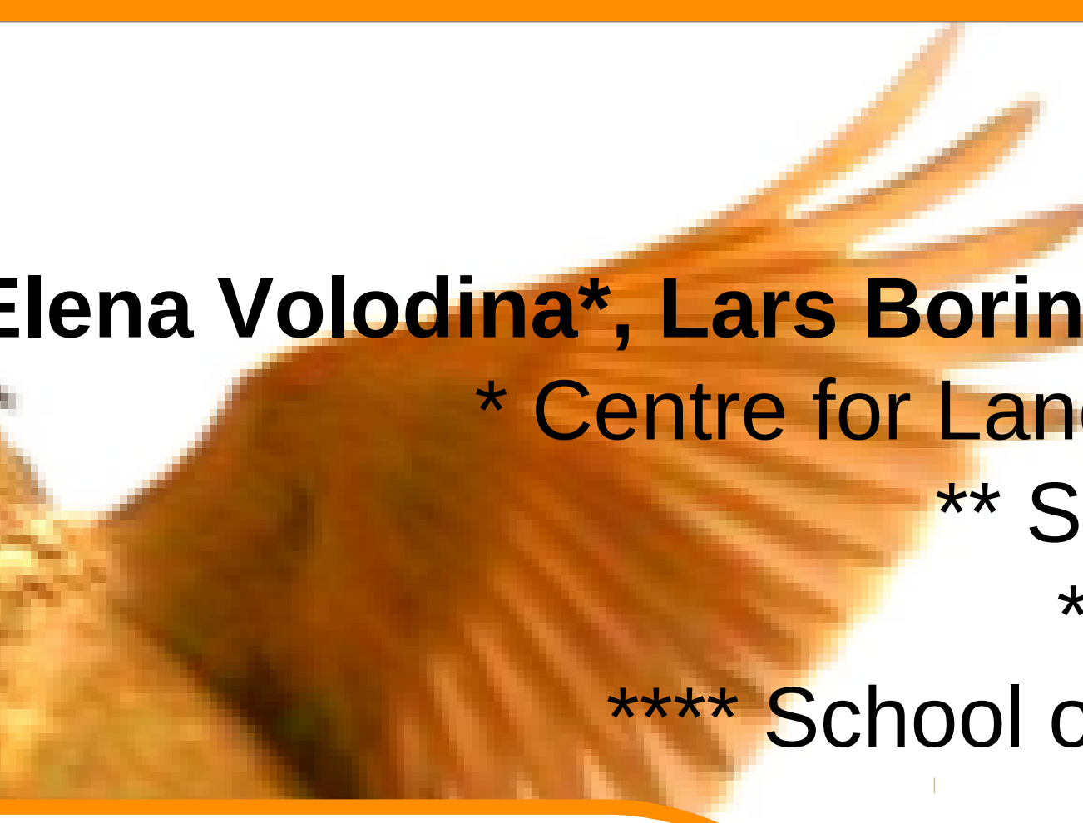
NORDPLUS

CLT

Towards a system architecture for ICALL

based on NLP component re-use



Elena Volodina*, Lars Borin*, Hrafn Loftsson, Birna Arnbjörnsdóttir***, Guðmundur Örn Leifsson******
* Centre for Language Technology, Språkbanken, University of Gothenburg, Sweden
** School of Computer Science, Reykjavík University, Iceland
*** School of Humanities, University of Iceland, Iceland
**** School of Engineering and Natural Sciences, University of Iceland, Iceland

ICALL: Intelligent Computer-Assisted Language Learning

Intelligence is ensured through the use of NLP tools and/or AI techniques

Systems Architecture for ICALL Funded by NordPlus Sprog

Partners: Reykjavik University, University of Gothenburg, University of Iceland

General aim: Encourage and facilitate the use of NLP tools and resources in CALL

Practical aims:

- Design a system architecture for re-use of existing reliable NLP tools/resources in CALL
- Implement applications for testing architecture

Principles for architecture and applications:

- Open-source, re-use, language independent
- Easy to adapt to new tasks
- Plug-and-play basis, modularity

NLP-based CALL:

- is characterized by use of NLP tools/ resources
- ensures linguistic analysis of the input data through use of NLP tools/ annotated resources
- adds generative power of applying the same analysis model to different (authentic) language samples over and over again, e.g. for generating exercises or detect errors in text production
- relieves teachers of monotonous tasks that can be modeled by computers
- supports self-learning for students where it is feasible and motivated
- popularizes NLP among CALL end-users

