

# Towards a Community-Driven Controlled Natural Languages Evolution

Martin LUTS, Monika SAARMANN, Daniel TIKKERBÄR, and Marius  
KUTATELADZE

*ELIKO Competence Centre in Electronics-, Information- and Communications  
Technologies, Estonia*

*Department of Informatics, Tallinn University of Technology, Estonia*

*M3D Ltd, Estonia*

*FocusIT Ltd, Estonia*

**CNL 2010: 2nd Workshop on Controlled Natural Languages**

Marettimo Island, Sicily, Italy

13-15 September 2010

# Agendum

- 1. Problem statement**
- 2. Inspiration:** IAL, pidgins, OSS development
- 3. Proposal** – process of a community-driven CNL evolution
- 4. Motivating Use-Cases**
  - Information retrieval: semantic similarity of documents based on CNL abstracts
  - community-MT „ Le Petit Prince“ from UNL into CNL-Est & CNL-\*
- 5. Discussion**, call for the community

# Problem statement

**Business opportunity:** exploit CNLs in software (SW) applications

**Problem 1:** SW developers / CNL experts

**Problem 2:** methodology for a CNL design component for SW development is needed

**Research question:** How to lower the barrier of incorporating CNL-components in a SW development projects?

**Solution:** to reglement the development of a CNL component in SW applications

- maturity, must ensure outcome/delivery
- reuse

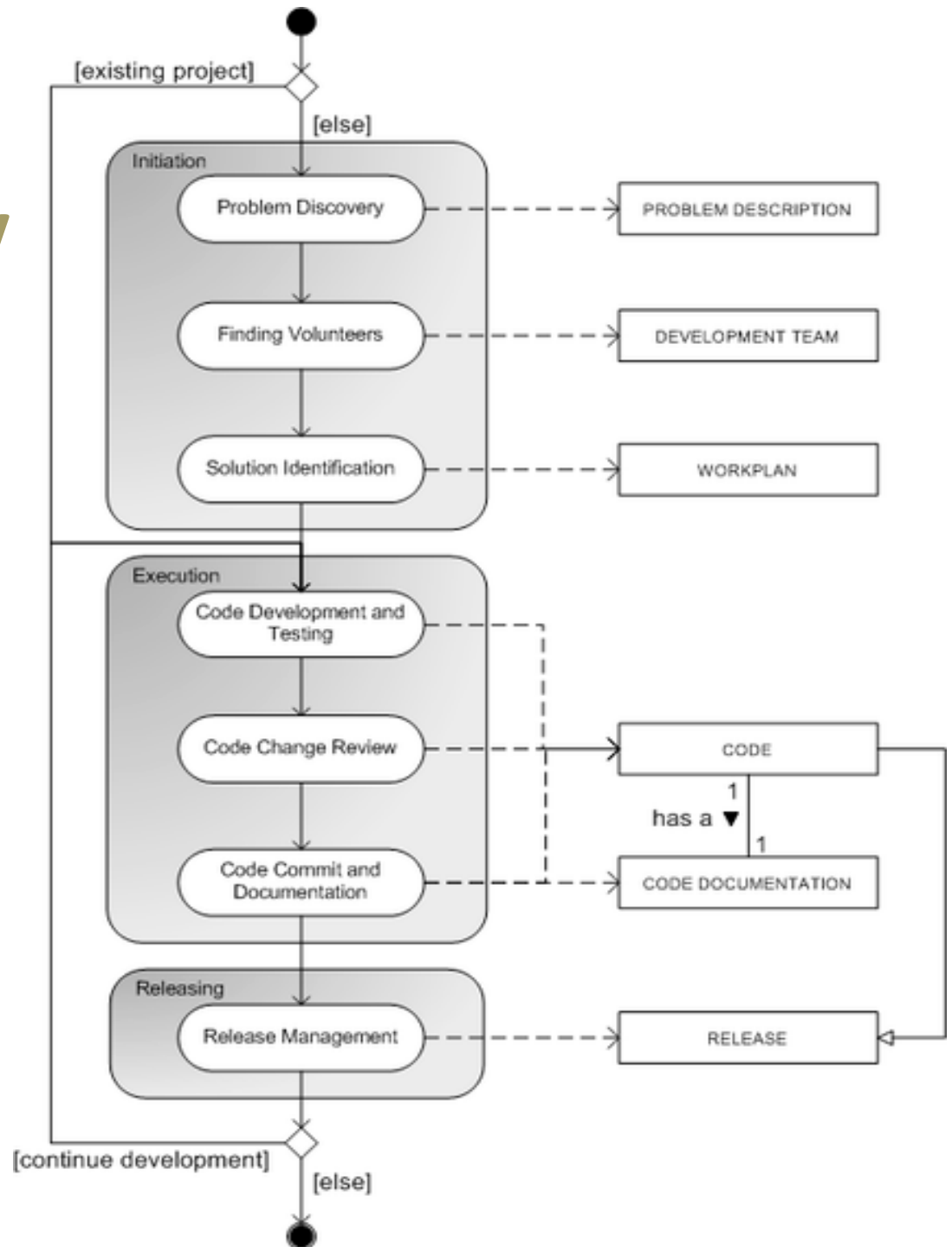
Attribute	Creole and pidgin languages	Planned languages	CNLs
Nature	Simplified languages		
Purpose	Communication between humans across language-barriers		
Creation	<ul style="list-style-type: none"> <li>• <b>Unconsciously</b> born</li> <li>• environment <b>spontaneously</b> gives birth</li> <li>• practical situation</li> <li>• <b>community driven</b></li> </ul>	Created <b>consciously</b>	
Creators <i>vrs</i> speakers	<ul style="list-style-type: none"> <li>• creators = speakers</li> <li>• no “The ONE”. Web2-style</li> <li>• creation while using</li> </ul>	Person/coherent project team => <b>draw</b> a group of speakers to learn	
Outcomes	“Creolization” of some pidgins has given us several lingua francas	<b>No viable language has emerged</b>	
Examples	洋泾浜英语, <b>Lingua franca of the Mediterranean</b> or <b>Sabir</b>	Esperanto, Volapük, Latino sine flexione, Interlingua de IALA, Lojban	

