr>
    <w id="219" align="121">.</w></sentence>
</parallel>
Importing semantic role labels: the first flow

2TXT_{EN,RO,S} > \text{SEM-LAB}_{EN} \text{ SRL}

Crosslingual Language Technology – Univ. Hamburg, 7-8 May 2012
Importing semantic role labels: the second flow

\[(2\text{TXT}_{EN,RO,S} \to_{\text{SEM-LAB}_{EN}} \text{SRL}), (2\text{TXT}_{EN,RO,S} \to_{\text{ALIGN}_{EN,RO,S}} \text{WA}_S) \to_{\text{WASRL}_{S} <_{\text{SIMPL}_{RO}} \text{SRL}}\]
Importing semantic role labels: the third flow

\[(2\text{TXT}_{\text{EN},\text{RO},\text{S}} >_{\text{SEM-LAB EN}} \text{SRL}), (\text{TXT}_{\text{EN},\text{RO},\text{S}} >_{\text{ALING EN,RO,S}} \text{WA}_S)) > \text{WASRL}_S > _{\text{IMPORT EN,RO,S}} \text{WASRL}_S <_{\text{SIMPL RO}} \text{SRL}\]
Instantiation conditions

• Each instantiation should satisfy the constraints:
  – on the input file,
  – on the output file,
  – on all processing steps.
Instantiation conditions: the first flow

• The input condition: a document containing a parallel EN-RO translation of a content S
  – such a document is given to the start node
    2TXT$_{EN,RO,S}$ $\rightarrow$ true;

• Edge instantiation conditions:
  – the restrictions of the pipeline edge $>$$_{SEM-LAB\_EN}$: sentences
    are marked with the attribute-value pair
    language=“EN” $\rightarrow$ true;

• The destination condition:
  – sentences marked with language = “RO” should
    include also SR markings $\rightarrow$ false.
Instantiation conditions:  
the second flow

• The input condition: a document containing a parallel EN-RO translation of a content S
  – same as above, such a document is given to the start node
    $2\text{TXT}_{\text{EN,RO,S}} \rightarrow \text{true}$;

• Edge instantiation conditions:
  – $\text{ALIGN}_{\text{EN,RO,S}}$: the two languages must be EN and RO and the texts have the same content $\rightarrow \text{true}$
  – the simplification edge $\text{SIMPL}_{\text{RO}} \rightarrow \text{prunes out all markings, because no language }= \text{"RO" exists which include also SR markups.}$

• The destination condition:
  – no language $= \text{"RO" sentences including also SR markings }\rightarrow \text{false.}$
Instantiation conditions: the third flow

• The input condition: a document containing a parallel EN-RO translation of a content S
  – same as above, such a document is given to the start node 2TXT\textsubscript{EN,RO,S} \rightarrow \text{true};

• Edge instantiation conditions:
  – a supplementary condition induced by IMPORT\textsubscript{EN,RO,S}: all sentences marked language=“EN” also include SR markups \rightarrow \text{true};
  – IMPORT\textsubscript{EN,RO,S} pipeline \rightarrow sentences marked language=“RO” and including SR markups
  – the simplification edge \textless \textsubscript{SIMP} \textsubscript{RO} \rightarrow leaves only them

• The destination condition:
  – language = “RO” sentences including also SR markings \rightarrow \text{true}.
Conclusions: the FMH

• Functions as a framework for recording processing tools, based on which workflows can be designed and visualised

• Can be shared by a community of researchers and is enriched any time a user “uploads” an annotated document/resource or a processing tool.

• Linguistic resources can be clearly positioned within the hierarchy of annotation schemes.
Conclusions: solving NLP problems

• The model needs two steps from the definition of a NLP problem until the preparation of the actual run:
  – flow computation: at times, the computed flows can be virtually the same irrespective of the language requirements
  – instantiation: differentiates among specific behaviours.
• In line with the trend in modern NLP to separate algorithms from linguistic details.
Conclusions: multilingual features

• Modules designed to perform specific tasks can be put to work on any language if fuelled with appropriate language resources:
  – a POS-tagger is powered by a specific language model (frequency of n-grams)
  – a syntactic parser can be powered by the grammar of a language to be effective in parsing sentences of that language
  – a shallow parser (implementing an abstract automata machinery) could recognize noun phases of one language if powered by a set of regular expressions specific to that language
  – in ATLAS: anaphora resolver, clause segmenter, discourse parser, summariser for BG, DE, EL, EN, PL, RO
Conclusions: multilingual features

• Similar processing for different languages: two or more languages can share the same processing chain, although the component modules may be instantiated differently by the language specific resources they require.
• Identical input-output schemes to accomplish the same task for different languages but different paths \(\Rightarrow\) distinct solutions.
• Processing of multi-language documents: parts of documents including passages in different languages can be the object of distinct processing.
• Snapshots of available processing power for particular languages: language filters can be applied to sieve only tools compatible with one language \(\Rightarrow\) at times: disconnected hyper-graphs (nodes which cannot be reached). Meaning: no solution!
Thank you!

Research partially supported by ATLAS and METANET4U.
Criticism

• The model considers processing tools as being compatible and that there exist a critical mass of tools out of which a sufficiently large hierarchy can be built.
  – however, standards still lack and there is sufficient evidence to believe that they will not appear too soon.
  – even worse, because the domain is so active, the research will always go ahead of any attempt of standardisation, and therefore, tools employing new formats will always exist.

• The answer to that is a line of research aiming to infer the semantic content of linguistic annotations.
  – copying human skills to recognise the significance of markings
  – enabling automatic classification of linguistic resources conforming to a general hierarchy
  – generating automatically a number of convertors, in order to assure interoperability where necessary.