Re-using NLP components

Most NLP components are:

- Monolithic and inflexible; need to be individually adapted to every new application
- Not readily available as the rights are held by individuals or institutions all over the world
- Physically located in different places
- Not interoperable via standardized interfaces

Strategies for making use of them:

- Rewrite in the target programming language
- Find chunks of similar code and build upon it using open-source initiatives, e.g. http://www.fsf.org
- Standardize communication between the tools and resources: e.g. initiatives for corpora (EAGLES, TEI, etc.); for e-learning (IMS Global Learning Consortium, SCORM, etc.); for NLP tools (GATE, NLTK, Apache UIMA) – still bound to programming languages (Java, Python)

SOA & web services: an approach to NLP component re-use

Service Oriented Architecture (SOA) principles:

- Modular services that can be re-used by others
- Communication layer with a well-defined interface for sending a request and getting a response
- Standardized data output format
- Well-documented interface and its service
- Services loosely coupled and can be re-combined

Web services as an implementation technology:

- Wrapper around a program making it accessible world-wide
- Can re-use other web services, databases, resources, etc.
- Access over Internet; the original software can still be residing on its original server
- Standardization initiative: trying to attract software and resource owners to provide web services

The Icelandic work

NLP and ICALL for Icelandic:

- IceNLP: Open source collection of tools for processing and analyzing Icelandic texts. Contains, e.g., a tokenizer, an unknown word guesser, a PoS tagger, and a shallow parser.
- Currently, no ICALL application exists for Icelandic
- *Icelandic Online* (IOL) is a CALL application (web course) with almost 90,000 registered users

ICALL platform:

- Being developed for supporting ICALL systems
- Connects various pre-existing NLP tools from IceNLP
- Uses the *Text Corpus Format* (TCF) for communication of information between components
- Individual components can be accessed through a web service

Output from the platform in TCF:

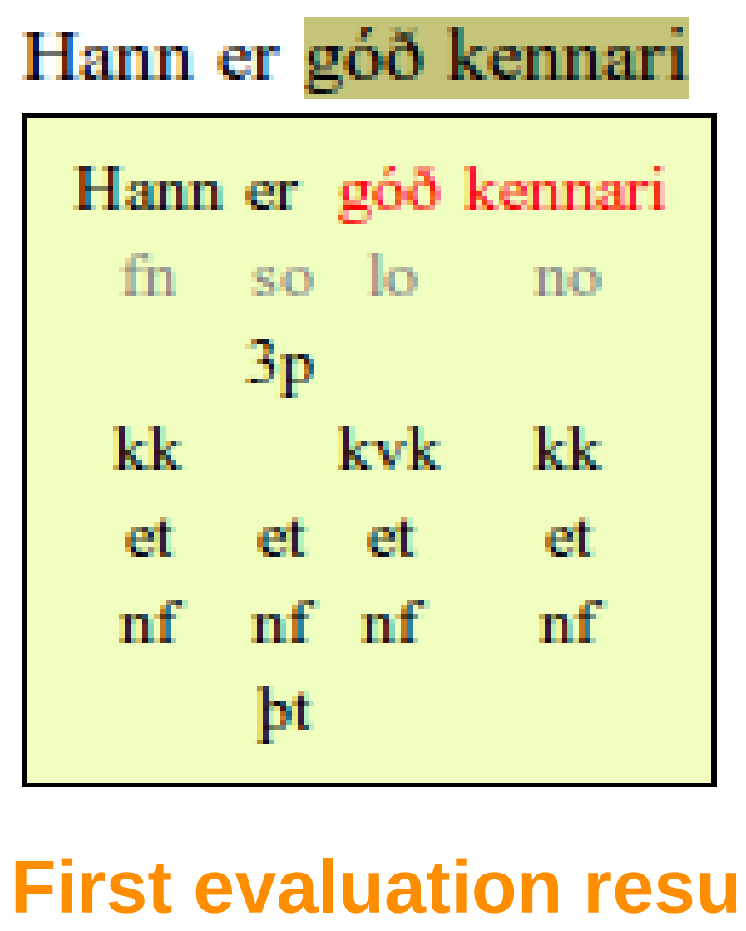
```
<text>Hann er góður kennari</text>
  <tokens>
   <token ID="t1">Hann</token>
   <token ID="t2">er</token>
   <token ID="t3">góður</token>
   <token ID="t4">kennari</token>
  </tokens>
  <POStags tagset="ifd">
   <tag tokenIDs="t1">fpken</tag>
   <tag tokenIDs="t2">sfg3en</tag>
   <tag tokenIDs="t3">lkensf</tag>
   <tag tokenIDs="t4">nken</tag>
  </POStags>
```

Writing support for second-language learners

- A web service for helping students (of IOL) correct second language grammar issues
- Detects particular types of grammatical errors in texts.
- Feature agreement errors:
  - (1) in noun phrases
  - (2) between subjects and verb complements
  - (3) between subjects and verbs
- (4) incorrect case selection of verb objects
- A student corrects potential errors, re-submits the text, and so on, before finally submitting it to teacher.