# Inspiration: Open/Community

**CNL engineering – a subfield of  
SW engineering?**  
(Zachmann's approach)

## Open Source SW Development Methodology

- In the past: unstructured, no clear development tools, phases, etc., every project had its own phases.
- More recently there has been much better progress, coordination, and communication within the open source community.



## Task: Adjusting the properties of the CNL



1. See Wyner, A., et al. Controlled Natural Languages: Properties and Prospects. Fuchs, N.E. (ed.). In Proceedings of the Workshop on Controlled Natural Languages (CNL 2009), Marettimo Island, Italy, 8-10 June, 2009. LNCS/LNAI, vol. 5972, Springer, 2010.

Disciplines: CNL component

Expand All Sections Collapse All Sec

### Relationships

<b>Roles</b>	Primary Performer: <ul style="list-style-type: none"><li>• CNL Expert</li></ul>	Additional Performers: <ul style="list-style-type: none"><li>• Product Owner</li><li>• Domain Expert</li></ul>
<b>Inputs</b>	Mandatory: <ol style="list-style-type: none"><li>1. Business case description.</li><li>2. Motivating use cases</li><li>3. Corpus of text samples</li><li>4. Sociolinguistic profiles, abilities and requirements of the user</li></ol>	Optional: <ul style="list-style-type: none"><li>• None</li></ul>
<b>Outputs</b>	<ul style="list-style-type: none"><li>• CNL dimensions in the Wyner framework</li></ul>	

# Process Steps

*Task 1:* Attracting the attention (CNL's business value).

*Task 2:* Elaborating motivating use cases (OpenUP process framework).

*Task 3:* Initial human-authored corpus of text samples.

*Task 4:* Composing the sociolinguistic profiles, abilities and requirements of the user.

*Task 5:* **Adjusting the properties of the CNL.**  
(Wyner et al).

*Task 6:* Selecting reusable components from a CNL repository.