How does it work?

- Using a web service, a web application instructs the platform to analyze a text and carry out error detection
- The platform calls the appropriate tools in IceNLP to carry out the given task
- The result in TCF is sent back to the web application
- The application displays the original text submitted along with error candidates highlighted and morphological information:



Hann er góð kennari

Hann	er	góð	kennari
fn	so	lo	no
	3p		
kk		kvk	kk
et	et	et	et
nf	nf	nf	nf
	þt		

The adjective "góð" ('good') does not agree in gender with the following noun "kennari" ('teacher') in the noun phrase "góð kennari".

First evaluation results

Error type	Precision	Recall
Errors in noun phrases	80.0%	100.0%
Errors between subjects and verb complements	100.0%	87.5%
Errors between subjects and verbs	42.9%	42.9%
Incorrect case selection of verb objects	100.0%	50.0%
All types	**76.0%**	**76.0%**

In general students found the system helpful for error detection and that it aided them in their writing.

The Swedish work

Lärka (Eng. Lark) – LÄR språket via KorpusAnalys:

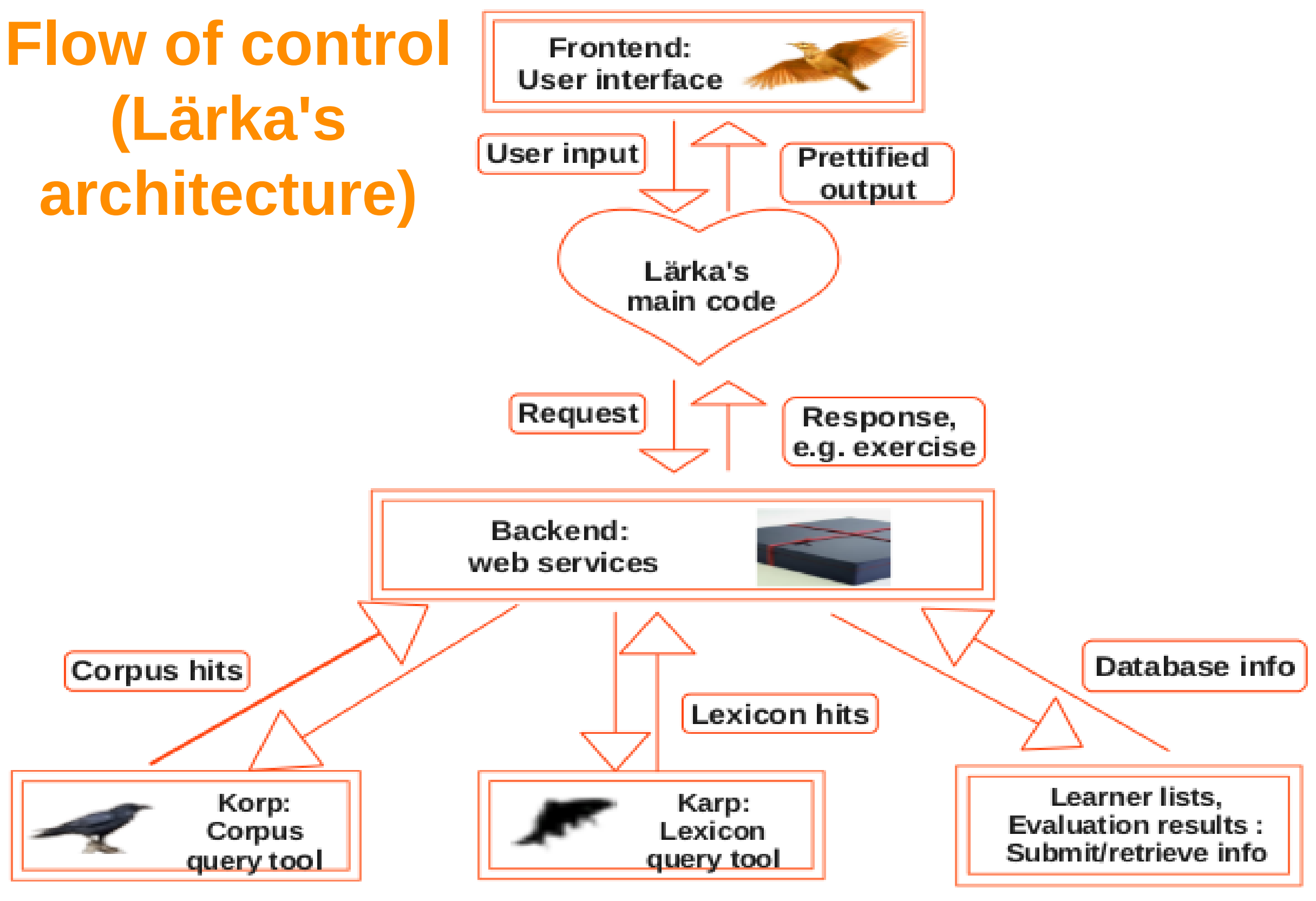
- ICALL platform, at the moment consisting of an exercise generator; eventually other learner-related activities, e.g. rating corpus hits, manipulating vocabulary lists, performing readability analysis, etc.