*Task 7:* Customizing these components to the needs outlined in steps 2-5.

# Discussion topics

*Wyner 2010/09/13: “... every CNL project starts from a scratch”*

---

- 1. Establishing a repository** for collecting reusable CNL components:
  - use-case descriptions
  - test-data
  - software
  - linguistic assets, etc.
- 2. Elaborating a “open-source” process description** for creating/ customizing CNLs.
  - EPFWiki?, ..



# Motivating-UCs

UC1. Information retrieval (IR) based on semantic similarity.

UC2. Tagging of digital items.

UC3. Machine translation

- Localizing (open) software

UC4. Communication with a smart environment.

UC5. Management of (controlled) vocabularies.

UC9. Web-service annotation with CNL-based SA-WSDL descriptions.

UC16. Creation and management of BPMN

# MUC1 – Information retrieval based on semantic similarity

- **Main idea:** document abstracts and (wiki) article summaries are written in a Controlled Natural Language,
  - enabling semantic search and article recommendation based on the semantic distance of CNL abstracts
- **Applicability:** search engines, recommendation engines, wiki engines, Electronic Document and Records Management Systems, etc.

# MUC1 – estimating semantic similarity (of scientific articles, documents, ..)

- Encoding a document in a vector is a very crucial step for any vector space model based IR system.
- In traditional document representation methods, a document is considered as a bag of words.
  - The fact that the words may be semantically related is not taken into account.
  - The feature vector representing the document is constructed from the frequency count of document terms.
- Improvement – preprocessing > RI/LDA/...
  - Generating feature vectors using the semantic relations between the words in a sentence.
  - The semantic relations are captured by the Universal Networking Language (UNL) //could be another CNL.

# Motivating-UCs

UC1. Information retrieval (IR) based on semantic similarity.

UC2. Tagging of digital items.

UC3. Machine translation

- Localizing (open) software

UC4. Communication with a smart environment.


UC5. Management of (controlled) vocabularies.

UC9. Web-service annotation with CNL-based SA-WSDL descriptions.

UC16. Creation and management of BPMN

Home » News » Call for Participation in the Project LPP

## Call for Participation in the Project LPP

Monday, 02 August 2010 13:46  Ronaldo Martins



The UNDL Foundation is extending the set of funded languages in the UNL Programme. Financial support will be initially granted to freelancers participating in the project Le Petit Prince (LPP). Any language is eligible, except those already funded by the UNDL Foundation (namely English, French, Arabic, Russian, Spanish, Portuguese and Armenian), which should be pursued in the [project MIR](#).

Contributions are paid through PayPal according to the UNL<sup>dots</sup> system. Tasks are distributed upon availability and will be carried out in a distance-working environment through a specific web interface. Candidates are not required to have any previous experience in natural language processing but are expected to have some acquaintance with descriptive Linguistics and a good knowledge of English. Undergraduate and graduate students of Language Studies and Translation Studies from minority and less-resourced languages are especially welcomed.

- Freelancers participating in the project Le Petit Prince (LPP)
- **Any language is eligible.**
- Tasks are distributed upon availability and will be carried out in a **distance-working environment.**





## Martin Luts

- Profile
- Assignments
- Checking Account
- Statistics
- Personal Data
- Change password
- Change email
- Join a language
- Join a project
- UNLarium
- VALERIE
- Delete Account



## Martin Luts

### Profile

Trainee

### UNLdots

780

### Level

A0

### Certificates

CLEA250 (UNL-NL Dictionary)	100.00%
CLEA450 (NL Dictionary)	0.00%
CLEA700 (Grammar)	0.00%
CUP500 (UNL-ization)	0.00%

### Working languages

Estonian

### Projects

Le Petit Prince

### Permissions

**Questions?  
Comments?**

**Thank you!**

[martin.luts@eesti.ee](mailto:martin.luts@eesti.ee)  
[monika.saarmann@eesti.ee](mailto:monika.saarmann@eesti.ee)

This research was supported by European Social Fund's Doctoral Studies and Internationalisation Programme DoRa