Characteristics:

- Web-based, modular, plug-and-play principle, SOA-based with web-service implementation
- Underlying corpora: SUC2, Talbanken, LäsBart
- Underlying lexicons: Saldo, Wikipedia, Wiktionary, Lexin
- Underlying word lists: Kelly list, Base Vocabulary pool, Lexin domain lists, Swadesh list, Academic word list, etc.
- Scope: exercise types for linguists and for L2 learners
- Modes: self-study and test

Feedback: in terms of correct/incorrect and lexicon & encyclopedia entries

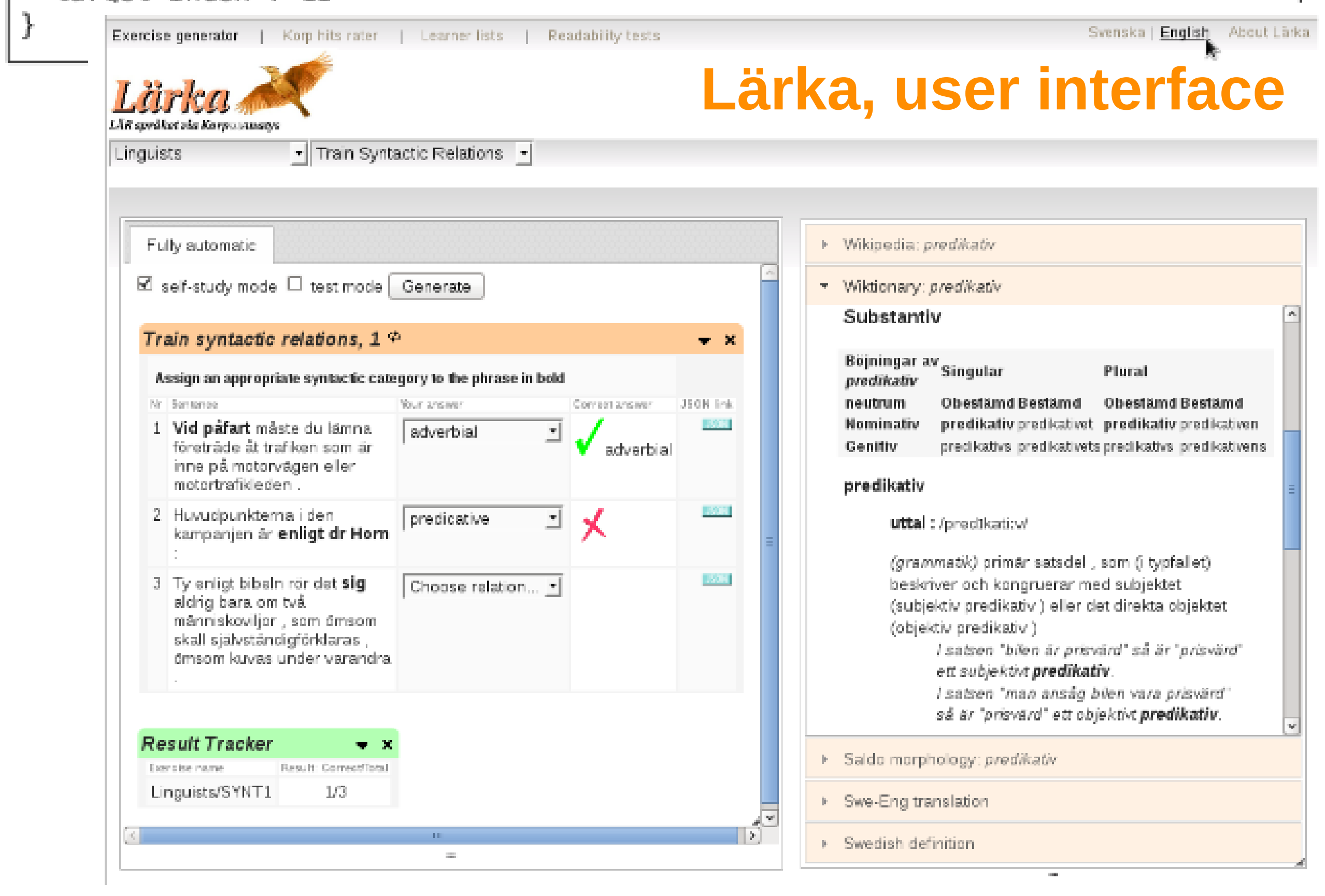
Flow of control (Lärka's architecture)



Output from Lärka's web service in JSON format

```
{
 "corpus": "TALBANKEN",
 "distractors": ["AG", "FV", "IO", "IV", "OO", "SP", "SS"],
 "distractors_en_sv": {
  "AG": {"en": "adverbial", "sv": "adverbial"},
  "FV": {"en": "finite verb", "sv": "finit verb"},
  "IO": {"en": "indirect object", "sv": "indirekt objekt"},
  "IV": {"en": "nonfinite verb", "sv": "infinit verb"},
  "OO": {"en": "object", "sv": "objekt"},
  "SP": {"en": "predicative", "sv": "predikativ"},
  "SS": {"en": "subject", "sv": "subjekt"}
 },
 "exetype": "synt1",
 "sent_index": 3440,
 "sentence_left": "Den ena är att man har en förebild som visar hur ",
 "sentence_right": "ska vara : enheten och kärleken mellan Kristus och
                   de kristna .",
 "target": "äktenskapet ",
 "target_deprel": "SS",
 "target_index": 11
}
```

Lärka, user interface



Assign an appropriate syntactic category to the phrase in bold

Nr	Sentence	Your answer	Correct answer	JSON link
1	**Vid påfart** måste du lämna företräde åt trafiken som är inne på motorvägen eller motortrafikleden .	adverbial	✓ adverbial	
2	Huvudpunkterna i den kampanjen är **enligt dr Horn** :	predicative	✗	
3	Ty enligt bibeln rör det **sig** aldrig bara om två människoviljor , som ömsom skall självständigförklaras , ömsom kuvas under varandra .	Choose relation...		

Exercise name	Result: Correct/total
Linguists/SYNT1	1/3

Substantiv

Böjningar av predikativ	Singular		Plural	
neutrum	Obestämd	Bestämd	Obestämd	Bestämd
Nominativ	predikativ	predikativet	predikativ	predikativen
Genitiv	predikativs	predikativets	predikativs	predikativens

predikativ

uttal : /predikati:v/

(grammatik) primär satsdel , som (i typfallet) beskriver och kongruerar med subjektet (subjektiv predikativ) eller det direkta objektet (objektiv predikativ)

I satsen "bilen är prisvärd" så är "prisvärd" ett subjektivt ***predikativ***.

I satsen "man ansåg bilen vara prisvärd" så är "prisvärd" ett objektivt ***predikativ***.

Lärka, future

- Expand exercise scope, e.g.: gap cloze, wordbank; morphological paradigm, semantic closeness, yes-no diagnostic test, spelling, naming grammatical features; word-order by syntactic group; word-building;
- Add learner lists/lexical database
- Enrich encyclopedia feedback
- Visualize syntactic tree for sentence-based exercises and eventually add exercises based on syntactic trees
- "Hit-ex" - rating corpus search hits (tests are ongoing)
- Readability analysis
- Half-automatic mode for exercise generation (feeding the system with the user choices/lists, etc.)
- Editable "mode" of exercise production – proofreading and modifying automatically created items; saving the items into a database
- Error typology and analysis of written texts etc.

Writing support: http://nlp.cs.ru.is/rithjalp **Lärka: http://spraakbanken.gu.se/larka**